InterpreTabNet: Distilling Predictive Signals from Tabular Data by Salient Feature Interpretation

Jacob Si [1] **Wendy Yusi Cheng*** [1] **Michael Cooper*** [1 2] **Rahul G. Krishnan** [1 2]

Abstract

Tabular data are omnipresent in various sectors of industries. Neural networks for tabular data such as TabNet have been proposed to make predictions while leveraging the attention mechanism for interpretability. However, the inferred attention masks are often dense, making it challenging to come up with rationales about the predictive signal. To remedy this, we propose InterpreTabNet, a variant of the TabNet model that models the attention mechanism as a latent variable sampled from a Gumbel-Softmax distribution. This enables us to regularize the model to learn distinct concepts in the attention masks via a KL Divergence regularizer. It prevents overlapping feature selection by promoting sparsity which maximizes the model's efficacy and improves interpretability to determine the important features when predicting the outcome. To assist in the interpretation of feature interdependencies from our model, we employ a large language model (GPT-4) and use prompt engineering to map from the learned feature mask onto natural language text describing the learned signal. Through comprehensive experiments on real-world datasets, we demonstrate that InterpreTabNet outperforms previous methods for interpreting tabular data while attaining competitive accuracy.

1. Introduction

Machine learning methods for tabular data enjoy broad applications in diverse settings like healthcare (Clore et al., 2014), insurance (Datta, 2020), and finance (Moro et al., 2012). While predictive performance is key in these settings, practitioners often aim to translate predictive models into intelligible insights. For example, a medical practitioner working with tabular Electronic Health Records may be interested in determining features that contribute to a patient's diagnosis. Furthermore, an insurance underwriter working with tabular client data focuses on determining crucial factors that influence a client's risk profile.

Despite commendable advancements made by existing models such as TabNet (Arik and Pfister, 2020), there remains a discernible gap in achieving an integration of accuracy and interpretability. TabNet's ability to generate learnable masks for salient feature interpretation is limited as its interpretation is ambiguous. The considerable overlap between multiple masks makes it challenging for a user to discern the salient features used by the model for reasoning at each decision step. Other means of interpreting tabular models, such as attention weights (Vaswani et al., 2017) and SHAP values (Lundberg and Lee, 2017) have been criticized for their inconsistency in providing meaningful insights (Roberts et al., 2022) and the computational intensity required to apply them to complex datasets (Jain and Wallace, 2019). Additionally, tree-boosting methods such as XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017) exhibited limitations in their interpretability aspects when analyzed through the lens of SHAP values. These models tend to distribute the prediction contribution across an extensive range of features (Madakkatel and Hyppönen, 2024), leading to a less sparse representation of feature importance, making it difficult to identify important features.

The objective of our work is to distill the predictive signals from tabular data by enhancing the interpretability of the established TabNet architecture while maintaining competitive accuracy on practical datasets. To do so, we introduce InterpreTabNet, a modified variant of the TabNet neural architecture, enabling us to sparsify the identity of the predictive signals. Our work is premised on the hypothesis that we can map the predictive signals from the TabNet model onto a collection of sparse attribution masks that encode instance-wise feature significance. The sparsity of our masks leads to quick and easy identification of the salient features in the data. Having achieved this, we then enable post-hoc, text-based interpretability, using large language models (LLMs) (OpenAI, 2023) to draw upon rich prior knowledge related to the application domain (Choi et al.,

*Equal contribution [1]University of Toronto [2]Vector Institute. Correspondence to: Jacob Si <jacobyhsi@cs.toronto.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

2022) and provide textual summaries of our sparse masks.

Our work makes the following contributions:

1. We devise *a regularization scheme that maximizes diversity between masks in the TabNet architecture*. This is in contrast to the default "sparsity regularizer" employed by TabNet (Grandvalet and Bengio, 2004); its reliance on entropy often leads to the reuse of features across attention masks within the architecture. Empirically, under our regularization scheme, the model learns to extract salient features and generate sparse masks, reducing these challenges implicit in interpreting the feature masks generated by TabNet. Furthermore, our method suffers from only a modest tradeoff between accuracy and interpretability: we find that our approach performs comparably to the other baselines in accuracy but outperforms them considerably in terms of interpretability.

2. Our regularization scheme relies on maximizing the KL divergence (Kullback and Leibler, 1951) between the distributions from which each TabNet attention mask is implicitly sampled. Whereas the original TabNet formulation does not explicitly characterize these distributions, we leverage tools from variational inference to *model the attention weights within TabNet as samples drawn from a Gumbel-Softmax distribution*. By reformulating the attention weights within TabNet as a latent variable model, we can directly control properties of the mask distributions (such as the KL divergence) using regularized gradient-based optimization. Our architecture can be found in Figure 1.

3. We show that by leveraging rich linguistic priors in a large language model we can capture the rich interdependencies between features that are needed to interpret model predictions in complex settings. We demonstrate how language models can relate the learned feature masks in our model to form detailed hypotheses about what is being learned at each step of the TabNet decision-making pipeline.

2. Related Works

Learning from Tabular Data. Early works on deep learning architecture for tabular data, such as TabNet, use a sequential attention mechanism for tabular data analysis (Arik and Pfister, 2020). Their prominent strength is the capability to outperform other neural networks and decision trees on tabular datasets while yielding some level of interpretability for feature selections. However, TabNet's self-attention transformers' inability to capture diversifying latent variables leads to suboptimal feature selection. To address this limitation, diversity-promoting regularizers and latent models attempted to solve this problem (Xie et al., 2017) (Xie et al., 2016). Subsequent works on tabular data include Net-DNF (Katzir et al., 2020), SubTab (Ucar et al., 2021), and TabTransformer (Huang et al., 2020). Net-DNF (Katzir et al., 2020) introduced an inductive bias that aligns model structures with disjunctive normal form (DNF) and emphasizes localized decisions. SubTab (Ucar et al., 2021) transformed tabular data into a multi-view representation learning task, enhancing latent representation. Furthermore, TabTransformer (Huang et al., 2020) is a deep tabular data modelling architecture built upon self-attention-based Transformers.

Latent Variable Models. Latent variable models like VAEs (Kingma and Welling, 2022) and their variations demonstrate attractive abilities to model complex distributions and produce latent values. DirVAE has more interpretable latent values with no collapsing issues (Joo et al., 2019), while the cVAE (Kristiadi, 2016) models random latent variables and observed data, which gains control of the data generation process on the VAE. Additionally, the cVAE generates diverse but realistic output representations using stochastic inference (Sohn et al., 2015). Transformer-based cVAE exhibits excellent representation learning capability and controllability (Fang et al., 2021). We draw inspiration from these VAE extensions and incorporate the cVAE into TabNet's architecture to capture and reconstruct discrete data. Recent work in approximate inference for categorical data includes Categorical Reparameterization with Gumbel-Softmax (Jang et al., 2016). In our paper, we leverage the Gumbel-Softmax distribution as a key component of our methodology to strike a balance between interpretability and performance.

Model Interpretability. Methods from interpretability aim to surface information about *why* a machine learning model is making certain predictions to the user. Broadly, there are two families of methods in model interpretability. *Intrinsic interpretability* refers to the scenario in which the user can directly leverage the parameters learned by the model to understand the rationale underlying the predictions. Linear models (Gauss, 1877), decision trees, Transformers (by means of their learned attention weights), and TabNet (Arik and Pfister, 2020), are all, to varying degrees, intrinsincally interpretable methods. In contrast, methods from *post-hoc interpretability* tackle the scenario in which the model may be black-box: these methods instead attempt to approximate the decision-making process underlying the model, which is then surfaced to the user. Methods like SHAP (Lundberg and Lee, 2017), LIME (Ribeiro et al., 2016), and Grad-CAM (Selvaraju et al., 2017) are methods for post-hoc interpretability. The central tradeoff between intrinsic and post-hoc interpretability is this: while an intrinsically interpretable model is (definitionally) faithful to its underlying decision rule, it may be necessary to make simplifying

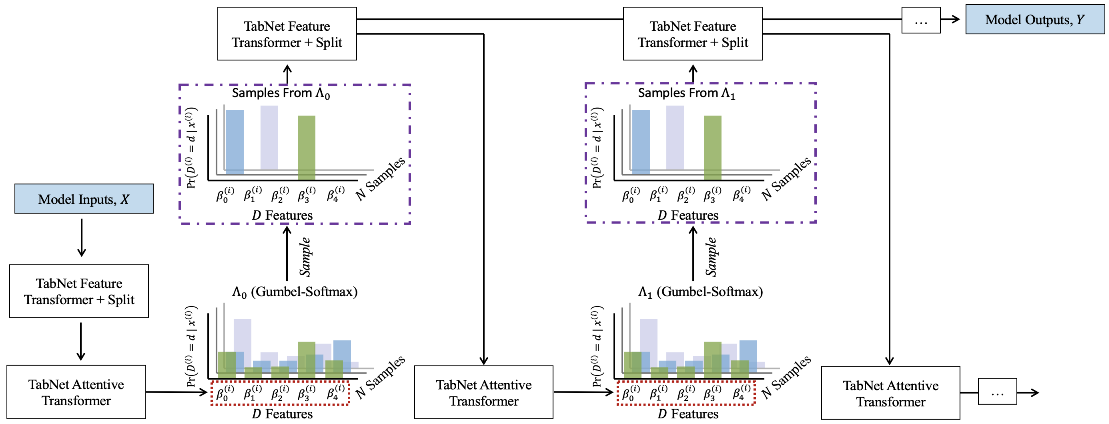


Figure 1: The InterpreTabNet architecture presents a variational formulation of the TabNet encoder. In our formulation, the weights of the attention masks produced by the TabNet encoder at each step k are treated as the parameters, $\beta_0^{(i)}, ..., \beta_{D-1}^{(i)}$, of a Gumbel-Softmax distribution, Λ_k, unique to each instance (shown by the red dotted rectangle). This distribution is then sampled to produce a single feature that is highlighted for each feature at each step (purple dot-dashed rectangle). This figure shows $k = 2$ steps of the encoder architecture, over $D = 5$ features, for $N = 3$ samples.

assumptions in the design of the model. Conversely, while post-hoc interpretability methods can interpret models of arbitrary complexity, the interpretable decision rule surfaced by such procedures is only an approximate one (Du et al., 2019). Our approach draws upon insights from both classes of methods: we leverage tools from variational inference to improve upon the intrinsic interpretability of TabNet, and we employ a large language model to provide a richer contextual interpretation of the learned features post-hoc.

3. The InterpreTabNet Model

Let $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathcal{X} \times \mathcal{Y}$ represent the covariates and an outcome that we want to model, respectively. As we are operating in the tabular data regime, assume that $X \in \mathbb{R}^{N \times D}$, where each $d \in [1, ..., D]$ corresponds to a single discrete feature in the data. Then, each $x^{(i)}, y^{(i)}$ represents D-vector and label corresponding to a particular example.[1] Let $P(\cdot|\cdot)$ denote true probability density functions, and $Q(\cdot|\cdot)$ denote variational approximations of those densities.

3.1. High-Level Approach

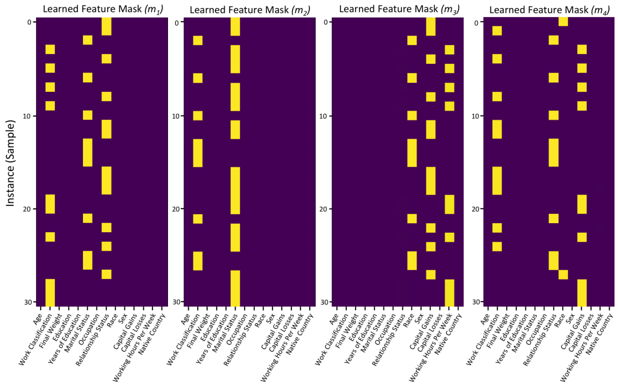
The TabNet encoder architecture models the predictive signal, $P(y \,|\, x)$, as a nonlinear combination of the covariates, x, and a sequence of k *learned attention masks*. Feature importance mask m_k depicts the feature selected at the k-th decision step. We learn each mask, m_k, by applying the TabNet Transformer in the encoder to the covariates and previous attention mask at each step of a multi-step decision process. Since the nonlinear combination is modeled using a multi-layer perceptron (Haykin, 1994), inference within TabNet's encoder can be expressed as:

$$\Pr(y \,|\, x) = f_\psi^{(\text{MLP})} \left(\sum_{k=0}^{K-1} f_\psi^{(\text{TabNet\_Transformer})}(m_k, x) \right), \quad (1)$$

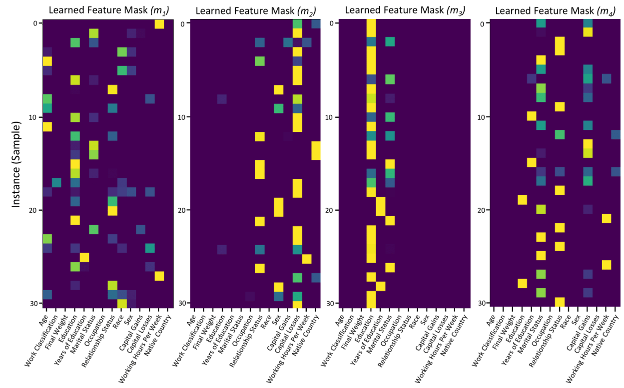
where $m_k = \varnothing$ if $k = 0$, and where ψ is a general-purpose variable to denote the parameters that are associated with a given MLP or TabNet Transformer sub-model. Our goal is to construct a version of this model wherein each mask is a latent variable in a deep generative model. Then we can learn the model via amortized variational inference by inferring m_k using some parametric distribution Q that admits backpropagation by means of the reparameterization trick. By specifying the form of the distribution mask samples are from, we can directly adjust the properties of this latent variable by regularizing the loss function. Specifically, as our objective is to promote sparsity among the masks, we will then aim to maximize the KL divergence between subsequent masks of the decision steps.

In the following sections, we demonstrate how we sample the masks in our architecture from a Gumbel-Softmax distribution (Jang et al., 2016). We choose Gumbel-Softmax as a natural sampling distribution for the masks because the salience of a feature can be treated as a categorical variable: for each example i in mask k, a feature j can either be "selected" ($m_{k_{ij}} = 1$), or "not selected" ($m_{k_{ij}} = 0$). The Gumbel-Softmax distribution offers a continuous relaxation of a categorical distribution, thus facilitating the application of the reparameterization trick under our method.

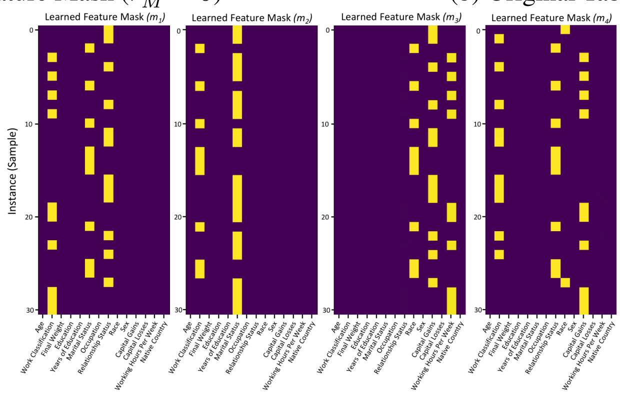
[1] Unless otherwise stated, our notation uses uppercase letters to refer to distribution-level quantities, such as the distribution over the covariates, and lowercase letters to refer to specific samples drawn from those distributions.



(a) InterpreTabNet Feature Mask ($r^*_M = 9$)



(b) Original TabNet Feature Mask



(c) Stacked InterpreTabNet Feature Masks

Figure 2: X/Y-axis labels denote the features and test samples for each respective mask at $N_{steps} = 4$ decision steps of the Adult Census Income dataset. **Left (a)**: Learned masks associated with InterpreTabNet. Observe how for each example, there is no overlap in the attention learned across different masks with high salience. This mutual exclusivity of attention across masks makes for easier visual interpretation of the learned signal that InterpreTabNet leverages in its predictions. **Right (b)**: Learned masks associated with TabNet. Observe how, for each example, there exist overlaps in the attention learned for each mask with no clear salience. This makes the masks challenging to interpret, as there is no obvious way to reconcile attention that is distributed across multiple masks in this manner. **Bottom (c)**: Stacked InterpreTabNet Feature Masks between subsequent feature masks (Left to Right: Masks 0 & 1, 1 & 2, 2 & 3, 3 & 0) outlining no overlap and sparsity in feature selection. More details can be found in Section 4.2.

3.2. Mask Sampling Process

The mask sampling process for InterpreTabNet is the following, where Y represents the predicted outcome, z represents the concatenation of all the m_k mask samples from a Gumbel-Softmax distribution, X represents the data, and Λ represents a Gumbel-Softmax distribution.

$$
\begin{aligned}
P(m_k|X) &\sim \Lambda_k(\text{TabNet\_Transformer}(X)), \\
&\quad \text{for } k = 0 \\
P(m_{k+1}|m_k, X) &\sim \Lambda_k(\text{TabNet\_Transformer}(m_k, X)), \\
&\quad \text{for } k \in [1, \dots, K-1].
\end{aligned}
$$

Unlike TabNet, InterpreTabNet does not only leverage its feature importance masks to make predictions; instead, the feature importance masks serve as the emissions of a stochastic process that we regularize in order to promote sparsity. InterpreTabNet utilizes these masks from the first iteration onwards (after the zeroth iteration) as latent variables. These latent variables serve as a rich source of embedded knowledge, allowing the model to improve its generalizations by acting as a stochastic process. Furthermore, sampling this latent variable from the Gumbel-Softmax distribution will act as a crucial component in improving interpretability (details explored in Section 3.4).

Let us represent the collection of all k masks, $[m_0, ..., m_{k-1}]$ as a single latent variable, $z \in \mathbb{R}^{N \times k}$, drawn from a Gumbel-Softmax distribution. Drawing samples z from a categorical distribution with class probabilities π is as follows.

$$
z = \text{one\_hot}\left(\arg\max_i(\beta_i + \log \pi_i)\right)
$$

where $\beta_0, ..., \beta_{D-1}$ are i.i.d samples drawn from a standard Gumbel distribution, Gumbel(0,1). As a small technical note, the original TabNet architecture requires a ReLU function to be applied to the embeddings between blocks. Our sampling scheme ensures nonnegative mask values, so this requirement is not necessary in our architecture.

The mask sampling process is characterized as a latent variable problem. Thus, this necessitates the implementation of inference techniques for effective learning.

3.3. Generating Predictions with the Conditional Variational Autoencoder

We interpret TabNet's encoder-decoder architecture as a conditional variational autoencoder (cVAE) (Kingma and Welling, 2022; Blei et al., 2017). We imagine an encoder conditioned on two variables, Y and X, which leverages the distribution $Q(z|Y,X)$ to sample the feature masks, z. Similarly, we imagine a decoder that conditions on the feature masks, z, and the data X, to predict a corresponding label drawn from $P(Y|z,X)$. Using this framework, we can derive a variational lower bound on this cVAE. We do so by modelling the outcome, $P(Y|X)$ as $\int P(Y|X,z)P(z|X)dz$, and inferring $P(z)$ through $P(z|Y)$ using $Q(z|Y)$. See Figure 3 for the graphical model. The derivation can be found in Appendix A.1.

$$\begin{aligned}&\log P(Y|X) - D_{KL}[Q(z|Y,X)||P(z|Y,X)] \\ &= E[\log P(Y|z,X)] - D_{KL}[Q(z|Y,X)||P(z|X)] \quad (2)\end{aligned}$$

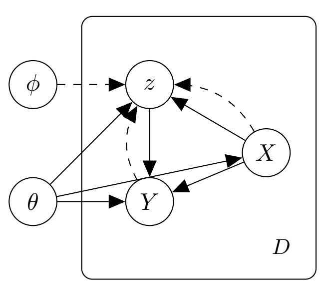


Figure 3: Graphical model of InterpreTabNet with D i.i.d samples. Solid lines denote the generative model $p_\theta(Y|z,X)p_\theta(z|X)$, dashed lines denote the variational approximation $q_\phi(z|X,Y)$ to the intractable posterior $p_\theta(z|X,Y)$. The variational parameters ϕ are learned jointly with the generative model parameters θ.

3.4. Sparsity-Promoting Regularization

Our formulation of TabNet as a stochastic cVAE allows us to directly promote mask sparsity by using the loss function to encourage variation in the Gumbel-Softmax distributions corresponding to adjacent masks. To do so, we incorporate a KL Divergence Sparsity Regularizer (r_M) in the model architecture. With the KL Divergence, we aim to maximize the difference between the distribution of masks that are subsequent to one another. This would reduce the number of selected features, ensuring that the features selected are independent between masks. Additionally, with a sparser feature selection, the model can focus on fewer high-salience features. Therefore, the ELBO of the InterpreTabNet model is as follows with r_M as a tunable regularizer weight.

$$\begin{aligned}&E[\log P(Y|z,X)] - \sum_i D_{KL}\Big(\big(Q\,(z_i|Y,X)\,\big)\,\Big|\Big|\,\big(P\,(z_i|X)\,\big)\Big) \\ &+ r_M \cdot \sum_{i \neq j} D_{KL}\Big(\big(Q\,(z_i|Y,X)\,\big)\,\Big|\Big|\,\big(Q\,(z_j|Y,X)\,\big)\Big) \quad (3)\end{aligned}$$

3.5. Sparsity Regularizer (r_M) Algorithm

To assess the level of interpretability a feature mask provides, we divide it into two sets of criteria. Note that there are more criteria than those we enumerate here. Those enumerated here are exemplars to aid understanding.

1. Number of selected features (e.g., number of important features must be at least 2-3).
2. "Salience" of each feature (e.g., percentage of importance captured by one feature in each mask must be between 20 and 25%).

Within a feature mask, we would like to swiftly identify the salient features that contribute to its prediction. Thus, our aim is to *minimize the number of selected features*, and only select those of *high salience*, while maintaining a competitive accuracy. This would yield an interpretable mask to determine the important features.

We propose an adaptive algorithm to optimize our KL Divergence Sparsity Regularizer, r_M, to improve the interpretability of the feature masks. Our method involves iterative training and evaluation of the InterpreTabNet model with varying values of r_M within a pre-defined range, to check the fulfilment of the above criteria. The end result is the optimal r_M value corresponding to a balance between an interpretable feature mask and classification accuracy, improving the overall efficacy of our model. The algorithm and the full set of criteria can be found in Appendix A.2.

4. Experiments and Discussions

We evaluate the performance of InterpreTabNet on real-world classification tasks both quantitatively and qualitatively. Our analysis sections (Sections 4.1 and 4.2) are based on the Adult Census Income (Becker and Kohavi, 1996) dataset for simplicity purposes. See Appendix C for full results from the other datasets.[2]

[2]The code is available on GitHub at: https://github.com/jacobyhsi/InterpreTabNet

Table 1: Test Accuracy Scores (mean and standard deviations across 20 random seed trails in %) across Different Models and Datasets with Optimal Mask Regularizer Values (r_M) for InterpreTabNet. InterpreTabNet achieves substantial improvements in interpretability across all the datasets and remains competitive in terms of accuracy in most datasets.

Model / Dataset	Adult Census	Forest Cover	Poker Hand	Mushroom	Blastchar	Diabetes	Higgs
InterpreTabNet	**87.42 ± 0.55**	**94.75 ± 0.53**	99.50 ± 0.48	96.62 ± 0.35	72.96 ± 0.56	55.37 ± 0.47	53.08 ± 0.56
Original TabNet	85.55 ± 0.56	94.18 ± 0.63	99.00 ± 0.62	99.94 ± 0.31	76.22 ± 0.34	56.91 ± 0.53	52.94 ± 0.45
XGBoost	86.60 ± 0.64	92.30 ± 0.62	75.57 ± 0.47	99.69 ± 0.39	77.29 ± 0.53	**61.44 ± 0.32**	**72.70 ± 0.35**
LightGBM	86.20 ± 0.43	86.38 ± 0.64	78.47 ± 0.36	**100.00 ± 0.37**	**77.86 ± 0.46**	60.87 ± 0.39	72.62 ± 0.35
TabTransformer	85.09 ± 0.39	82.55 ± 0.39	**99.81 ± 0.31**	**100.00 ± 0.57**	73.17 ± 0.56	44.45 ± 0.34	51.97 ± 0.54
MLP	79.76 ± 0.65	84.89 ± 0.56	99.70 ± 0.56	99.82 ± 0.56	75.16 ± 0.61	53.99 ± 0.46	63.17 ± 0.36

Datasets. The real-world tabular datasets we use in our experiments are from the UCI Machine Learning Repository (Kelly et al., 2023) and OpenML (Vanschoren et al., 2013). These datasets were selected since they were utilized to evaluate the existing methods (baselines). Additionally, they vary in size and nature, with both categorical and continuous features, to ensure a holistic evaluation of our methodology across multiple domains and scenarios. The training/validation/testing proportion of the datasets for each split is 80/10/10% apart from the Higgs dataset. Due to the inherently large Higgs dataset, we adhere to TabNet's method of data splitting with 500k training samples, 100k validation samples, and 100k testing samples. Details of the datasets can be found in Appendix B.2.

Baselines: Accuracy. We compare our model against five other ML methods for tabular classification. These include the Original TabNet, XGBoost (Chen and Guestrin, 2016), LightGBM (Ke et al., 2017), TabTransformer (Huang et al., 2020), and multi-layer perceptrons (MLP) (Haykin, 1994). For each model, we utilize the recommended hyperparameters mentioned by the authors of their respective papers. Furthermore, we also conduct a grid search within the range of the recommended hyperparameters to optimize the models, selecting the best-performing hyperparameter configuration.

Baselines: Interpretability. We compare our model against four other ML methods to determine which model allows the user to easily determine the important features when predicting the outcome. These include the Original TabNet, XGBoost, LightGBM, and TabTransformer. We excluded MLPs as they perform notably worse than the other models in accuracy. The interpretability figures for InterpreTabNet, Original TabNet, XGBoost, and LightGBM are feature masks whereas TabTransformer uses an attention mask. In order to compare the interpretability of feature masks between InterpreTabNet, Original TabNet, XGBoost, and LightGBM, we conduct row-wise normalization on the absolute SHAP values from XGBoost and LightGBM. This yields the same feature importance scale (relative importance of each feature within each sample's prediction) as InterpreTabNet and Original TabNet.

4.1. Quantitative Analysis

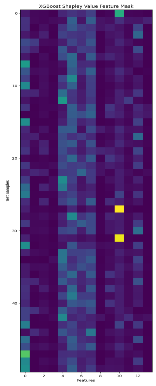
Performance against Baselines. The performance of our method relative to the baselines for tabular learning is shown in Table 1. We achieve the best performance in 2/7 datasets while maintaining a competitive accuracy for the remaining 5/7 datasets. Our most notable contribution is achieving a significant improvement in interpretability.

Model Faithfulness (Quantitative). We conducted synthetic dataset experiments to ensure that InterpreTabNet is faithful to its predictions. We used the same synthetic data generation model in TabNet (Arik and Pfister, 2020) and INVASE (Yoon et al., 2019). InterpreTabNet outperforms existing methods in 4 out of 6 synthetic datasets, indicating that its predictions adhere to the ground truth quantitatively. See Table 9 Appendix D for more details.

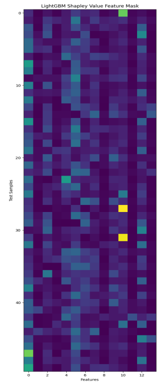
Computational Efficiency. Our model necessitates an additional computation through the Gumbel-Softmax reparameterization and conditioning on the mask from the previous time step when compared to TabNet. Nonetheless, this extra step incurs a minimal cost, leading to a mere several-minute increase in training time. Furthermore, likewise to TabNet, our model maintains greater computational efficiency than other baseline models without necessitating an extensive search for fine-grained hyperparameters.

4.2. Qualitative Analysis

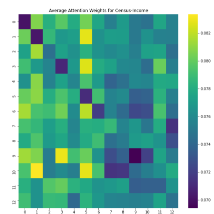
Interpretability Evaluation. Figure 2 highlights the learned masks associated with InterpreTabNet using a sparsity regularizer value of $r_M = 9$ compared to those of TabNet. The rows of each mask represent individual data samples, while the columns represent discrete features in the tabular data. Values of feature importance for each test sample/row sum up to 1. Thus, bright yellow squares indicate values close to/equal to 1, dark purple squares indicate values close to/equal to 0, and other colored squares indicate values between 0 and 1. As observed in Figure 2, our InterpreTabNet model highlights mutually exclusive features of high importance that are easily interpretable. Practitioners can easily identify the salient features contributing to the



(a) XGBoost SHAP



(b) LightGBM SHAP

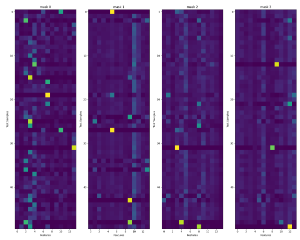


(c) TabTrans. Att. Wgts.

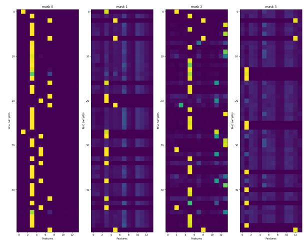
Figure 4: (a), (b), and (c) depicts the XGBoost SHAP Feature Mask, LightGBM SHAP Feature Mask, and TabTransformer Attention Weights for the Adult Census Income Dataset, respectively. X/Y-axis labels denote the features and test samples for the feature masks in (a) and (b) whereas only features for the attention weights in (c).

outcome prediction. On the contrary, feature masks of the Original TabNet are more difficult to interpret since each mask highlights multiple features for a given data sample.

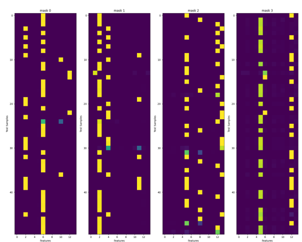
When comparing against the other baselines, InterpreTabNet leverages sequential decision-making that allows users to understand how the model's focus shifts and how different features interact over the decision process. Figure 4 illustrates the complex pathways of model interpretation inherent in our baseline architectures like XGBoost, LightGBM, and TabTransformer. These models necessitate additional tools to render interpretative insights. Both XGBoost and LightGBM are augmented with SHAP values derived from external SHAP packages to achieve interpretability while TabTransformer relies on attention weights. The reliance on SHAP values in XGBoost and LightGBM found in Figures 4(a) and 4(b) distribute the contribution of the prediction across all features. This leads to a less sparse representation of feature importance. In practice, this means that while each feature's contribution to the prediction is identified, the significance of each feature is not as distinct. This results in an interpretation where barely any feature stands out, especially in models with a large number of features where many contributed incrementally to the final prediction. This lack of sparsity makes it challenging for practitioners to pinpoint a concise set of features for understanding and analysis. Note that we utilized feature masks to illustrate the SHAP values of XGBoost and LightGBM to ensure a consistent basis for comparison with InterpreTabNet. The attention mechanism of TabTransformer in Figure 4(c) provides a form of interpretability by capturing relationships between features. However, it is unable to pinpoint a set of important features. Furthermore, attention weights are typically dense, meaning that most features will get some level of attention.



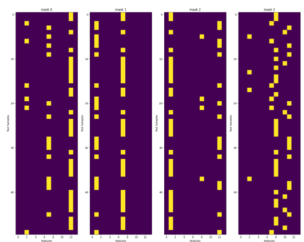
(a) $r_M = 0$



(b) $r_M = 1$



(c) $r_M = 100$



(d) $r_M^* = 1000$

Figure 5: InterpreTabNet Sparsity Regularizer r_M Feature Mask Ablation. As the r_M value increases, both feature mask sparsity and feature importance increase, improving the interpretability of the masks.

Ablation: Affect of r_M Regularizer. Figure 5 illustrates an ablation study on how varying r_M values affect our masks. We notice that at low r_M values, test accuracy is high but feature selection diversity is poor. This makes mask interpretation difficult since almost all features are selected in the decision-making process. On the other hand, at high r_M values, the masks are sparse and are easily interpretable but at the cost of accuracy. Therefore, we ensure that our selected r_M using the Sparsity Regularizer Algorithm in Section 3.5 provides us with a compromise of a competitive accuracy while having the best interpretability against the baseline models.

Model Faithfulness (Qualitative). To reinforce InterpreTabNet's ability to generate faithful feature masks, we assess our synthetic data generation process qualitatively. Depicted in Figure 12 Appendix D, InterpreTabNet determines features 2-5 to be the most salient features as shown by the yellow bars in most of its masks, aligning with the ground truth where features 2-5 are used to generate the synthetic data's predictions. On the other hand, TabNet illustrates lower salience levels in features 2-5 when conducting its decision process for the prediction, indicating its uncertainty in reasoning when predicting the ground truth.

Robustness and Reliability of Feature Masks. To assess the robustness and reliability of InterpreTabNet's feature masks, we examine if salient feature identification differs across subgroups. We divide the Adult dataset into male and female subpopulations, with the feature importance visualization depicted in Figure 13 Appendix E.1. The salient features match well with known socio-economic factors affecting income. For males, occupation and capital gains highlight the importance of job roles and investments. For females, marital status notably influences income, likely due to the socio-economic dynamics and potential household income sharing.

Rationale of Maximizing Feature Mask Diversity. One could argue that if TabNet's feature masks often utilize the same features across stages, it might indicate that fewer features or stages are required for accurate predictions. Maximizing diversity could potentially introduce features that do not contribute to the prediction, thus complicating the model unnecessarily. However, we observe in Figure 14 Appendix E.2 that even when we select the least possible decision steps, $N_{steps} = 2$, TabNet utilizes every single feature aggregated across the two masks. On the contrary, InterpreTabNet selects clear salient features in its decision-making process, without introducing features that do not contribute to the prediction.

Training Stability. In Figure 15 Appendix E.3, we observe that InterpreTabNet's training loss shows a general downward trend. Although InterpreTabNet exhibits higher variability in loss reduction across epochs, it suggests a more exploratory learning process to determine the salient features. Overall, InterpreTabNet's training process is relatively smooth compared to TabNet.

Human Evaluation Survey on Interpretability. We conducted a small-scale human evaluation survey on 20 Ph.D. and Masters students combined, with a machine learning background. It is conducted in a blind format where the identities of the models are anonymous to ensure that our data is trusted and not biased. The survey asks: *"Which figure do you think is the best method to determine the important features?"*.

Table 2: Survey on Interpretability

Model	Vote Percentage	Number of Votes
InterpreTabNet	**65%**	**13**
TabNet	15%	3
XGBoost	5%	1
LightGBM	5%	1
TabTransformer	10%	2
Total	100%	20

In Table 2, 65% of respondents prefer InterpreTabnet to highlight salient features. This validates our motivation that sparse feature selection simplifies the complexity of the data into a more understandable form for practical applications. InterpreTabNet provides a concise set of important features, making it easier for users to understand the underlying reasons for predictions, trust the model's outputs, and explain these outcomes to stakeholders.[3]

4.3. Capturing Feature Interdependencies by Prompting LLMs

We have generated interpretable feature masks where users can determine the important features. However, one issue is that our approach does not grasp the *interrelationships among features* required to explain model predictions in complicated scenarios. Therefore, we leveraged an LLM such as GPT-4 to *incorporate extensive linguistic priors into the interpretation process* that help mitigate the issues.

Table 3: Prompt Structure Design

Section	Description
Dataset Description	The Adult Census Income dataset is considered...
Mask Description	At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 1, 5, and 7...
In-Context Example 1	The Poker Hand dataset is considered...
In-Context Output 1	Output: {"Mask 0": "Initially, the rank of card 2 is recognized...}
In-Context Example 2	The Forest Cover Type dataset is considered...
In-Context Output 2	Output: {"Mask 0": "The initial feature selection identifies...}
GPT-4 Output	*{"Mask 0": "This suggests that early in the feature selection process, the model finds that work-related...}*

Interpreting Feature Masks with GPT-4. To generate a precise output mapping, we provided instructions to GPT-4 where the extracted salient features are formatted into a dictionary. Each mask corresponds to an individual analysis, followed by an aggregate analysis of all masks. Furthermore, a statement to ensure that GPT-4 produced no other natural language generation was added to maintain a consistent output map.

Finally, GPT-4 was provided with in-context examples to enable prompt tuning through few-shot learning. This was conducted via 3-fold cross-validation where datasets D1 and D2 were used as part of the prompt for tuning on D3,

[3]Link to survey: https://forms.gle/87PDZo56RUtHqFSb9

D2, and D3 as part of the prompt for tuning on D1, and so on. Only a 3-fold CV was conducted since increasing the subsets will decrease GPT-4's performance as it was unable to process extremely long sequences of texts.

Overall, GPT-4 improves the analysis of salient features extracted from InterpreTabNet by explaining their interdependencies. The structure of the designed prompt can be found in Table 3. The full prompts and outputs can be found in Appendix C.2 and C.3 respectively.

Human Evaluation Survey on GPT-4's Analysis. We conducted another survey on GPT-4's analysis in the same format as our survey on interpretability. The survey asks: *"Which model provides a more relevant and reasonable interpretation?"* when comparing between InterpreTabNet and TabNet. In Table 4, 60% of the respondents prefer InterpreTabnet as the model with a more interpretable output from GPT-4. The clear-cut features allow GPT-4 to easily comprehend the information, providing a clear and concise output that explains the outcomes to the stakeholders. The prompt and GPT-4 output for TabNet can be found in Appendix C.4.[4]

Table 4: Survey on LLM-Generated Interpretations from InterpreTabNet vs. TabNet

Model	Vote Percentage	Number of Votes
InterpreTabNet	**60%**	**12**
TabNet	40%	8
Total	100%	20

Though the results of our surveys are promising, future work can replicate these findings on a larger sample size.

Justifying GPT-4's Interpretation Capabilities. A potential concern is whether GPT-4 actually interprets the model's internal behavior rather than merely rephrasing the prompt input. To address this, we conduct the following experiments to demonstrate that the model exhibits a genuine understanding of the data.

We test the *integrity* of our prompt with a definition check on "feature mask". Figure 16 in Appendix E.8 aligns with our expectations, providing an accurate and detailed explanation. Next, we verify the *reliability* of GPT-4's analysis by prompting it to interpret synthetic datasets generated in the manner from (Yoon et al., 2019). The results in Appendix E.4 indicate that the analysis is indeed robust since it does not show any signs of hallucinations or mistakes hence, verifying the integrity of GPT-4's ability to interpret feature masks.

[4] Link to survey: https://forms.gle/ZbGeXNF1HcSEYSNP7

Additionally, we try *prompt diversification* to strengthen the trust in the generated explanations while determining which prompt design yields the most insightful and accurate explanations from GPT-4. We explore explanations in different formats and varying levels of detail. The results can be found in Tables 10 and 11 of Appendix E.5. Our analysis indicates that our original prompt structure leads to a higher level of detail and attempts to deduce deeper meanings from the prominent features, as opposed to simply categorizing them. This suggests a level of interpretive understanding by GPT-4 that went beyond basic rephrasing, thereby reinforcing the effectiveness of our prompt design.

To determine whether *GPT-4's interpretations correlate with human interpretations*, we run an experiment to identify the variability in interpretations based on what an expert in the domain might care about. We summarize the variation of results in Appendix E.6 — what we find is that the expert interpretations do not vary significantly compared to the original interpretations given different "expert" prompts. Therefore, we can anticipate that the interpretations are "robust".

Lastly, we test if *GPT-4 alone without InterpreTabNet* could identify the salient features and elucidate their relationships from the Adult dataset. However, as observed in Table 14 Appendix E.7, GPT-4 is unable to determine the salient features even when the dataset information is provided. For the first prompt, it is unable to compile any aggregate analysis. In the second prompt, the extracted important features are not accurate as well as being dense, selecting more than 50% (8/14) of the features.

5. Conclusion

We propose an interpretable variant of the TabNet neural network that is as expressive in learning the distributions of tabular data while enabling an enhanced level of interpretability. This model is designed by blending a Gumbel-Softmax distribution with a KL divergence sparsity regularizer between the attention-based feature masks to create a sparse and semantically meaningful decomposition of the predictive signals. Relative to our baselines, our model outputs more interpretable feature masks to determine salient features while maintaining its competitive accuracy across most datasets. The salient features from our masks are channeled into GPT-4 via prompts that encourage a careful analysis of the features' interdependencies. For practitioners, InterpreTabNet distills the predictive signals allowing it to stand as a practical toolkit for understanding where tabular data comes from. It bridges the often challenging gap between intricate machine learning outputs and real-world decision-making, ensuring that insights are not just extracted but also intuitively understood and readily actionable.

Impact Statement

This paper introduces InterpreTabNet to improve the interpretability of machine learning models that handle tabular data. Its primary societal impact lies in offering more transparent, understandable deep-learning predictions and decisions. This is crucial in high-stakes human-oriented sectors such as healthcare and finance, where decision-making impacts human lives. Ethically, InterpreTabNet represents a step towards responsible AI, as it allows users to understand how and why specific decisions are made, enhancing trust and reducing the 'black box' nature of complex models. However, there are potential risks. An example could be an over-reliance on model interpretations, leading to neglecting other important factors not captured by the model.

Acknowledgements

We thank the University of Toronto for supporting this research. MC is supported by a Vector Institute Graduate Student Fellowship and a CIHR Health Systems Impact Fellowship. RGK is supported by a Tier II Canada Research Chair and Canada CIFAR AI Chair.

References

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

B. Becker and R. Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

BlastChar. Telco customer churn, Feb 2018. URL https://www.kaggle.com/datasets/blastchar/telco-customer-churn.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

R. Cattral and F. Oppacher. Poker Hand. UCI Machine Learning Repository, 2007. DOI: https://doi.org/10.24432/C5KW38.

J. Chen, L. Song, M. J. Wainwright, and M. I. Jordan. Learning to explain: An information-theoretic perspective on model interpretation, 2018.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

K. Choi, C. Cundy, S. Srivastava, and S. Ermon. Lmpriors: Pre-trained language models as task-specific priors. *arXiv preprint arXiv:2210.12530*, 2022.

J. Clore, K. Cios, J. DeShazo, and B. Strack. Diabetes 130-US hospitals for years 1999-2008. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5230J.

A. Datta. Us health insurance dataset, Feb 2020. URL https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset.

M. Du, N. Liu, and X. Hu. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1): 68–77, 2019.

D. Dua and C. Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

L. Fang, T. Zeng, C. Liu, L. Bo, W. Dong, and C. Chen. Transformer-based conditional variational autoencoder for controllable story generation, Jul 2021. URL https://arxiv.org/abs/2101.00828v2.

C. F. Gauss. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*, volume 7. FA Perthes, 1877.

P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Mach. Learn.*, 63(1):3–42, apr 2006. ISSN 0885-6125. doi: 10.1007/s10994-006-6226-1. URL https://doi.org/10.1007/s10994-006-6226-1.

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.

S. Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.

S. Jain and B. C. Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax, 2016. URL https://arxiv.org/pdf/1611.01144.pdf.

W. Joo, W. Lee, S. Park, and I.-C. Moon. Dirichlet variational autoencoder, Jan 2019. URL https://arxiv.org/abs/1901.02739v1.

L. Katzir, G. Elidan, and R. El-Yaniv. Net-dnf: Effective deep modeling of tabular data. In *International conference on learning representations*, 2020.

G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

M. Kelly, R. Longjohn, and K. Nottingham. The UCI machine learning repository, 2023. URL https://archive.ics.uci.edu.

D. P. Kingma and M. Welling. Auto-encoding variational bayes, 2022.

A. Kristiadi. Conditional variational autoencoder: Intuition and implementation, 2016. URL https://agustinus.kristia.de/techblog/2016/12/17/conditional-vae/.

S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86, 1951.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

I. Madakkatel and E. Hyppönen. Llpowershap: Logistic loss-based automated shapley values feature selection method, 2024.

S. Moro, P. Rita, and P. Cortez. Bank Marketing. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C5K306.

OpenAI. Gpt-4 technical report, 2023.

U. M. L. Repository. Mushroom. UCI Machine Learning Repository, 1987. DOI: https://doi.org/10.24432/C5959T.

M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. URL http://arxiv.org/abs/1602.04938.

C. V. Roberts, E. Elahi, and A. Chandrashekar. On the bias-variance characteristics of lime and shap in high sparsity movie recommendation explanation tasks. *arXiv preprint arXiv:2206.04784*, 2022.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/8d55a249e6baa5c06772297520da2051-Paper.pdf.

T. Ucar, E. Hajiramezanali, and L. Edwards. Subtab: Subsetting features of tabular data for self-supervised representation learning. *Advances in Neural Information Processing Systems*, 34:18853–18865, 2021.

J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo. Openml: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013. doi: 10.1145/2641190.2641198. URL http://doi.acm.org/10.1145/2641190.2641198.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

D. Whiteson. HIGGS. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5V312.

P. Xie, J. Zhu, and E. Xing. Diversity-promoting bayesian learning of latent variable models. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 59–68, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/xiea16.html.

P. Xie, J. Zhu, and E. P. Xing. Diversity-promoting bayesian learning of latent variable models, Nov 2017. URL https://arxiv.org/abs/1711.08770.

J. Yoon, J. Jordon, and M. van der Schaar. INVASE: Instance-wise variable selection using neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJg_roAcK7.

Appendix

A. Proofs and Algorithms

A.1. Proof: cVAE Evidence Lower Bound

$$
\begin{aligned}
D_{KL}[Q(z|Y,X)||P(z|Y,X)] &= \sum_z Q(z|Y,X)\log\frac{Q(z|Y,X)}{P(z|Y,X)} \\
&= E[\log\frac{Q(z|Y,X)}{P(z|Y,X)}] \\
&= E[\log Q(z|Y,X) - \log P(z|Y,X)] \\
&\text{using Bayes' rule} \\
&= E[\log Q(z|Y,X) - \log\frac{P(z,Y,X)}{P(Y,X)}] \\
&= E[\log Q(z|Y,X) - \log\frac{P(Y|z,X)P(z,X)}{P(Y,X)}] \\
&= E[\log Q(z|Y,X) - \log\frac{P(Y|z,X)P(z|X)P(X)}{P(Y,X)}] \\
&= E[\log Q(z|Y,X) - \log\frac{P(Y|z,X)P(z|X)P(X)}{P(Y|X)P(X)}] \\
&= E[\log Q(z|Y,X) - \log\frac{P(Y|z,X)P(z|X)}{P(Y|X)}] \\
&= E[\log Q(z|Y,X) - (\log P(Y|z,X) + \log P(z|X) - \log P(Y|X))] \\
&= E[\log Q(z|Y,X) - \log P(Y|z,X) - \log P(z|X) + \log P(Y|X)] \\
&= E[\log Q(z|Y,X) - \log P(Y|z,X) - \log P(z|X)] + \log P(Y|X) \\
D_{KL}[Q(z|Y,X)||P(z|Y,X)] - \log P(Y|X) &= E[\log Q(z|Y,X) - \log P(Y|z,X) - \log P(z|X)] \\
&\text{rearranging the sign to rewrite RHS as another KL Divergence} \\
\log P(Y|X) - D_{KL}[Q(z|Y,X)||P(z|Y,X)] &= E[\log P(Y|z,X) - (\log Q(z|Y,X) - \log P(z|X))] \\
\log P(Y|X) - D_{KL}[Q(z|Y,X)||P(z|Y,X)] &= E[\log P(Y|z,X)] - E[\log Q(z|Y,X) - \log P(z|X)] \\
\log P(Y|X) - D_{KL}[Q(z|Y,X)||P(z|Y,X)] &= E[\log P(Y|z,X)] - E[\log\frac{Q(z|Y,X)}{P(z|X)}] \\
\log P(Y|X) - D_{KL}[Q(z|Y,X)||P(z|Y,X)] &= E[\log P(Y|z,X)] - D_{KL}[Q(z|Y,X)||P(z|X)] \\
&\text{which is the cVAE objective function}
\end{aligned}
$$

A.2. Algorithm: KL Divergence Sparsity Regularizer r_M

The algorithm analyzes the model's feature importance masks to validate that they meet a set criterion. This criterion is to validate that the masks are sparse and that the features the model selects are important. Upon fulfilling the criterion a specific number of times, the algorithm terminates. To increase efficiency, the algorithm also employs a recursive search to narrow down the value range around the current best r_M, thereby reducing computational overhead.

Algorithm 1 Our proposed algorithm for interpretability optimization. Good default settings for the tested machine learning problems are $\alpha = 0$, $\beta = [0, 10000000]$, $\delta = [0.20, 0.25]$, $\gamma = [2, 3]$ $\epsilon = [3, 5]$. For β, δ and γ, it would depend on the nature of the dataset. More samples require higher parameter values.

Require: α: Starting range (start)
Require: β: Ending range (end)
Require: δ: Percentage of feature importance captured by one feature in each feature mask (col_threshold_val)
Require: γ: Number of columns that satisfies δ in each feature mask (col_threshold)
Require: ι: Number of complete-feature masks that passes the algorithm's feature selection criteria (all_mask_pass)
Require: ϵ: Threshold for the number of complete-feature masks that passes the algorithm's feature selection criteria (all_mask_pass_thresh)
Require: ζ: Step size computed using a logarithmic scale at high levels (step_size)
Require: θ: Dictionary storing r_M-accuracy pairs (reg_m_acc_dict)
Require: λ: Flag for recursion (is_recursive)
Ensure: Optimal regularization parameter r_M^*

Initialize θ if θ is None.
Initialize ι if ι is None.
if $\iota = \epsilon$ **then**
    $r_M^* = \arg\max(\theta)$
    return r_M^*
end if
while $\alpha \leq \beta$ and $\iota < \epsilon$ **do**
    Train TabNet, Compute Accuracy and Generate Masks
    ▷ Inner loop evaluating each feature mask here.
    if Criteria for updating θ and ι are met **then**
        Update θ, ι
    end if
    if λ **then**
        $\alpha = \alpha + \zeta$
    else if $\alpha = 0$ **then**
        $\alpha = 10$
    else
        $\alpha * = 10$
    end if
end while
if r_M^* is Not None & Length of $\theta = 1$ **then**
    Recurse with updated boundaries.
else
    $r_M^* = \arg\max(\theta)$
    return r_M^*
end if

B. Experimental Setup and Datasets

B.1. Reproducibility

Code Release. The code for InterpreTabNet and files to reproduce the experiments are available on GitHub at: `https://github.com/jacobyhsi/InterpreTabNet`.

Availability of Datasets. The datasets used in this paper are all freely accessible on OpenML. `OpenML.org` and `UCI Machine Learning Repository`. Download links and additional statistical details about the datasets can be found in Appendix B.2 of the paper.

GPT-4 Version. The GPT-4 version used in our experiments is "gpt-4-1106" with training data up to Apr 2023 and a context window of 128,000 tokens.

B.2. Additional Dataset Information

We evaluated our model on 7 datasets. These datasets contain 4 binary classification tasks and 3 multi-class classification tasks. We provided statistical details in Table 5, and download links in Table 6. In each of our datasets, we applied label encoding to the categorical features to transform textual values into numerical representations. Additionally, we introduced a distinct token to handle missing data within these categorical columns. This uniform preprocessing approach was applied consistently across all datasets, ensuring compatibility and reliability for subsequent machine learning analyses.

Table 5: Datasets used for evaluation

Dataset	Task	# Features	# Categorical	# Instances	# Classes	# NaNs
Adult Census Income	Binary	14	8	32,560	2	0
Forest Cover Type	Multi-Class	54	44	581,012	7	0
Poker Hand	Multi-Class	10	10	1,025,010	10	0
Mushroom	Binary	22	22	8,124	2	0
Blastchar	Binary	20	17	7,043	2	0
Diabetes	Multi-Class	49	39	101,766	3	0
Higgs	Binary	28	0	11,000,000	2	0

Table 6: Dataset Links

Dataset Name	Dataset Link
Adult Census Income	`https://archive.ics.uci.edu/dataset/2/adult`
Forest Cover Type	`https://archive.ics.uci.edu/dataset/31/covertype`
Poker Hand	`https://archive.ics.uci.edu/dataset/158/poker+hand`
Mushroom	`https://archive.ics.uci.edu/dataset/73/mushroom`
Blastchar	`https://www.kaggle.com/datasets/blastchar/telco-customer-churn`
Diabetes	`https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008`
Higgs	`https://archive.ics.uci.edu/dataset/280/higgs`

B.3. Hyperparameters

GUIDELINES

Hyperparameters such as $N_d = N_a$, N_{steps}, γ, and learning rate are tuned in the range per TabNet's recommendations. In terms of the sparsity regularizer for InterpreTabNet, r_M, we recommend a smaller range e.g. [0,10000] for datasets with a low to moderate number of features and samples (Adult dataset), and a larger range e.g. [0, 1,000,000,000,000] for datasets with a larger number of features and samples (Higgs dataset). Within the sparsity algorithm itself (Appendix A.2), parameters such as the number of salient features and feature importance threshold can be adjusted to the user's preference.

SEARCH SPACE

We provided hyperparameter search spaces for all models in Table 7. For TabTransformer, we used the same hyperparameter space mentioned in their paper (Huang et al., 2020). XGboost and LightGBM were designed from scratch and used common hyperparameter choices with suggestions from the official documentation (Chen and Guestrin, 2016) (Ke et al., 2017). For MLP, we followed the exact hyperparameter search space as (Huang et al., 2020).

Table 7: Hyperparameter spaces for all models

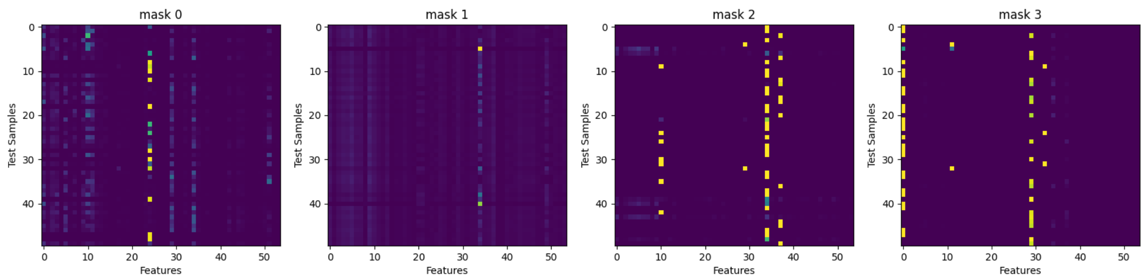
Model	**Hyperparameter Space**
InterpreTabNet	$N_d = N_a$ (output dimension): [16, 32, 128], N_{steps}: [3, 4, 5], γ: [1.0, 1.2, 1.5, 2.0], Learning Rate: [0.005, 0.01, 0.02, 0.025], r_M: range from [0, 1,000,000,000,000]
Original TabNet	$N_d = N_a$ (output dimension): [16, 32, 128], N_{steps}: [3, 4, 5], γ: [1.0, 1.2, 1.5, 2.0], Learning Rate: [0.005, 0.01, 0.02, 0.025]
TabTransformer	Hidden Dimension: [32, 54, 128, 256], Number of Layers: [1, 2, 3, 6, 12], Number of Attention Heads: [2, 4, 8], MLP First Hidden Layer: $x = m \times l, m \in \mathbb{Z} \| 1 \leq m \leq 8$, where l is the input size, MLP Second Hidden Layer: $x = m \times l, m \in \mathbb{Z} \| 1 \leq m \leq 3$, where l is the input size
XGBoost	learning_rate: [0.01, 0.1, 0.2], max_depth: [3, 4, 5, 6], n_estimators: [50, 100, 200], subsample: [0.8, 0.9], colsample_bytree: [0.8, 0.9], min_child_weight: [1, 2, 3]
LightGBM	num_leaves: [20, 30, 40], learning_rate: [0.05, 0.1, 0.2], n_estimators: [100, 200], subsample: [0.8, 0.9], colsample_bytree: [0.8, 0.9]
MLP	First Hidden Layer: $x = m \times l, m \in \mathbb{Z} \| 1 \leq m \leq 8$, where l is the input size, Second Hidden Layer: $x = m \times l, m \in \mathbb{Z} \| 1 \leq m \leq 3$, where l is the input size

C. Additional Experimental Results from Other Datasets

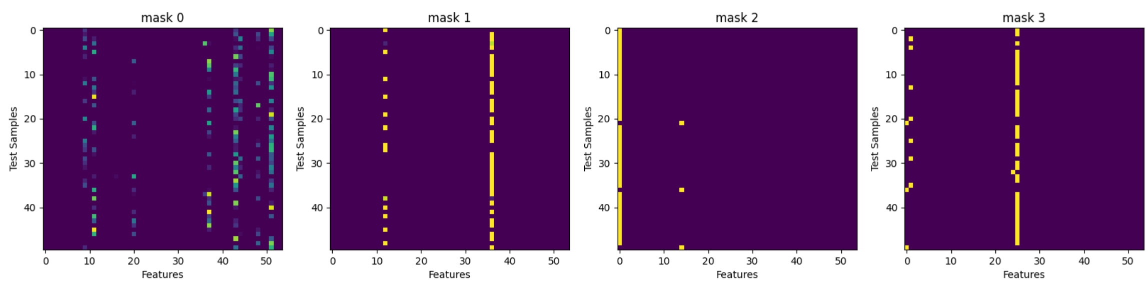
C.1. Accuracies & Masks

FOREST COVER TYPE (DUA AND GRAFF, 2017)

Model	**Test Accuracy (%)**
InterpreTabNet ($r_M^* = 900$)	94.75 ± 0.53
Original TabNet	94.18 ± 0.63
XGBoost	92.30 ± 0.62
LightGBM	86.38 ± 0.64
TabTransformer	82.55 ± 0.39
MLP	84.89 ± 0.56



(a) InterpreTabNet

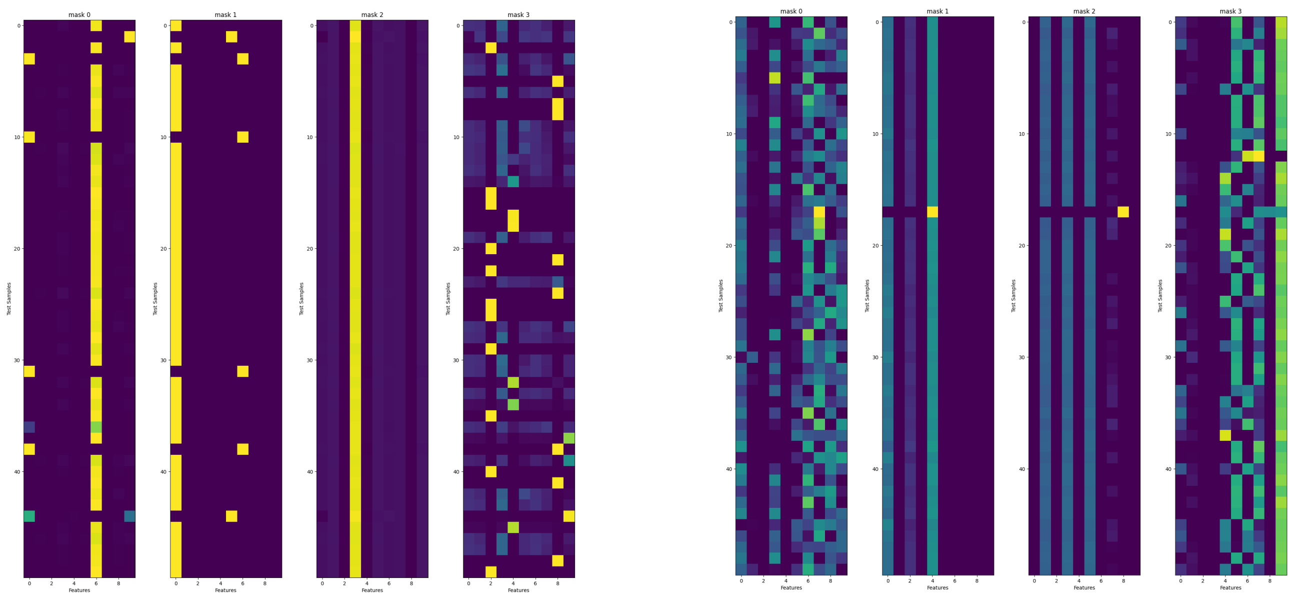


(b) Original TabNet Model

Figure 6: (a) Best performing model using InterpreTabNet $r_M^* = 900$ with an accuracy of 94.75% on the Forest Cover Type Dataset. (b) The baseline performance using the Original TabNet model, attaining an accuracy of 94.18%.

POKER HAND (CATTRAL AND OPPACHER, 2007)

Model	**Test Accuracy (%)**
InterpreTabNet ($r_M^* = 1000$)	99.50 ± 0.48
Original TabNet	99.00 ± 0.62
XGBoost	75.57 ± 0.47
LightGBM	78.47 ± 0.36
TabTransformer	99.81 ± 0.31
MLP	99.70 ± 0.56

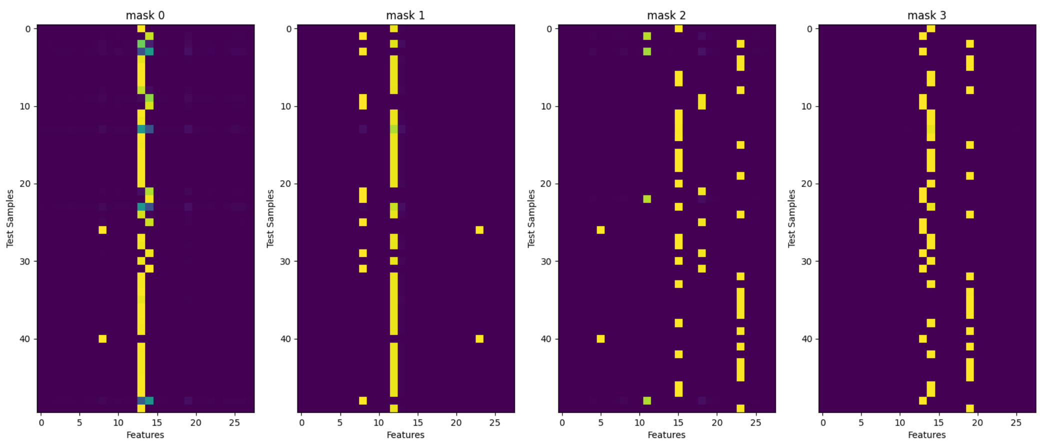


(a) InterpreTabNet (b) Original TabNet Model

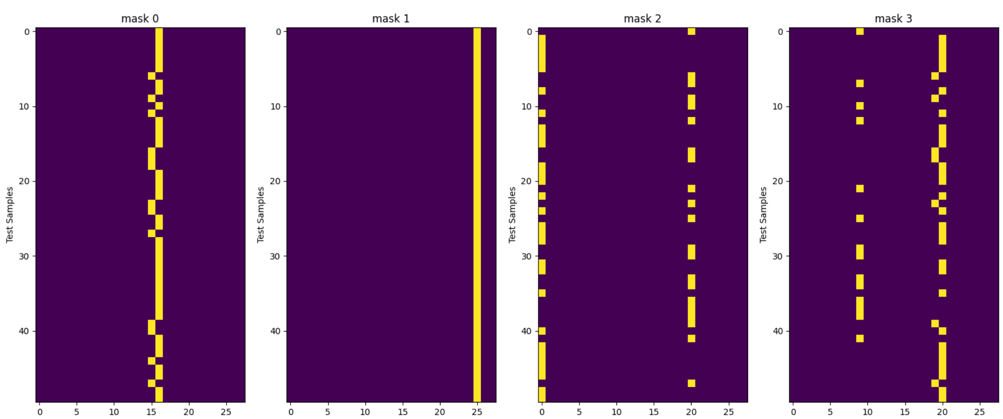
Figure 7: (a) Best performing model using InterpreTabNet $r_M^* = 1000$ with an accuracy of 99.13% on the Poker Hand Dataset. (b) The baseline performance using the Original TabNet model, attaining an accuracy of 99.23%.

HIGGS (WHITESON, 2014)

Model	**Test Accuracy (%)**
InterpreTabNet ($r_M^* = 10000$)	53.08 ± 0.56
Original TabNet	52.94 ± 0.45
XGBoost	72.70 ± 0.35
LightGBM	72.62 ± 0.35
TabTransformer	51.97 ± 0.54
MLP	63.17 ± 0.36



(a) InterpreTabNet

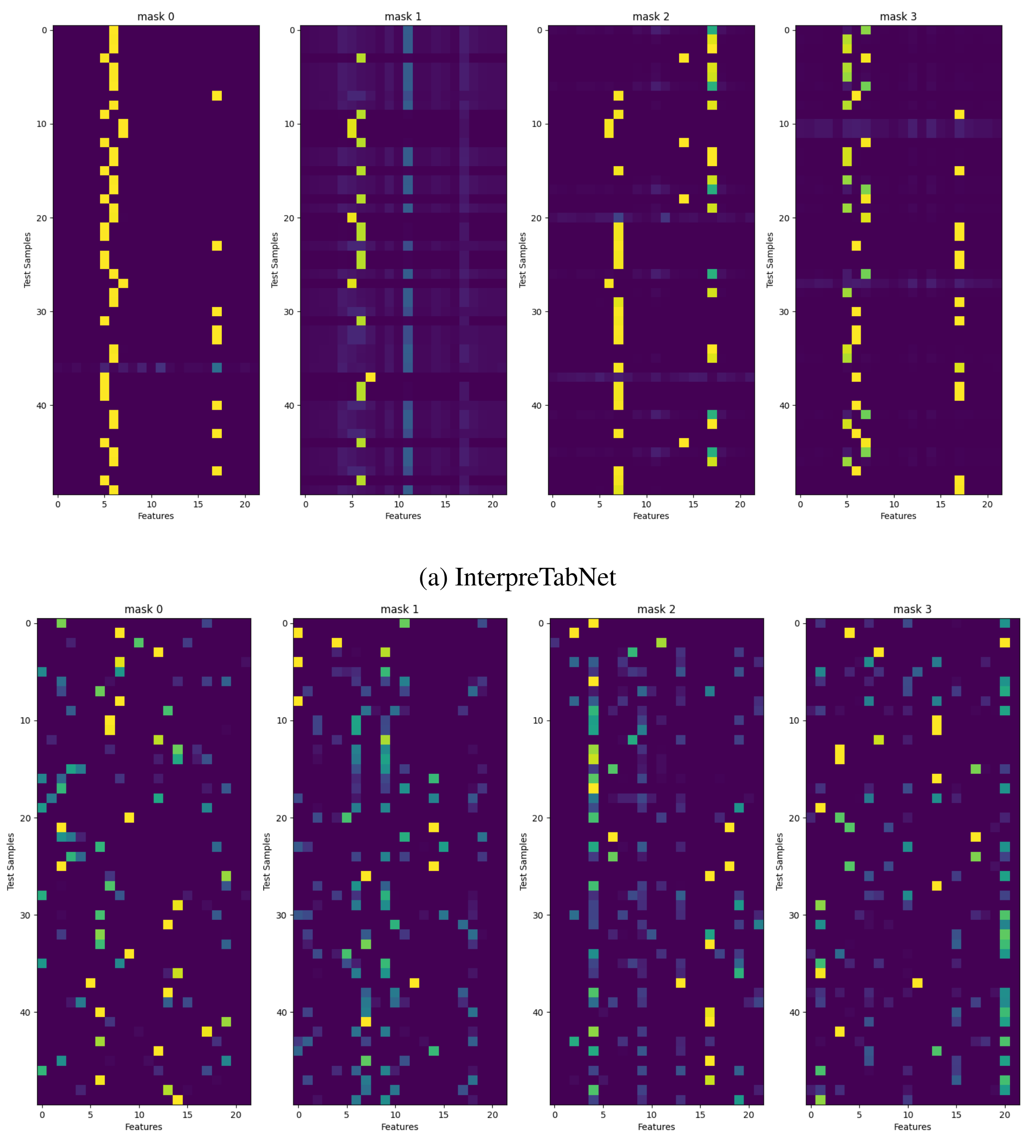


(b) Original TabNet Model

Figure 8: (a) Best performing model using InterpreTabNet $r_M^* = 10000$ with an accuracy of 53.08% on the Higgs Dataset. (b) The baseline performance using the Original TabNet model, attaining an accuracy of 60.22%.

MUSHROOM (REPOSITORY, 1987)

Model	**Test Accuracy (%)**
InterpreTabNet ($r_M^* = 10,000,000,000,000$)	96.62 ± 0.35
Original TabNet	99.94 ± 0.31
XGBoost	99.69 ± 0.39
LightGBM	100.00 ± 0.37
TabTransformer	100.00 ± 0.57
MLP	99.82 ± 0.56



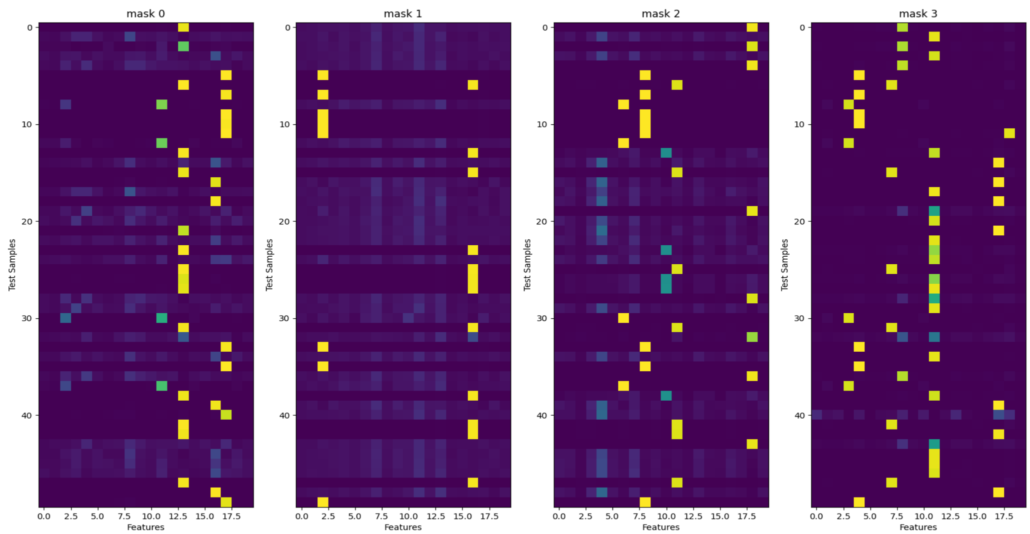
(a) InterpreTabNet

(b) Original TabNet Model

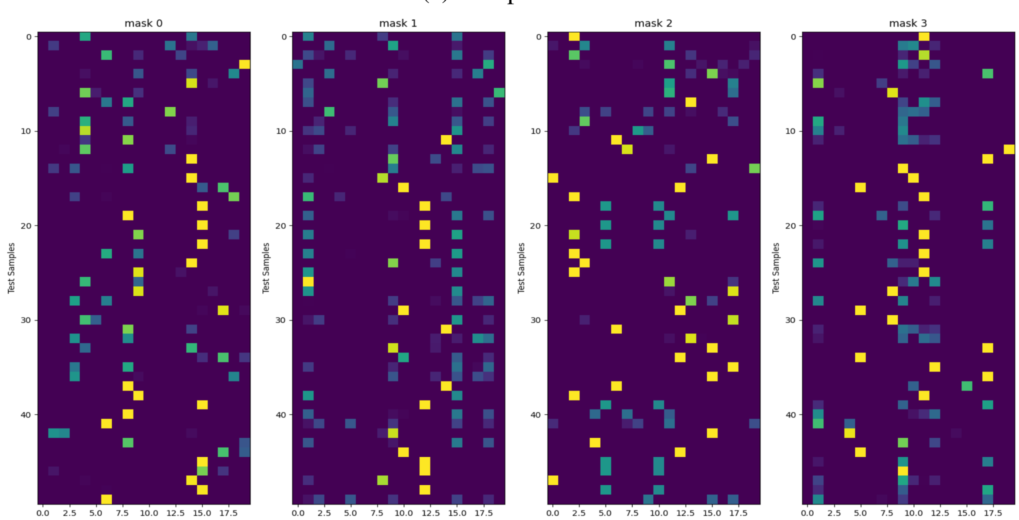
Figure 9: (a) Best performing model using InterpreTabNet $r_M^* = 1,000,000,000,0000$ with an accuracy of 96.62% on the Mushroom Dataset. (b) The baseline performance using the Original TabNet model, attaining an accuracy of 99.94%.

BLASTCHAR (BLASTCHAR, 2018)

Model	Test Accuracy (%)
InterpreTabNet ($r_M^* = 10,000,000,000,000$)	72.96 ± 0.56
Original TabNet	76.22 ± 0.34
XGBoost	77.29 ± 0.53
LightGBM	77.86 ± 0.46
TabTransformer	73.17 ± 0.56
MLP	75.16 ± 0.61



(a) InterpreTabNet

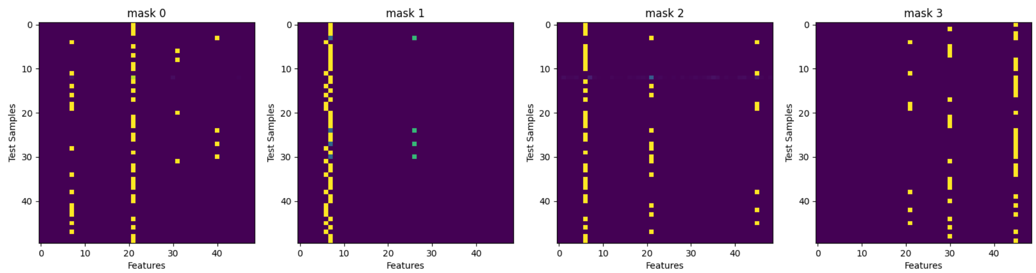


(b) Original TabNet Model

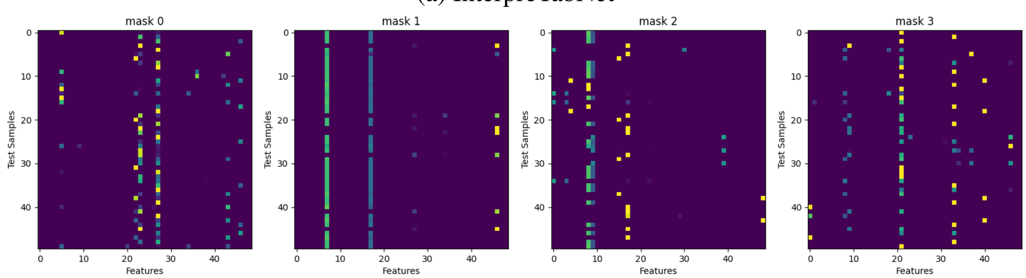
Figure 10: (a) Best performing model using InterpreTabNet $r_M^* = 10,000,000,000,000$ with an accuracy of 72.96% on the Blastchar Dataset. (b) The baseline performance using the Original TabNet model, attaining an accuracy of 76.22%.

DIABETES (CLORE ET AL., 2014)

Model	Test Accuracy (%)
InterpreTabNet ($r_M^* = 100,000,000$)	55.37 ± 0.47
Original TabNet	56.91 ± 0.53
XGBoost	61.44 ± 0.32
LightGBM	60.87 ± 0.39
TabTransformer	44.45 ± 0.34
MLP	53.99 ± 0.46



(a) InterpreTabNet



(b) Original TabNet Model

Figure 11: (a) Best performing model using InterpreTabNet $r_M^* = 100,000,000$ with an accuracy of 55.38% on the Diabetes Dataset. (b) The baseline performance using the Original TabNet model, attaining an accuracy of 56.91%.

C.2. InterpreTabNet: Prompts for All Datasets

ADULT CENSUS INCOME

Conduct aggregate analysis on the description of the following feature mask. Please output ONLY a dictionary and no other natural language generation when generating the sentence as shown in the in-context example below. Please use single-word classification that encapsulates the meaning of the features if possible.

The Adult Census Income dataset is considered which predicts whether an individual's income exceeds $50,000 per year. There are 50 test samples and 14 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 1, 5 and 7 which are workclass, marital status and relationship. At the 1st step of feature selection, we observe mask 1 with the main features highlighted as 1 and 5 which are workclass and marital status. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 8, 10 and 12 which are race, capital gain and hours-per-week. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 1, 7 and 10 which are workclass, relationship and capital-gain.

Here are in-context examples for few-shot learning.

The Poker Hand dataset is considered, which is a classification of poker hands into 10 distinct types based on attributes like rank of individual cards, suits, and card combinations. There are 50 test samples and 10 features.

At the 0th step of feature selection, we observe mask 0 with the main feature highlighted as 3 which is the rank of the card 2. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 7 which is the rank of card 3. At the 2nd step of feature selection, we observe mask 2 with the main feature highlighted as 6 which is the suit of card 3. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 3, 6, 7 and 10 which are the rank of cards 1 and 3 as well as the suits of cards 3 and 5.

Output: {"Mask 0": "Initially, the rank of card 2 is recognized as a significant predictor in the classification of poker hands. The rank of individual cards seems crucial at this stage.", "Mask 1": "At this step, the rank of card 3 is highlighted, further emphasizing the importance of individual card ranks in determining poker hand types.", "Mask 2": "Here, besides the rank of card 3, the suit of the same card is considered, suggesting that both rank and suit play roles in classifying poker hands.", "Mask 3": "This step combines multiple features, both ranks (cards 1 and 3) and suits (cards 3 and 5), indicating that a combination of these factors provides a broader understanding of poker hand classifications.", "Aggregate": "The rank of cards, specifically card 2 and card 3, appears consistently across different masks, indicating their primary role in poker hand classification. Suits, particularly of card 3, also have importance but are introduced later in the feature selection. The combined importance of both ranks and suits in the final step suggests that while individual card ranks are pivotal in initial classifications, understanding the relationship between card ranks and their respective suits provides a more comprehensive insight into the poker hand types. The consistent presence of card 3's attributes (both rank and suit) underscores its pivotal role in determining poker hand categories."}

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

Output: {"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape's nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn't be overlooked."}

FOREST COVER TYPE

Conduct aggregate analysis on the description of the following feature mask. Please output ONLY a dictionary and no other natural language generation when generating the sentence as shown in the in-context example below. Please use single-word classification that encapsulates the meaning of the features if possible.

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

Here are in-context examples for few-shot learning.

The Adult Census Income dataset is considered which predicts whether an individual's income exceeds $50,000 per year. There are 50 test samples and 14 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 1, 5 and 7 which are workclass, marital status and relationship. At the 1st step of feature selection, we observe mask 0 with the main features highlighted as 1 and 5 which are workclass and marital status. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 8, 10 and 12 which are race, capital gain and hours-per-week. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 1, 7 and 10 which are workclass, relationship and capital-gain.

Output: {"Mask 0": "This suggests that early in the feature selection process, the model finds that work-related factors (like the type of work) and personal relationships (both marital status and other relationships) are significant predictors for income exceeding $50,000.", "Mask 1": "Here, the model emphasizes even more on work nature and marital status, suggesting these two as the primary drivers for the prediction. Relationship as a feature has been excluded in this step, which might indicate a possible correlation with marital status.", "Mask 2": "At this step, the model seems to be looking beyond just personal and work nature to more financial and demographic aspects. Capital gain, which is directly related to financial health, is considered along with hours-per-week, potentially pointing to a relationship between longer working hours and higher income. The inclusion of race might indicate underlying societal or economic factors at play.", "Mask 3": "This combination brings back features from both the initial and second masks. Workclass and relationship are revisited, which further underscores their importance. Capital gain remains consistent from the second step, emphasizing its importance as a predictor for income.", "Aggregate": "The features that consistently appear across multiple masks are: workclass (masks 0, 1, and 3), relationship (masks 0 and 3), and capital gain (masks 2 and 3). This highlights that these features play a pivotal role in predicting whether an individual's income exceeds $50,000 per year. Marital status is seen in the initial stages but disappears later, while hours-per-week and race come into play at the second step. This could indicate that while some features provide foundational predictive power, others may provide more nuanced insight or refinement to the model's predictions. Work-related aspects, personal relationships, and financial metrics seem to be the main predictors of an individual's income level in this dataset. It would be crucial to further analyze these features' exact impacts and understand any potential correlations among them for a more comprehensive understanding."}

The Poker Hand dataset is considered, which is a classification of poker hands into 10 distinct types based on attributes like rank of individual cards, suits, and card combinations. There are 50 test samples and 10 features.

At the 0th step of feature selection, we observe mask 0 with the main feature highlighted as 3 which is the rank of the card 2. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 7 which is the rank of card 3. At the 2nd step of feature selection, we observe mask 2 with the main feature highlighted as 6 which is the suit of card 3.

Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 3, 6, 7 and 10 which are the rank of cards 1 and 3 as well as the suits of cards 3 and 5.

Output: {"Mask 0": "Initially, the rank of card 2 is recognized as a significant predictor in the classification of poker hands. The rank of individual cards seems crucial at this stage.", "Mask 1": "At this step, the rank of card 3 is highlighted, further emphasizing the importance of individual card ranks in determining poker hand types.", "Mask 2": "Here, besides the rank of card 3, the suit of the same card is considered, suggesting that both rank and suit play roles in classifying poker hands.", "Mask 3": "This step combines multiple features, both ranks (cards 1 and 3) and suits (cards 3 and 5), indicating that a combination of these factors provides a broader understanding of poker hand classifications.", "Aggregate": "The rank of cards, specifically card 2 and card 3, appears consistently across different masks, indicating their primary role in poker hand classification. Suits, particularly of card 3, also have importance but are introduced later in the feature selection. The combined importance of both ranks and suits in the final step suggests that while individual card ranks are pivotal in initial classifications, understanding the relationship between card ranks and their respective suits provides a more comprehensive insight into the poker hand types. The consistent presence of card 3's attributes (both rank and suit) underscores its pivotal role in determining poker hand categories."}

Poker Hand Prompt

Conduct aggregate analysis on the description of the following feature mask. Please output ONLY a dictionary and no other natural language generation when generating the sentence as shown in the in-context example below. Please use single-word classification that encapsulates the meaning of the features if possible.

The Poker Hand dataset is considered, which is a classification of poker hands into 10 distinct types based on attributes like rank of individual cards, suits, and card combinations. There are 50 test samples and 10 features.

At the 0th step of feature selection, we observe mask 0 with the main feature highlighted as 3 which is the rank of the card 2. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 7 which is the rank of card 3. At the 2nd step of feature selection, we observe mask 2 with the main feature highlighted as 6 which is the suit of card 3. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 3, 6, 7 and 10 which are the rank of cards 1 and 3 as well as the suits of cards 3 and 5.

Here are in-context examples for few-shot learning.

The Adult Census Income dataset is considered which predicts whether an individual's income exceeds $50,000 per year. There are 50 test samples and 14 features. At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 1, 5 and 7 which are workclass, marital status and relationship. At the 1st step of feature selection, we observe mask 0 with the main features highlighted as 1 and 5 which are workclass and marital status. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 8, 10 and 12 which are race, capital gain and hours-per-week. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 1, 7 and 10 which are workclass, relationship and capital-gain.

Output: {"Mask 0": "This suggests that early in the feature selection process, the model finds that work-related factors (like the type of work) and personal relationships (both marital status and other relationships) are significant predictors for income exceeding $50,000.", "Mask 1": "Here, the model emphasizes even more on work nature and marital status, suggesting these two as the primary drivers for the prediction. Relationship as a feature has been excluded in this step, which might indicate a possible correlation with marital status.", "Mask 2": "At this step, the model seems to be looking beyond just personal and work nature to more financial and demographic aspects. Capital gain, which is directly related to financial health, is considered along with hours-per-week, potentially pointing to a relationship between longer working hours and higher income. The inclusion of race might indicate underlying societal or economic factors at play.", "Mask 3": "This combination brings back features from both the initial and second masks. Workclass and relationship are revisited, which further underscores their importance. Capital gain remains consistent from the second step, emphasizing its importance as a predictor for income.", "Aggregate": "The features that consistently appear across multiple masks are: workclass (masks 0, 1, and 3), relationship (masks 0 and 3), and capital gain (masks 2 and 3). This highlights that these features play a pivotal role in predicting whether an individual's income exceeds $50,000 per year. Marital status is seen in the initial stages but disappears later, while hours-per-week and race come into play at the second step. This could indicate that while some features provide foundational predictive power, others may provide more nuanced insight or refinement to the model's predictions. Work-related aspects, personal relationships, and financial metrics seem to be the main predictors of an individual’s income level in this dataset. It would be crucial to further analyze these features’ exact impacts and understand any potential correlations among them for a more comprehensive understanding."}

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

Output: {"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape’s nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn’t be overlooked."}

MUSHROOM PROMPT

Conduct aggregate analysis on the description of the following feature masks. Start off with an analysis of the individual masks, followed by an aggregate analysis of all masks combined. Please format the output into a dictionary as shown in the in-context examples. The output should only contain the formatted output, no other natural language generation is required.

The Mushroom dataset is considered, which is a classification of mushrooms into edible or poisonous categories based on attributes like cap shape, gill color, stalk length, and other morphological characteristics. There are 50 test samples and 22 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 5, 6 and 17 which are bruises, odor and veil-type. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 5, 6, 11 which are bruises, odor and stalk-shape. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 7, 14, 17 which are gill-attachment, stalk-surface-below-ring and veil-type. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 5, 6, 7 and 17 which are bruises, odor, gill-attachment and veil-type.

Here are in-context examples for few-shot learning.

The Adult Census Income dataset is considered which predicts whether an individual’s income exceeds $50,000 per year. There are 50 test samples and 14 features. At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 1, 5 and 7 which are workclass, marital status and relationship. At the 1st step of feature selection, we observe mask 0 with the main features highlighted as 1 and 5 which are workclass and marital status. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 8, 10 and 12 which are race, capital gain and hours-per-week. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 1, 7 and 10 which are workclass, relationship and capital-gain.

Output: {"Mask 0": "This suggests that early in the feature selection process, the model finds that work-related factors (like the type of work) and personal relationships (both marital status and other relationships) are significant predictors for income exceeding $50,000.", "Mask 1": "Here, the model emphasizes even more on work nature and marital status, suggesting these two as the primary drivers for the prediction. Relationship as a feature has been excluded in this step, which might indicate a possible correlation with marital status.", "Mask 2": "At this step, the model seems to be looking beyond just personal and work nature to more financial and demographic aspects. Capital gain, which is directly related to financial health, is considered along with hours-per-week, potentially pointing to a relationship between longer working hours and higher income. The inclusion of race might indicate underlying societal or economic factors at play.", "Mask 3": "This combination brings back features from both the initial and second masks. Workclass and relationship are revisited, which further underscores their importance. Capital gain remains consistent from the second step, emphasizing its importance as a predictor for income.", "Aggregate": "The features that consistently appear across multiple masks are: workclass (masks 0, 1, and 3), relationship (masks 0 and 3), and capital gain (masks 2 and 3). This highlights that these features play a pivotal role in predicting whether an individual's income exceeds $50,000 per year. Marital status is seen in the initial stages but disappears later, while hours-per-week and race come into play at the second step. This could indicate that while some features provide foundational predictive power, others may provide more nuanced insight or refinement to the model's predictions. Work-related aspects, personal relationships, and financial metrics seem to be the main predictors of an individual's income level in this dataset. It would be crucial to further analyze these features' exact impacts and understand any potential correlations among them for a more comprehensive understanding."}

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

Output: {"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape's nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn't be overlooked."}

BLASTCHAR PROMPT

Conduct aggregate analysis on the description of the following feature masks. Start off with an analysis of the individual masks, followed by an aggregate analysis of all masks combined. Please format the output into a dictionary as shown in the in-context examples. The output should only contain the formatted output, no other natural language generation is required.

The BlastChar Telco Customer Churn dataset is considered, which is a classification of customers into retained or churned categories based on attributes like gender, seniority, tenure, service subscriptions, contract type, billing methods, and charges, among others. There are 50 test samples and 21 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 13, 16, and 17 which are StreamingTV, PaperlessBilling and PaymentMethod. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 2 and 16 which are SeniorCitizen and PaperlessBilling. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 8, 11, and 18 which are InternetService, DeviceProtection, and MonthlyCharges. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 3, 11, and 17 which are Partner, DeviceProtection, and PaymentMethod.

Here are in-context examples for few-shot learning.

The Adult Census Income dataset is considered which predicts whether an individual's income exceeds $50,000 per year. There are 50 test samples and 14 features. At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 1, 5 and 7 which are workclass, marital status and relationship. At the 1st step of feature selection, we observe mask 0 with the main features highlighted as 1 and 5 which are workclass and marital status. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 8, 10 and 12 which are race, capital gain and hours-per-week. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 1, 7 and 10 which are workclass, relationship and capital-gain.

Output: {"Mask 0": "This suggests that early in the feature selection process, the model finds that work-related factors (like the type of work) and personal relationships (both marital status and other relationships) are significant predictors for income exceeding $50,000.", "Mask 1": "Here, the model emphasizes even more on work nature and marital status, suggesting these two as the primary drivers for the prediction. Relationship as a feature has been excluded in this step, which might indicate a possible correlation with marital status.", "Mask 2": "At this step, the model seems to be looking beyond just personal and work nature to more financial and demographic aspects. Capital gain, which is directly related to financial health, is considered along with hours-per-week, potentially pointing to a relationship between longer working hours and higher income. The inclusion of race might indicate underlying societal or economic factors at play.", "Mask 3": "This combination brings back features from both the initial and second masks. Workclass and relationship are revisited, which further underscores their importance. Capital gain remains consistent from the second step, emphasizing its importance as a predictor for income.", "Aggregate": "The features that consistently appear across multiple masks are: workclass (masks 0, 1, and 3), relationship (masks 0 and 3), and capital gain (masks 2 and 3). This highlights that these features play a pivotal role in predicting whether an individual's income exceeds $50,000 per year. Marital status is seen in the initial stages but disappears later, while hours-per-week and race come into play at the second step. This could indicate that while some features provide foundational predictive power, others may provide more nuanced insight or refinement to the model's predictions. Work-related aspects, personal relationships, and financial metrics seem to be the main predictors of an individual's income level in this dataset. It would be crucial to further analyze these features' exact impacts and understand any potential correlations among them for a more comprehensive understanding."}

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

Output: {"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape's nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn't be overlooked."}

DIABETES PROMPT

Conduct aggregate analysis on the description of the following feature masks. Start off with an analysis of the individual masks, followed by an aggregate analysis of all masks combined. Please format the output into a dictionary as shown in the in-context examples. The output should only contain the formatted output, no other natural language generation is required.

The Diabetes 130-US hospitals for years 1999-2008 dataset is considered, which is a classification of patient encounters into readmitted or not readmitted categories based on attributes like the number of laboratory tests performed, the number of medications prescribed, diagnoses, and other clinical and administrative data. There are 50 test samples and 50 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 7, 21, 31, and 40 which are discharge disposition id, number diagnoses, glyburide, and citoglipton. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 6 and 7 which are admission type id and discharge disposition id. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 6, 21, and 45 which are admission type id, number diagnoses, and metformin. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 21, 30, 45 which are number diagnoses, glipizide, and metformin.

Here are in-context examples for few-shot learning.

The Adult Census Income dataset is considered which predicts whether an individual's income exceeds $50,000 per year. There are 50 test samples and 14 features. At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 1, 5 and 7 which are workclass, marital status and relationship. At the 1st step of feature selection, we observe mask 0 with the main features highlighted as 1 and 5 which are workclass and marital status. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 8, 10 and 12 which are race, capital gain and hours-per-week. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 1, 7 and 10 which are workclass, relationship and capital-gain.

Output: {"Mask 0": "This suggests that early in the feature selection process, the model finds that work-related factors (like the type of work) and personal relationships (both marital status and other relationships) are significant predictors for income exceeding $50,000.", "Mask 1": "Here, the model emphasizes even more on work nature and marital status, suggesting these two as the primary drivers for the prediction. Relationship as a feature has been excluded in this step, which might indicate a possible correlation with marital status.", "Mask 2": "At this step, the model seems to be looking beyond just personal and work nature to more financial and demographic aspects. Capital gain, which is directly related to financial health, is considered along with hours-per-week, potentially pointing to a relationship between longer working hours and higher income. The inclusion of race might indicate underlying societal or economic factors at play.", "Mask 3": "This combination brings back features from both the initial and second masks. Workclass and relationship are revisited, which further underscores their importance. Capital gain remains consistent from the second step, emphasizing its importance as a predictor for income.", "Aggregate": "The features that consistently appear across multiple masks are: workclass (masks 0, 1, and 3), relationship (masks 0 and 3), and capital gain (masks 2 and 3). This highlights that these features play a pivotal role in predicting whether an individual's income exceeds $50,000 per year. Marital status is seen in the initial stages but disappears later, while hours-per-week and race come into play at the second step. This could indicate that while some features provide foundational predictive power, others may provide more nuanced insight or refinement to the model's predictions. Work-related aspects, personal relationships, and financial metrics seem to be the main predictors of an individual's income level in this dataset. It would be crucial to further analyze these features' exact impacts and understand any potential correlations among them for a more comprehensive understanding."}

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

Output: {"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape's nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn't be overlooked."}

HIGGS PROMPT

Conduct aggregate analysis on the description of the following feature mask. Please output ONLY a dictionary and no other natural language generation when generating the sentence as shown in the in-context example below. Please use single-word classification that encapsulates the meaning of the features if possible.

The Higgs dataset is considered, which is a classification of events into signal and background processes based on attributes like jet kinematics, lepton momentum, and other high-energy physics variables. There are 50 test samples and 28 features. At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 22 and 25 which are the invariant mass of the lepton, the missing energy, and the leading jet, as well as the four leading jets. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 6 which is the Azimuthal angle for the second leading jet. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 15 and 25 which are the transverse momentum for the second leading jet and the invariant mass of the lepton, the missing energy, and the leading jet.. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 8 and 16 which are the B-tagging score for the second leading jet and the pseudorapidity of the lepton.

Here are in-context examples for few-shot learning.

The Adult Census Income dataset is considered which predicts whether an individual's income exceeds $50,000 per year. There are 50 test samples and 14 features. At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 1, 5 and 7 which are workclass, marital status and relationship. At the 1st step of feature selection, we observe mask 0 with the main features highlighted as 1 and 5 which are workclass and marital status. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 8, 10 and 12 which are race, capital gain and hours-per-week. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 1, 7 and 10 which are workclass, relationship and capital-gain.

Output: {"Mask 0": "This suggests that early in the feature selection process, the model finds that work-related factors (like the type of work) and personal relationships (both marital status and other relationships) are significant predictors for income exceeding $50,000.", "Mask 1": "Here, the model emphasizes even more on work nature and marital status, suggesting these two as the primary drivers for the prediction. Relationship as a feature has been excluded in this step, which might indicate a possible correlation with marital status.", "Mask 2": "At this step, the model seems to be looking beyond just personal and work nature to more financial and demographic aspects. Capital gain, which is directly related to financial health, is considered along with hours-per-week, potentially pointing to a relationship between longer working hours and higher income. The inclusion of race might indicate underlying societal or economic factors at play.", "Mask 3": "This combination brings back features from both the initial and second masks. Workclass and relationship are revisited, which further underscores their importance. Capital gain remains consistent from the second step, emphasizing its importance as a predictor for income.", "Aggregate": "The features that consistently appear across multiple masks are: workclass (masks 0, 1, and 3), relationship (masks 0 and 3), and capital gain (masks 2 and 3). This highlights that these features play a pivotal role in predicting whether an individual's income exceeds $50,000 per year. Marital status is seen in the initial stages but disappears later, while hours-per-week and race come into play at the second step. This could indicate that while some features provide foundational predictive power, others may provide more nuanced insight or refinement to the model's predictions. Work-related aspects, personal relationships, and financial metrics seem to be the main predictors of an individual's income level in this dataset. It would be crucial to further analyze these features' exact impacts and understand any potential correlations among them for a more comprehensive understanding."}

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

Output: {"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape's nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn't be overlooked."}

C.3. InterpreTabNet: GPT-4 Output Text on All Datasets for Mask Interpretation

Dataset	Output from GPT-4
Adult Census Income	{"Mask 0": "This suggests that early in the feature selection process, the model finds that work-related factors (like the type of work) and personal relationships (both marital status and other relationships) are significant predictors for income exceeding $50,000.", "Mask 1": "Here, the model emphasizes even more on work nature and marital status, suggesting these two as the primary drivers for the prediction. Relationship as a feature has been excluded in this step, which might indicate a possible correlation with marital status.", "Mask 2": "At this step, the model seems to be looking beyond just personal and work nature to more financial and demographic aspects. Capital gain, which is directly related to financial health, is considered along with hours-per-week, potentially pointing to a relationship between longer working hours and higher income. The inclusion of race might indicate underlying societal or economic factors at play.", "Mask 3": "This combination brings back features from both the initial and second masks. Workclass and relationship are revisited, which further underscores their importance. Capital gain remains consistent from the second step, emphasizing its importance as a predictor for income.", "Aggregate": "The features that consistently appear across multiple masks are: workclass (masks 0, 1, and 3), relationship (masks 0 and 3), and capital gain (masks 2 and 3). This highlights that these features play a pivotal role in predicting whether an individual's income exceeds $50,000 per year. Marital status is seen in the initial stages but disappears later, while hours-per-week and race come into play at the second step. This could indicate that while some features provide foundational predictive power, others may provide more nuanced insight or refinement to the model's predictions. Work-related aspects, personal relationships, and financial metrics seem to be the main predictors of an individual's income level in this dataset. It would be crucial to further analyze these features' exact impacts and understand any potential correlations among them for a more comprehensive understanding."}

Forest Cover Type	{"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape's nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn't be overlooked."}
Poker Hand	{"Mask 0": "Initially, the rank of card 2 is recognized as a significant predictor in the classification of poker hands. The rank of individual cards seems crucial at this stage.", "Mask 1": "At this step, the rank of card 3 is highlighted, further emphasizing the importance of individual card ranks in determining poker hand types.", "Mask 2": "Here, besides the rank of card 3, the suit of the same card is considered, suggesting that both rank and suit play roles in classifying poker hands.", "Mask 3": "This step combines multiple features, both ranks (cards 1 and 3) and suits (cards 3 and 5), indicating that a combination of these factors provides a broader understanding of poker hand classifications.", "Aggregate": "The rank of cards, specifically card 2 and card 3, appears consistently across different masks, indicating their primary role in poker hand classification. Suits, particularly of card 3, also have importance but are introduced later in the feature selection. The combined importance of both ranks and suits in the final step suggests that while individual card ranks are pivotal in initial classifications, understanding the relationship between card ranks and their respective suits provides a more comprehensive insight into the poker hand types. The consistent presence of card 3's attributes (both rank and suit) underscores its pivotal role in determining poker hand categories."}

Mushroom	{"Mask 0": "At the outset of the feature selection process, the model identifies physical attributes such as bruises, odor, and veil-type as significant indicators for classifying mushrooms into edible or poisonous categories. This suggests that visible physical changes and inherent characteristics play an essential role in the initial classification stages.", "Mask 1": "In this step, stalk-shape is introduced while retaining bruises and odor from the previous mask. Here, the model seems to emphasize the physical appearance (bruises and stalk-shape) in conjunction with inherent characteristics (odor) for the classification. Veil-type, although significant in the initial step, is omitted here, hinting at its reduced relevance or a possible correlation with other selected features.", "Mask 2": "The model shifts its attention here, emphasizing gill-attachment and stalk-surface-below-ring while still retaining the veil-type. This shift indicates the significance of features directly related to the mushroom's growth structure and its underlying layers. The absence of previously highlighted features suggests a refined approach to feature importance at this step.", "Mask 3": "This combination merges the significant features from the earlier masks, highlighting bruises, odor, gill-attachment, and veil-type. The recurrence of these features underscores their consistent importance in classifying mushrooms. The introduction of gill-attachment here, combined with its presence in the previous mask, further emphasizes its role.", "Aggregate": "The features that consistently appear across multiple masks are: bruises (masks 0, 1, and 3), odor (masks 0, 1, and 3), and veil-type (masks 0, 2, and 3). This suggests that these attributes play a crucial role in distinguishing between edible and poisonous mushrooms in this dataset. Gill-attachment, although appearing in the latter stages (masks 2 and 3), indicates its significance in the classification process. The dataset appears to highlight the importance of physical attributes (like bruises and stalk-shape), inherent characteristics (like odor), and growth-related features (like gill-attachment and stalk-surface-below-ring) in determining the edibility of mushrooms. Further investigation into these features' interactions and exact impacts would provide deeper insights into the classification process."}

Blastchar	{ "Mask 0": "The initial feature selection suggests that aspects related to customer service preferences such as StreamingTV and billing and payment methods like PaperlessBilling and PaymentMethod are significant predictors for customer churn. These features might be pointing towards a customer's comfort and adaptability with the service's digital offerings.", "Mask 1": "In this step, demographic aspects come into play with the inclusion of SeniorCitizen, suggesting that age or seniority might have a bearing on customer churn. The continued emphasis on PaperlessBilling also underlines its importance in this prediction.", "Mask 2": "At this juncture, the model considers the type of service subscribed by the customer, as seen from the inclusion of InternetService and DeviceProtection. MonthlyCharges also comes into the picture, indicating that the amount being charged could play a role in a customer's decision to stay or leave the service. It's possible that customers assess the value they receive from these services against what they are being charged.", "Mask 3": "In this combination, there's a mix of service-related (DeviceProtection), demographic (Partner), and payment (PaymentMethod) features. The appearance of Partner might indicate the influence of family decisions or joint subscriptions on customer churn.", "Aggregate": "PaperlessBilling (masks 0 and 1) and DeviceProtection (masks 2 and 3) consistently appear in multiple steps, highlighting their significance in predicting customer churn. PaymentMethod also appears in masks 0 and 3, suggesting the importance of billing methods in this context. On the other hand, unique features like SeniorCitizen and InternetService give insight into the potential impact of demographics and service type respectively. In essence, customer service preferences, billing methods, and charges, along with certain demographic elements, are key in predicting customer churn in this dataset. For a comprehensive understanding, a deeper dive into the interaction of these features and their respective weights in the prediction model would be beneficial." }

Diabetes	{ "Mask 0": "In the initial step, the model identifies discharge disposition id, number of diagnoses, and specific medications (glyburide, and citoglipton) as key features. This suggests a combination of administrative data (discharge process), clinical data (diagnoses), and prescribed medications play a significant role in predicting patient readmission.", "Mask 1": "Here, the model focuses on the patient's admission type and discharge disposition, emphasizing the importance of the hospital admission and discharge process in determining the likelihood of readmission. The absence of clinical and medication features in this step indicates a potential higher influence of administrative processes at this stage of feature selection.", "Mask 2": "The model reintroduces clinical data (number of diagnoses) and adds a new medication (metformin), while retaining admission type id. This combination underscores a nuanced interplay between the nature of hospital admission, the complexity of patient's health conditions, and specific treatments in predicting readmissions.", "Mask 3": "This step focuses on the number of diagnoses and two medications (glipizide and metformin). The consistency of number of diagnoses across masks highlights its pivotal role, while the introduction of glipizide and retention of metformin point towards an evolving emphasis on medication in conjunction with clinical data.", "Aggregate": "Number diagnoses (masks 0, 2, and 3) and discharge disposition id (masks 0 and 1) are recurrent, underscoring their persistent importance in predicting readmissions. Admission type id also appears consistently (masks 1 and 2), linking the administrative process tightly with readmission predictions. The changing set of medications across the masks (glyburide, citoglipton, metformin, glipizide) points to the dynamic nature of treatment's role in the prediction, warranting a closer look into specific medication combinations and their interactions with clinical and administrative data. The aggregate analysis reveals a complex interplay between administrative processes, clinical data, and medication in predicting patient readmissions, suggesting a multifaceted approach for model refinement and interpretation." }

Higgs	{ "Mask 0": "The initial feature selection emphasizes on variables related to kinematics and energy in the Higgs dataset, particularly invariant mass of the lepton, missing energy, and jets, implying these measurements play a fundamental role in distinguishing signal from background processes.", "Mask 1": "At this step, the model focuses on the Azimuthal angle for the second leading jet, suggesting that the orientation or direction of this jet may have a unique significance in the classification process.", "Mask 2": "Here, the model considers the transverse momentum of the second leading jet alongside the invariant mass of the lepton, missing energy, and leading jet. This suggests an interplay between momentum and energy variables as critical determinants in the classification.", "Mask 3": "In this step, the B-tagging score for the second leading jet and the pseudorapidity of the lepton are highlighted. B-tagging, used to identify the presence of a b-quark, combined with the pseudorapidity, indicates the model's attempt to understand particle behavior and properties for classification.", "Aggregate": "Over the feature selection steps, emphasis is consistently placed on high-energy physics variables, particularly those related to jets and leptons. The invariant mass of the lepton and missing energy are repeated features, appearing in both masks 0 and 2, showcasing their importance in the classification process. Momentum and angular measurements, like the Azimuthal angle and transverse momentum, also play a pivotal role. As the steps progress, there's an evident shift from energy-related features to more particle-specific characteristics, such as B-tagging. Overall, the classification in the Higgs dataset relies heavily on a combination of energy measurements, momentum, and particle properties."}

C.4. TabNet: Prompt and GPT-4 Output on Adult Dataset

Prompt Conduct aggregate analysis on the description of the following feature mask. Please output ONLY a dictionary and no other natural language generation when generating the sentence, as shown in the in-context example below. Please use single-word classification that encapsulates the meaning of the features if possible.

The Adult Census Income dataset is considered which predicts whether an individual's income exceeds $50,000 per year. There are 50 test samples and 14 features.

At the 0th step of feature selection, we observe mask 0 with barely any main features highlighted. At most feature 3 which is education is somewhat highlighted. At the 1st step of feature selection, we observe mask 1 with the main features highlighted as somewhat 7 and mainly 11 which are relationship and capital loss. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 3 and 5 which are education and marital-status. Lastly, at the 3rd step of feature selection, we observe mask 3 with difficulties in discerning the important features.

Here are in-context examples for few-shot learning.

The Poker Hand dataset is considered, which is a classification of poker hands into 10 distinct types based on attributes like rank of individual cards, suits, and card combinations. There are 50 test samples and 10 features.

At the 0th step of feature selection, we observe mask 0 with the main feature highlighted as 3 which is the rank of the card 2. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 7 which is the rank of card 3. At the 2nd step of feature selection, we observe mask 2 with the main feature highlighted as 6 which is the suit of card 3. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 3, 6, 7 and 10 which are the rank of cards 1 and 3 as well as the suits of cards 3 and 5.

Output: {"Mask 0": "Initially, the rank of card 2 is recognized as a significant predictor in the classification of poker hands. The rank of individual cards seems crucial at this stage.", "Mask 1": "At this step, the rank of card 3 is highlighted, further emphasizing the importance of individual card ranks in determining poker hand types.", "Mask 2": "Here, besides the rank of card 3, the suit of the same card is considered, suggesting that both rank and suit play roles in classifying poker hands.", "Mask 3": "This step combines multiple features, both ranks (cards 1 and 3) and suits (cards 3 and 5), indicating that a combination of these factors provides a broader understanding of poker hand classifications.", "Aggregate": "The rank of cards, specifically card 2 and card 3, appears consistently across different masks, indicating their primary role in poker hand classification. Suits, particularly of card 3, also have importance but are introduced later in the feature selection. The combined importance of both ranks and suits in the final step suggests that while individual card ranks are pivotal in initial classifications, understanding the relationship between card ranks and their respective suits provides a more comprehensive insight into the poker hand types. The consistent presence of card 3's attributes (both rank and suit) underscores its pivotal role in determining poker hand categories."}

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

Output: {"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape's nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn't be overlooked."}

GPT-4 Output {"Mask 0": "Initially, only education is somewhat highlighted, suggesting its early recognition as a potential factor in predicting income levels.",

"Mask 1": "This step brings relationship and capital loss into focus, indicating their significant roles in determining income above $50,000.",

"Mask 2": "Education and marital-status are emphasized, reinforcing the importance of educational background and marital status in income prediction.",

"Mask 3": "Difficulties in discerning important features suggest a complexity in determining clear predictors by this stage.",

"Aggregate": "Across the steps, education appears consistently, underscoring its critical role in income prediction. Relationship and capital loss also emerge as key factors, with their introduction in mask 1 highlighting their importance. The repeated emphasis on marital status alongside education in mask 2 further indicates the significance of personal and social factors in determining income levels. However, the challenges faced in mask 3 reveal the complexities in identifying definitive predictors as feature selection progresses."}

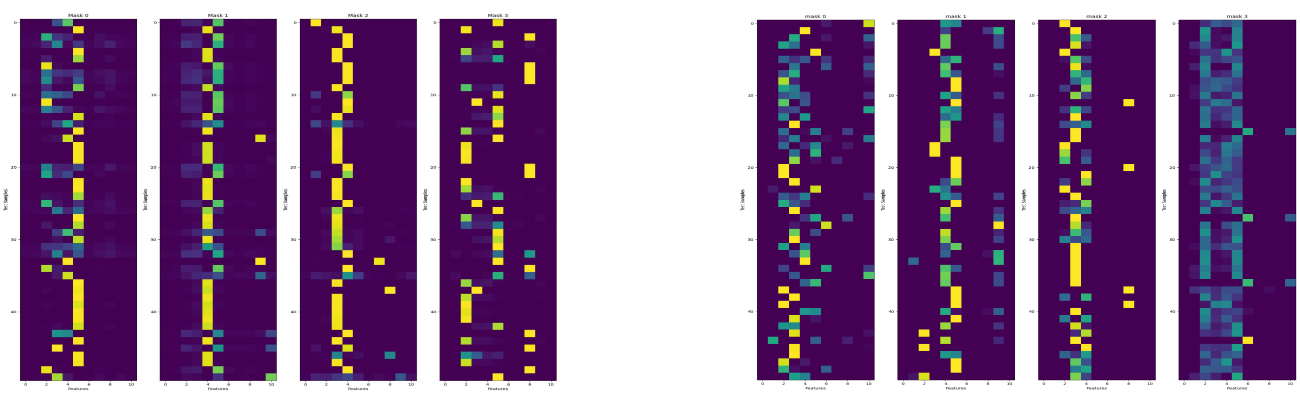
D. Synthetic Dataset Experiments

Table 9: AUC (classification task) mean and standard deviation (20 random seed trails) scores on 6 synthetic datasets from (Chen et al., 2018). Results from the other models are reported from the TabNet paper which includes TabNet (Arik and Pfister, 2020) itself, INVASE (Yoon et al., 2019), L2X (Chen et al., 2018) and Tree Ensembles (Geurts et al., 2006). "Global" refers to the use of only globally salient features. "No Selection" refers to the use of all features without any feature selection. ↑ indicates that the higher the score, the better the performance. The corresponding InterpreTabNet r_M values for each Syn1-Syn6 datasets are as follows: 5250, 4750, 4750, 3250, 5000, 3000. Red bolded numbers denote the best performance for each dataset.

Methods	**Syn1**	**Syn2**	**Syn3**	**Syn4**	**Syn5**	**Syn6**
	AUC ↑	AUC ↑	AUC ↑	AUC ↑	AUC ↑	AUC ↑
InterpreTabNet	**.696±.005**	.885±.003	.899±.004	**.790±.008**	**.791±.006**	**.880±.003**
TabNet	.682±.005	**.892±.004**	.897±.003	.776±.017	.789±.009	.878±.004
INVASE	.690±.006	.877±.003	**.902±.003**	.787±.004	.784±.005	.877±.003
L2X	.498±.005	.823±.029	.862±.009	.678±.024	.709±.008	.827±.017
Lasso-regularized	.498±.006	.555±.061	.886±.003	.512±.031	.691±.024	.727±.025
Tree	.574±.101	.872±.003	.899±.001	.684±.017	.741±.004	.771±.031
Global	.686±.005	.873±.003	.900±.003	.774±.006	.784±.005	.858±.004
No Selection	.578±.004	.789±.003	.854±.004	.558±.021	.662±.013	.692±.015

Synthetic datasets are crucial in determining whether the feature selection process is faithful towards the model's predictions. We use the same synthetic data generation model in TabNet (Arik and Pfister, 2020) and L2X (Chen et al., 2018), as well as a sample size of 10K training and 10K testing. The datasets are formulated where a subset of features determines the prediction. For Syn1-Syn3, they only depend on their specified salient features i.e. Syn1 only depends on features 0 and 1. For Syn4-Syn6, the salient features depend on instance-wise features i.e. the output of Syn4, relies on either features 0-1 or features 2-5 depending on the value of feature 10.

Table 9 indicates that InterpreTabNet with r_M values 5250, 4750, 4750, 3250, 5000, 3000 for each Syn1-Syn6 datasets outperforms existing methods in 4 out of 6 of the synthetic datasets. More notably in all Syn4-Syn6 while remaining competitive in the other two Syn2 and Syn3 datasets. With respect to Syn1-Syn3, InterpreTabNet's competitiveness indicates that it is able to accurately achieve global feature selection. Additionally, for Syn4-Syn6, the SOTA performance indicate that InterpreTabnet is the best at performing global feature selection when instance-wise redundant features are removed.



(a) InterpreTabNet

(b) Original TabNet Model

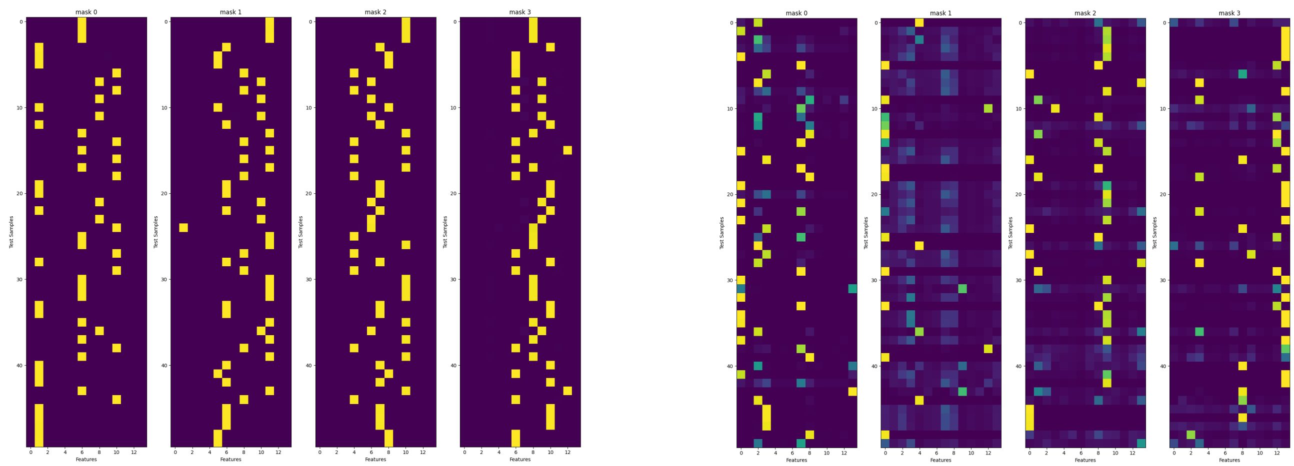
Figure 12: (a) Best performing model using InterpreTabNet $r_M^* = 4,750$ with an AUC of 0.899 on Syn3 Dataset. (b) The baseline performance using the Original TabNet model, attaining an AUC of 0.897.

Syn3 uses features 2-5 to generate its predictions. As observed in Figure 12, InterpreTabNet does indeed determine features 2-5 to be the most salient features as shown by the yellow bars in most of its masks. On the other hand, TabNet depicts lower salience levels when conducting its decision process for the prediction, indicating its uncertainty in reasoning when predicting the ground truth.

E. Justifications and Sanity Checks for Faithfulness and Robustness

E.1. InterpreTabNet Subgroup Feature Importance Analysis for Varying Subpopulation Characteristics

ADULT CENSUS INCOME MASK FIGURES



(a) InterpreTabNet (Sex Subgroup: Male)

(b) InterpreTabNet (Sex Subgroup: Female)

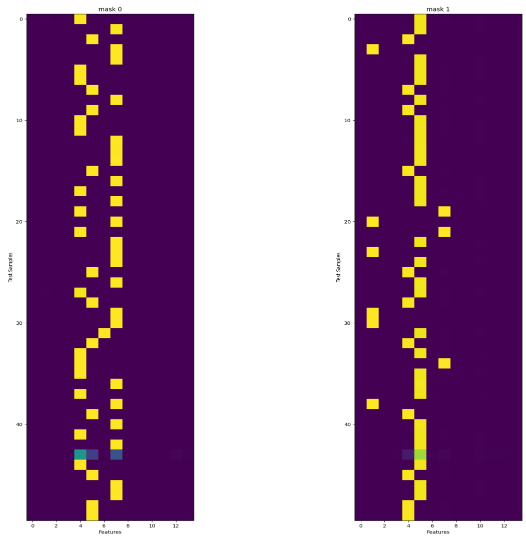
Figure 13: Feature importance visualization across different education subgroups in InterpreTabNet.

To assess InterpreTabNet's robustness and reliability, we conducted a subgroup analysis, examining if salient feature identification differs across subgroups. We divided the adult census income dataset into male and female subpopulations, with the feature importance visualization depicted in Figure 13. For males, mask 0 highlighted work class, occupation, race, and capital gain (features 1, 6, 8, and 10, respectively) as most influential for income prediction. Conversely, mask 0 for females pinpointed age, education, and marital status (features 0, 3, and 5) as salient features contributing to prediction.

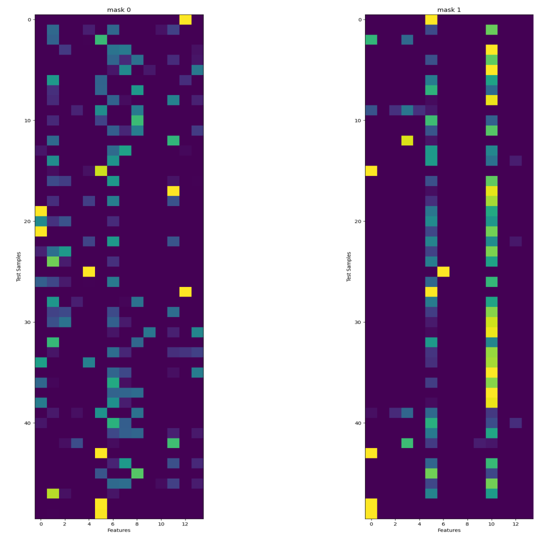
The salient features match well with known socio-economic factors affecting income. For males, occupation and capital gains highlight the importance of job roles and investments. For females, marital status notably influences income, likely due to the socio-economic dynamics and potential household income sharing.

Subsequent masks for males consistently highlight features 6 (occupation), 8 (race), and 10 (capital gain). For females, mask 1 identifies feature 0 (age), mask 2 both feature 0 (age) and 8 (race), and mask 3 points feature 13 (native country). The minimum overlap in identified features suggests our model's robustness in identifying differences in important features between different subpopulations of the data.

E.2. Rationale of Maximising Feature Diversity



(a) InterpreTabNet



(b) Original TabNet Model

Figure 14: (a) Best performing model using $r_M^* = 4$ for InterpreTabNet has an accuracy of 87.48 on the Adult Dataset. (b) The baseline performance using the Original TabNet model, attains an accuracy of 86.87%.

We observe in Figure 14 that even when we select the least possible decision steps $N_{steps} = 2$, TabNet utilizes every single feature aggregated across the two masks. On the other hand, InterpreTabNet selects clear salient features in its decision-making process.

E.3. Training Stability of InterpreTabNet vs. TabNet

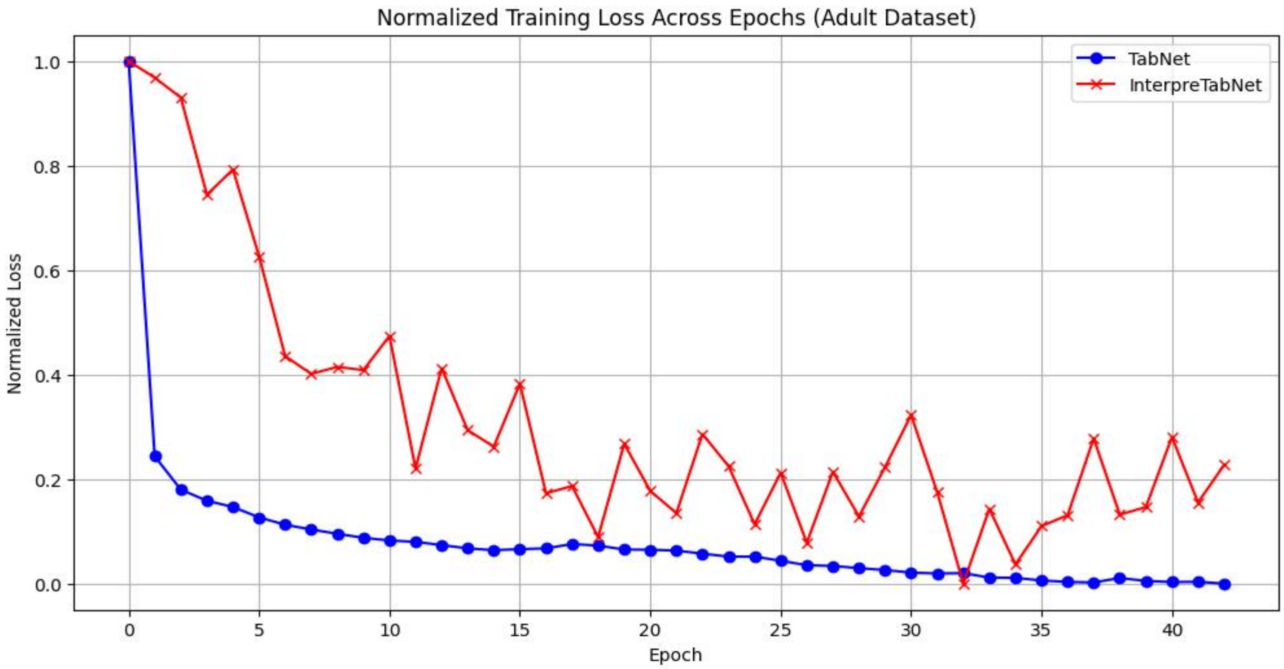


Figure 15: Normalized Training Loss of InterpreTabNet vs. TabNet for the Adult Income Dataset

Figure 15 demonstrates the normalized training loss across epochs for two models: TabNet and InterpreTabNet. TabNet's loss decreases rapidly in the initial epochs and stabilizes quickly. In comparison, InterpreTabNet exhibits higher variability in loss reduction across epochs, which suggests a more exploratory learning process to determine the salient features. Despite the fluctuations, InterpreTabNet's loss shows a general downward trend, which is a positive indicator of learning. Overall, while TabNet appears to have a smoother training process, InterpreTabNet remains competitive.

E.4. Verifying Reliability of GPT-4's Analysis on Synthetic Dataset.

We verify the reliability of GPT-4's analysis by prompting it to interpret synthetic datasets generated in the manner from (Yoon et al., 2019). A synthetic dataset ($Syn1$) is considered where input features are generated from an 11-dimensional Gaussian distribution with no correlations across the features ($X \sim N(0, I)$). The label Y is sampled as a Bernoulli random variable with $\mathbb{P}(Y = 1|X) = \frac{1}{1+logit(X)}$, where $logit(X)$ is varied to create a synthetic dataset as follows: $Syn1 = exp(X_0, X_1)$. As shown in the GPT-4 output below, the analysis is indeed robust and trustworthy since it does not show any signs of hallucinations or mistakes hence, verifying the integrity of GPT-4's ability to interpret feature masks.

Prompt. Conduct aggregate analysis on the description of the following feature mask. Please output ONLY a dictionary and no other natural language generation when generating the sentence as shown in the in-context example below. Please use single-word classification that encapsulates the meaning of the features if possible.

A synthetic dataset (Syn1) is considered where input features are generated from an 11-dimensional Gaussian distribution with no correlations across the features $X \sim N(0, I)$. The label Y is sampled as a Bernoulli random variable with $\mathbb{P}(Y = 1|X) = \frac{1}{1+logit(X)}$, where logit(X) is varied to create a synthetic dataset as follows: Syn1 = $exp(X_0, X_1)$. Make sure to analyze the effect of Syn1 = $exp(X_0, X_1)$ in your aggregate analysis.

At the 0th step of feature selection, we observe mask 0 with the main feature highlighted as 0 which is X_0. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 1 which is X_1. Lastly, at the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 0 and 1 which are X_1 and X_2.

Here are in-context examples for few-shot learning.

The Poker Hand dataset is considered, which is a classification of poker hands into 10 distinct types based on attributes like rank of individual cards, suits, and card combinations. There are 50 test samples and 10 features.

At the 0th step of feature selection, we observe mask 0 with the main feature highlighted as 3 which is the rank of the card 2. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 7 which is the rank of card 3. At the 2nd step of feature selection, we observe mask 2 with the main feature highlighted as 6 which is the suit of card 3. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 3, 6, 7 and 10 which are the rank of cards 1 and 3 as well as the suits of cards 3 and 5.

Output: {"Mask 0": "Initially, the rank of card 2 is recognized as a significant predictor in the classification of poker hands. The rank of individual cards seems crucial at this stage.", "Mask 1": "At this step, the rank of card 3 is highlighted, further emphasizing the importance of individual card ranks in determining poker hand types.", "Mask 2": "Here, besides the rank of card 3, the suit of the same card is considered, suggesting that both rank and suit play roles in classifying poker hands.", "Mask 3": "This step combines multiple features, both ranks (cards 1 and 3) and suits (cards 3 and 5), indicating that a combination of these factors provides a broader understanding of poker hand classifications.", "Aggregate": "The rank of cards, specifically card 2 and card 3, appears consistently across different masks, indicating their primary role in poker hand classification. Suits, particularly of card 3, also have importance but are introduced later in the feature selection. The combined importance of both ranks and suits in the final step suggests that while individual card ranks are pivotal in initial classifications, understanding the relationship between card ranks and their respective suits provides a more comprehensive insight into the poker hand types. The consistent presence of card 3's attributes (both rank and suit) underscores its pivotal role in determining poker hand categories."}

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

Output: {"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape's nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn't be overlooked."}

GPT-4 Output. {"Mask 0": "Initially, the feature X_0 is identified as a significant predictor for the dataset. This step suggests the individual effect of X_0 on the label Y is crucial.",

"Mask 1": "At this step, feature X_1 is highlighted, indicating its individual significance in influencing the label Y, similar to X_0's initial impact.",

"Mask 2": "Here, both features X_0 and X_1 are highlighted together, suggesting that the combined effect of these features significantly impacts the label Y, beyond their individual contributions.",

"Aggregate": "The progression of feature selection indicates a transition from individual to combined feature importance. Initially, X_0 and then X_1 are identified as significant on their own. However, by the final step, the combination of X_0 and X_1 is recognized as crucial, highlighting the importance of interactions between features in influencing the label Y. This suggests that while individual features have their significance, the interaction between X_0 and X_1, especially considering the synthetic dataset's nature (Syn1 = $exp(X_0, X_1)$), is vital for predicting Y. The synthetic dataset's structure emphasizes the multiplicative interaction between X_0 and X_1, illustrating the complex relationship between features in modeling."}

E.5. Prompt Variability

Table 10: GPT-4 Prompts in different formats

Section	Description
Goal & Format	Generate a dictionary analysis of the feature masks described below. The output should consist solely of a dictionary, with no additional explanatory text. Keys should be the mask identifiers, and values should be lists containing single-word descriptors that capture the essence of the features identified at each step of feature selection. The descriptors should be as succinct as possible.
Dataset Description	For the Adult Census Income dataset, which predicts whether an individual's income exceeds $50,000 per year based on 50 test samples and 14 features, the following observations were made:
Mask Description	At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 1, 5 and 7 which are workclass, marital status and relationship. At the 1st step of feature selection, we observe mask 0 with the main features highlighted as 1 and 5 which are workclass and marital status. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 8, 10 and 12 which are race, capital gain and hours-per-week. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 1, 7 and 10 which are workclass, relationship and capital-gain.
GPT-4 Output	*{ "Mask 0": ["Employment", "Marital", "Relationship"], "Mask 1": ["Employment", "Marital"], "Mask 2": ["Ethnicity", "Income", "Worktime"], "Mask 3": ["Employment", "Relationship", "Income"] }*

Table 11: GPT-4 Prompts in varying detail levels

Section	Description
Less Detailed Goal	Generate a dictionary with key insights from feature masks. Output a dictionary with keys as 'Mask <step>' and values as single-word descriptors of the most significant feature(s) at each step. The descriptors should summarize the feature's impact on predicting an individual's income in the Adult Census Income dataset.
Moderately Detailed Goal	Construct a dictionary summarizing the significance of feature masks in predicting income levels in the Adult Census Income dataset. For each mask, output a dictionary with a list of descriptors, each a concise term that captures the importance of the features at that step of selection.
More Detailed Goal	Create a detailed dictionary explaining the role of each feature mask in the context of the Adult Census Income dataset. Each dictionary entry should include the step number as the key and a list of phrases as the value, capturing the nuanced significance of each feature in predicting whether an individual's income exceeds $50,000 per year.
GPT-4 Less Detailed Output	*{ "Mask 0": ["Primary"], "Mask 1": ["Secondary"], "Mask 2": ["Tertiary"], "Mask 3": ["Quaternary"] }*
GPT-4 Moderately Detailed Output	*{"Mask 0": ["Work", "Family", "Relations"], "Mask 1": ["Work", "Family"], "Mask 2": ["Ethnicity", "Investment", "Labor"], "Mask 3": ["Work", "Relations", "Investment"]}*
GPT-4 More Detailed Output	*{"Mask 0": ["Employment class", "Marital bonds", "Social ties"], "Mask 1": ["Employment status", "Marital status"], "Mask 2": ["Ethnicity diversity", "Capital growth", "Work duration"], "Mask 3": ["Employment nature", "Social relationship", "Capital investment"]}*

E.6. Prompting GPT-4 with Persona of a Domain Expert

We conditioned GPT-4 on a prompt where the model was asked to comment on the feature masks as if they were an expert in the broad topic that the predictive model was built for i.e. "Assume that you are an Economist - Adult Dataset", "Assume that you are a Physician - Diabetes Dataset". The goal of this experiment is to give the reader a qualitative assessment of the variability in the interpretations based on what an expert in the domain might care about.

We summarize the variation of results in Sections E.6 and E.6 — what we find is that the expert interpretations do not vary significantly compared to the original interpretations given different "expert" prompts. Therefore, we can anticipate that the interpretations are "robust".

ADULT DATASET - ECONOMIST

Dataset & Expert Persona	**Output from GPT-4**
Adult Dataset, Economist	{ "Mask 0": "Initial selection highlights workclass, marital status, and relationship as key predictors of income levels, indicating that employment type, marital situation, and family roles are crucial for income classification.", "Mask 1": "This step maintains focus on workclass and marital status, reaffirming their importance in predicting income levels and suggesting a consistent pattern in the significance of employment type and marital situation.", "Mask 2": "The focus shifts to race, capital gain, and hours-per-week, introducing new dimensions such as ethnic background, financial investment returns, and work hours as significant factors influencing income.", "Mask 3": "Reintroduces workclass and relationship along with capital gain, indicating that while employment type and family roles remain pertinent, financial investments emerge as a critical predictor.", "Aggregate": "Workclass and marital status appear as consistent predictors across multiple masks, emphasizing the role of employment type and marital situation in income determination. Relationship and capital gain also emerge as significant, with the latter highlighting financial investments' importance. The inclusion of race and hours-per-week in one mask suggests a broader perspective, considering ethnic background and work dedication. Overall, employment type, marital status, financial investments, and to a lesser extent, family roles, ethnic background, and work hours, are key to predicting income levels." }

Adult Dataset, No Persona	{"Mask 0": "This suggests that early in the feature selection process, the model finds that work-related factors (like the type of work) and personal relationships (both marital status and other relationships) are significant predictors for income exceeding $50,000.", "Mask 1": "Here, the model emphasizes even more on work nature and marital status, suggesting these two as the primary drivers for the prediction. Relationship as a feature has been excluded in this step, which might indicate a possible correlation with marital status.", "Mask 2": "At this step, the model seems to be looking beyond just personal and work nature to more financial and demographic aspects. Capital gain, which is directly related to financial health, is considered along with hours-per-week, potentially pointing to a relationship between longer working hours and higher income. The inclusion of race might indicate underlying societal or economic factors at play.", "Mask 3": "This combination brings back features from both the initial and second masks. Workclass and relationship are revisited, which further underscores their importance. Capital gain remains consistent from the second step, emphasizing its importance as a predictor for income.", "Aggregate": "The features that consistently appear across multiple masks are: workclass (masks 0, 1, and 3), relationship (masks 0 and 3), and capital gain (masks 2 and 3). This highlights that these features play a pivotal role in predicting whether an individual's income exceeds $50,000 per year. Marital status is seen in the initial stages but disappears later, while hours-per-week and race come into play at the second step. This could indicate that while some features provide foundational predictive power, others may provide more nuanced insight or refinement to the model's predictions. Work-related aspects, personal relationships, and financial metrics seem to be the main predictors of an individual's income level in this dataset. It would be crucial to further analyze these features' exact impacts and understand any potential correlations among them for a more comprehensive understanding."}

Prompt *Assume that you are an Economist.* Conduct aggregate analysis on the description of the following feature mask. Please output ONLY a dictionary and no other natural language generation when generating the sentence as shown in the in-context example below. Please use single-word classification that encapsulates the meaning of the features if possible.

The Adult Census Income dataset is considered which predicts whether an individual's income exceeds $50,000 per year. There are 50 test samples and 14 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 1, 5 and 7 which are workclass, marital status and relationship. At the 1st step of feature selection, we observe mask 1 with the main features highlighted as 1 and 5 which are workclass and marital status. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 8, 10 and 12 which are race, capital gain and hours-per-week. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 1, 7 and 10 which are workclass, relationship and capital-gain.

Here are in-context examples for few-shot learning.

The Poker Hand dataset is considered, which is a classification of poker hands into 10 distinct types based on attributes like rank of individual cards, suits, and card combinations. There are 50 test samples and 10 features.

At the 0th step of feature selection, we observe mask 0 with the main feature highlighted as 3 which is the rank of the card 2. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 7 which is the rank of card 3.

At the 2nd step of feature selection, we observe mask 2 with the main feature highlighted as 6 which is the suit of card 3. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 3, 6, 7 and 10 which are the rank of cards 1 and 3 as well as the suits of cards 3 and 5.

Output: {"Mask 0": "Initially, the rank of card 2 is recognized as a significant predictor in the classification of poker hands. The rank of individual cards seems crucial at this stage.", "Mask 1": "At this step, the rank of card 3 is highlighted, further emphasizing the importance of individual card ranks in determining poker hand types.", "Mask 2": "Here, besides the rank of card 3, the suit of the same card is considered, suggesting that both rank and suit play roles in classifying poker hands.", "Mask 3": "This step combines multiple features, both ranks (cards 1 and 3) and suits (cards 3 and 5), indicating that a combination of these factors provides a broader understanding of poker hand classifications.", "Aggregate": "The rank of cards, specifically card 2 and card 3, appears consistently across different masks, indicating their primary role in poker hand classification. Suits, particularly of card 3, also have importance but are introduced later in the feature selection. The combined importance of both ranks and suits in the final step suggests that while individual card ranks are pivotal in initial classifications, understanding the relationship between card ranks and their respective suits provides a more comprehensive insight into the poker hand types. The consistent presence of card 3's attributes (both rank and suit) underscores its pivotal role in determining poker hand categories."}

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

Output: {"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape's nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn't be overlooked."}

Diabetes Dataset - Physician

Dataset & Expert Persona	Output from GPT-4
Diabetes Dataset, Physician	{"Mask 0": "This initial selection highlights the importance of discharge disposition, the number of diagnoses, and specific medications (glyburide and citoglipton) in predicting readmissions. It suggests that patient outcomes and the specific treatment protocols, including medication use, play a crucial role in the likelihood of a patient being readmitted.", "Mask 1": "At this stage, the model focuses on the admission type and discharge disposition, indicating the significance of the nature of admission and the patient's discharge status in predicting readmissions. This could reflect how initial hospitalization circumstances and post-discharge plans impact patient outcomes.", "Mask 2": "Here, the inclusion of admission type, number of diagnoses, and metformin (another medication) suggests a continued emphasis on the circumstances of admission and the complexity of the patient's health condition, as well as the role of specific diabetes medication in influencing readmission risks.", "Mask 3": "By focusing on the number of diagnoses and medications (glipizide and metformin), this step underscores the critical nature of the patient's health status and the treatment of diabetes as key factors in predicting hospital readmissions. It highlights the significance of managing complex health conditions and diabetes treatment.", "Aggregate": "Across the selection process, the consistent features are the number of diagnoses and the emphasis on specific diabetes medications (metformin appearing in masks 2 and 3, and other diabetes-related medications in masks 0 and 3). This indicates a strong relationship between the complexity of a patient's health (as represented by the number of diagnoses) and the treatment of diabetes in predicting readmissions. Discharge disposition and admission type are also significant but might offer insights into the logistical aspects of care that affect readmission risks. The aggregate analysis underscores the complexity of managing diabetes within the hospital setting and its impact on readmission rates, highlighting the importance of both clinical and administrative factors."}

Diabetes Dataset, No Persona	{ "Mask 0": "In the initial step, the model identifies discharge disposition id, number of diagnoses, and specific medications (glyburide, and citoglipton) as key features. This suggests a combination of administrative data (discharge process), clinical data (diagnoses), and prescribed medications play a significant role in predicting patient readmission.", "Mask 1": "Here, the model focuses on the patient's admission type and discharge disposition, emphasizing the importance of the hospital admission and discharge process in determining the likelihood of readmission. The absence of clinical and medication features in this step indicates a potential higher influence of administrative processes at this stage of feature selection.", "Mask 2": "The model reintroduces clinical data (number of diagnoses) and adds a new medication (metformin), while retaining admission type id. This combination underscores a nuanced interplay between the nature of hospital admission, the complexity of patient's health conditions, and specific treatments in predicting readmissions.", "Mask 3": "This step focuses on the number of diagnoses and two medications (glipizide and metformin). The consistency of number of diagnoses across masks highlights its pivotal role, while the introduction of glipizide and retention of metformin point towards an evolving emphasis on medication in conjunction with clinical data.", "Aggregate": "Number diagnoses (masks 0, 2, and 3) and discharge disposition id (masks 0 and 1) are recurrent, underscoring their persistent importance in predicting readmissions. Admission type id also appears consistently (masks 1 and 2), linking the administrative process tightly with readmission predictions. The changing set of medications across the masks (glyburide, citoglipton, metformin, glipizide) points to the dynamic nature of treatment's role in the prediction, warranting a closer look into specific medication combinations and their interactions with clinical and administrative data. The aggregate analysis reveals a complex interplay between administrative processes, clinical data, and medication in predicting patient readmissions, suggesting a multifaceted approach for model refinement and interpretation." }

Prompt *Assume you are a Physician.* Conduct aggregate analysis on the description of the following feature masks. Start off with an analysis of the individual masks, followed by an aggregate analysis of all masks combined. Please format the output into a dictionary as shown in the in-context examples. The output should only contain the formatted output, no other natural language generation is required.

The Diabetes 130-US hospitals for years 1999-2008 dataset is considered, which is a classification of patient encounters into readmitted or not readmitted categories based on attributes like the number of laboratory tests performed, the number of medications prescribed, diagnoses, and other clinical and administrative data. There are 50 test samples and 50 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 7, 21, 31, and 40 which are discharge disposition id, number diagnoses, glyburide, and citoglipton. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 6 and 7 which are admission type id and discharge disposition id. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 6, 21, and 45 which are admission type id, number diagnoses, and metformin. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 21, 30, 45 which are number diagnoses, glipizide, and metformin.

Here are in-context examples for few-shot learning.

The Adult Census Income dataset is considered which predicts whether an individual's income exceeds $50,000 per year.

There are 50 test samples and 14 features. At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 1, 5 and 7 which are workclass, marital status and relationship. At the 1st step of feature selection, we observe mask 0 with the main features highlighted as 1 and 5 which are workclass and marital status. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 8, 10 and 12 which are race, capital gain and hours-per-week. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 1, 7 and 10 which are workclass, relationship and capital-gain.

Output: {"Mask 0": "This suggests that early in the feature selection process, the model finds that work-related factors (like the type of work) and personal relationships (both marital status and other relationships) are significant predictors for income exceeding $50,000.", "Mask 1": "Here, the model emphasizes even more on work nature and marital status, suggesting these two as the primary drivers for the prediction. Relationship as a feature has been excluded in this step, which might indicate a possible correlation with marital status.", "Mask 2": "At this step, the model seems to be looking beyond just personal and work nature to more financial and demographic aspects. Capital gain, which is directly related to financial health, is considered along with hours-per-week, potentially pointing to a relationship between longer working hours and higher income. The inclusion of race might indicate underlying societal or economic factors at play.", "Mask 3": "This combination brings back features from both the initial and second masks. Workclass and relationship are revisited, which further underscores their importance. Capital gain remains consistent from the second step, emphasizing its importance as a predictor for income.", "Aggregate": "The features that consistently appear across multiple masks are: workclass (masks 0, 1, and 3), relationship (masks 0 and 3), and capital gain (masks 2 and 3). This highlights that these features play a pivotal role in predicting whether an individual's income exceeds $50,000 per year. Marital status is seen in the initial stages but disappears later, while hours-per-week and race come into play at the second step. This could indicate that while some features provide foundational predictive power, others may provide more nuanced insight or refinement to the model's predictions. Work-related aspects, personal relationships, and financial metrics seem to be the main predictors of an individual's income level in this dataset. It would be crucial to further analyze these features' exact impacts and understand any potential correlations among them for a more comprehensive understanding."}

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

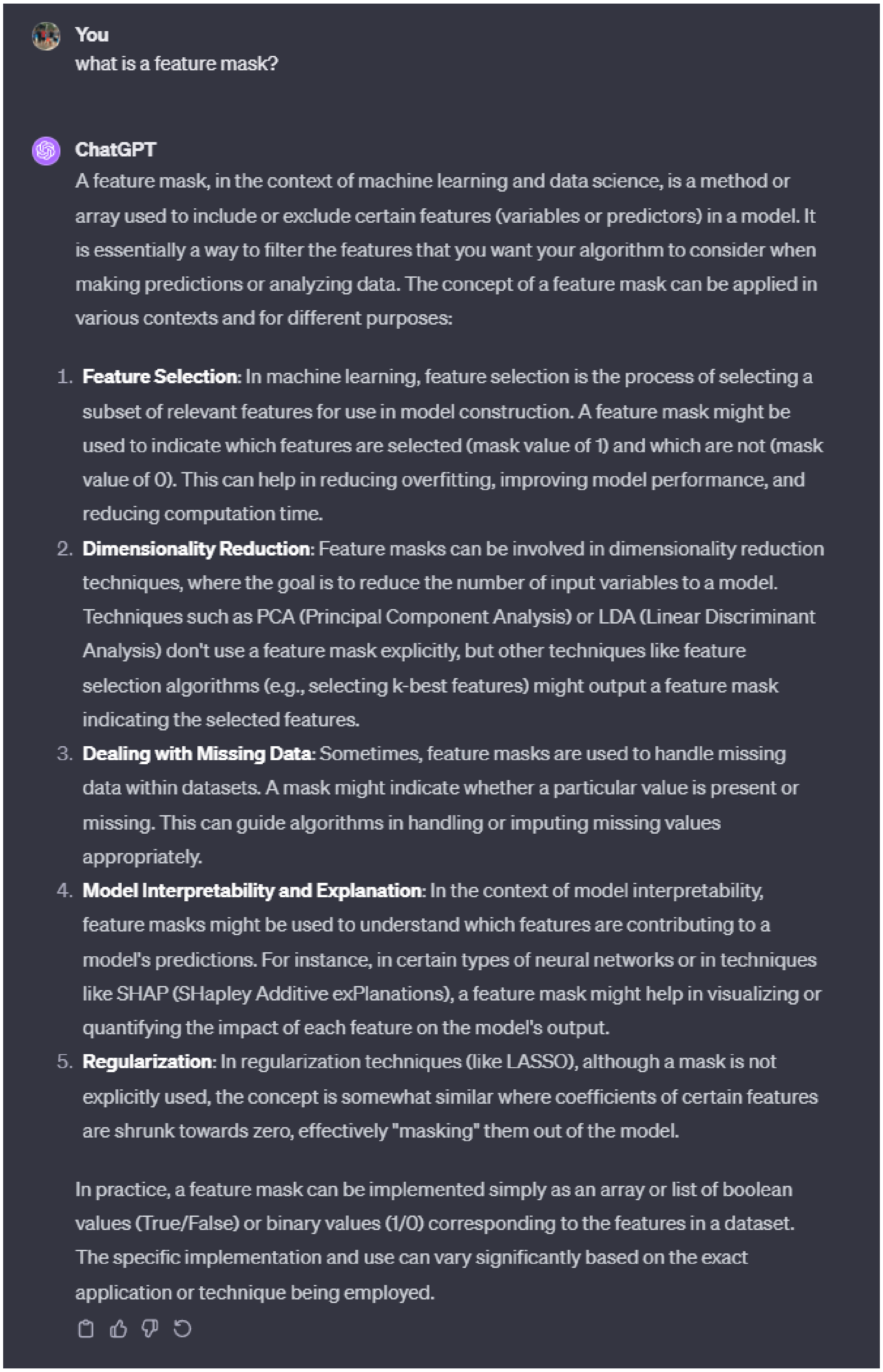
Output: {"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape's nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn't be overlooked."}

E.7. GPT-4 Analysis without InterpreTabNet

Table 14: GPT-4 Analysis without InterpreTabNet

Prompts	Outputs
Identify the most important features of the following dataset and conduct aggregate text analysis on it. Please output ONLY a dictionary and no other natural language generation when generating the sentence. The Adult Census Income dataset is considered which predicts whether an individual's income exceeds $50,000 per year. The features are as follows: Age. Workclass. Final Weight. Education. Education Number of Years. Marital-status. Occupation. Relationship. Race. Sex. Capital-gain. Capital-loss. Hours-per-week. Native-country.	{ "Important Features": "Numerical": ["Age", "Final Weight", "Education Number of Years", "Capital-gain", "Capital-loss", "Hours-per-week"], "Categorical": ["Workclass", "Education", "Marital-status", "Occupation", "Relationship", "Race", "Sex", "Native-country"] , "Aggregate Text Analysis": { "Income Prediction": "The Adult Census Income dataset is used to predict if an individual's income exceeds $50,000 per year based on various demographic and financial features." } }
Indicate the most important features of the following dataset. Please output ONLY a dictionary and no other natural language generation when generating the sentence. The Adult Census Income dataset is considered which predicts whether an individual's income exceeds $50,000 per year. The features are as follows: Age. Workclass. Final Weight. Education. Education Number of Years. Marital-status. Occupation. Relationship. Race. Sex. Capital-gain. Capital-loss. Hours-per-week. Native-country.	{ "Important Features": ["Age", "Education Number of Years", "Marital-status", "Occupation", "Relationship", "Capital-gain", "Capital-loss", "Hours-per-week"] }

E.8. Feature Mask Definition Check



You
what is a feature mask?

ChatGPT
A feature mask, in the context of machine learning and data science, is a method or array used to include or exclude certain features (variables or predictors) in a model. It is essentially a way to filter the features that you want your algorithm to consider when making predictions or analyzing data. The concept of a feature mask can be applied in various contexts and for different purposes:

1. **Feature Selection**: In machine learning, feature selection is the process of selecting a subset of relevant features for use in model construction. A feature mask might be used to indicate which features are selected (mask value of 1) and which are not (mask value of 0). This can help in reducing overfitting, improving model performance, and reducing computation time.
2. **Dimensionality Reduction**: Feature masks can be involved in dimensionality reduction techniques, where the goal is to reduce the number of input variables to a model. Techniques such as PCA (Principal Component Analysis) or LDA (Linear Discriminant Analysis) don't use a feature mask explicitly, but other techniques like feature selection algorithms (e.g., selecting k-best features) might output a feature mask indicating the selected features.
3. **Dealing with Missing Data**: Sometimes, feature masks are used to handle missing data within datasets. A mask might indicate whether a particular value is present or missing. This can guide algorithms in handling or imputing missing values appropriately.
4. **Model Interpretability and Explanation**: In the context of model interpretability, feature masks might be used to understand which features are contributing to a model's predictions. For instance, in certain types of neural networks or in techniques like SHAP (SHapley Additive exPlanations), a feature mask might help in visualizing or quantifying the impact of each feature on the model's output.
5. **Regularization**: In regularization techniques (like LASSO), although a mask is not explicitly used, the concept is somewhat similar where coefficients of certain features are shrunk towards zero, effectively "masking" them out of the model.

In practice, a feature mask can be implemented simply as an array or list of boolean values (True/False) or binary values (1/0) corresponding to the features in a dataset. The specific implementation and use can vary significantly based on the exact application or technique being employed.

Figure 16: Feature Mask Definition Check