InterpreTabNet: Distilling Predictive Signals From Tabular Data

Anonymous Authors[1]

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. We find that the inferred attention masks on high-dimensional data are often dense, hindering interpretability. To remedy this, we propose the 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 automate the interpretation of feature interdependencies from our model, we employ 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 our InterpreTabNet Model 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 and Strack, 2014), insurance (Datta, 2020), and finance (Moro and Cortez, 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, from a large collection of features, those 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 selection 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 models of tabular data, 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 less sparse representation of feature importance, making it difficult to identify important features.

The objective of our work is to *enhance 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., 2022) and provide textual summaries of our sparse masks.

Our work makes the following contributions:

---

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

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 a reuse of features across attention masks within the architecture. Empirically, under our regularization scheme, the model learns to distill the predictive signals 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.

3. Our method simplifies the learned importance masks generated under TabNet, one potential concern is that our method is unable to capture the rich interdependencies between features that are needed to interpret model predictions in complex settings. We show that *leveraging rich linguistic priors in interpretation by means of a large language model largely ameliorates these concerns*. We demonstrate how language models can relate the learned feature masks to a world model underlying the LLM (Hao et al., 2023) 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, uses 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 can lead to suboptimal feature selection. To address this limitation, diversity-promoting regularizers and latent models attempt to solve this problem (Xie et al., 2017) (Xie et al., 2016). Subsequent work on tabular data includes Net-DNF (Katzir et al., 2020), SubTab (Ucar et al., 2021), and TabTransformer (Huang et al., 2020). Net-DNF (Katzir et al., 2020) introduces an inductive bias that aligns model structures with disjunctive normal form (DNF) and emphasizes localized decisions. SubTab (Ucar et al., 2021) transforms 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 have demonstrated 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 latent variables and observed data, both on random variables, which gain control of the data generation process on the VAE. Additionally, the cVAE also generates diverse but realistic output representations using stochastic inference (Sohn et al., 2015). Transformer-based cVAE demonstrates its 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 works in approximate inference for categorical data include 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 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

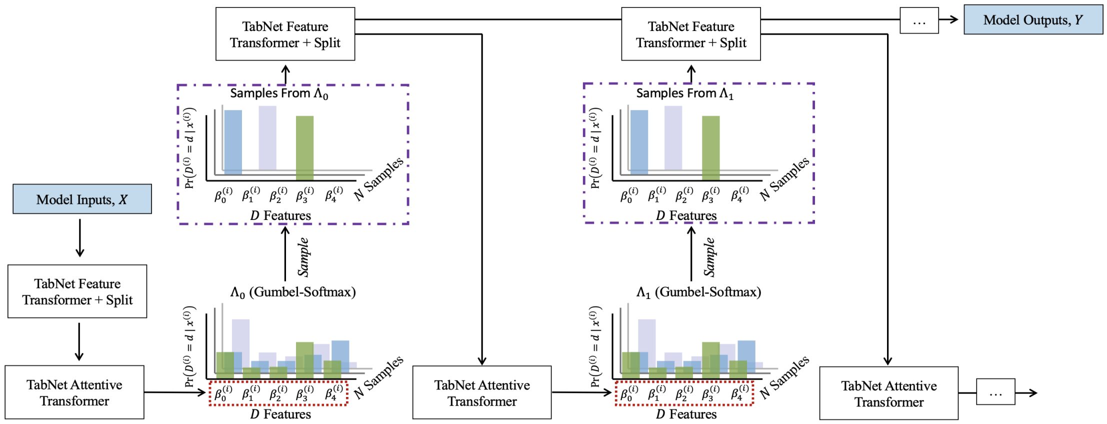

TabNet Feature Transformer + Split
TabNet Feature Transformer + Split
…
Model Outputs, Y
Samples From Λ0
Samples From Λ1
Pr(D(i) = d | x(i))
β0(i) β1(i) β2(i) β3(i) β4(i)
N Samples
D Features
Model Inputs, X
Sample
Λ0 (Gumbel-Softmax)
Λ1 (Gumbel-Softmax)
TabNet Feature Transformer + Split
TabNet Attentive Transformer
TabNet Attentive Transformer
TabNet Attentive Transformer
…

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.

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 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 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 prediction process, $P(y \,|\, x)$, 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:

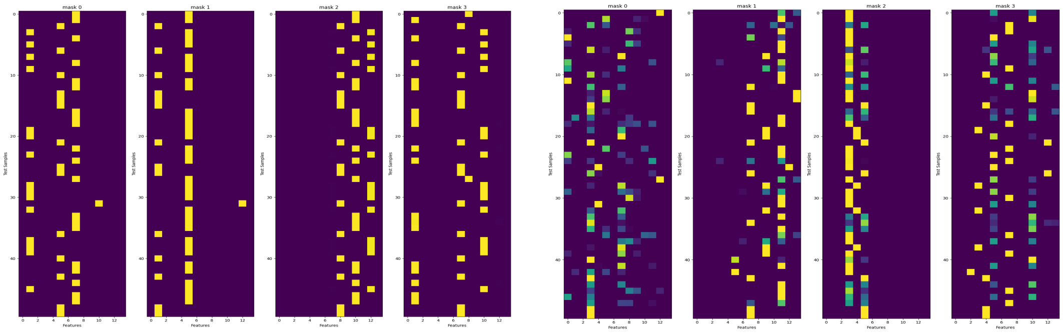
$$\Pr(y \,|\, x) = f_\psi^{\text{(MLP)}} \left(\sum_{k=0}^{K-1} f_\psi^{\text{(TabNet\_Transformer)}}(m_k, x) \right), \quad (1)$$

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 application of the reparameterization trick under our method.

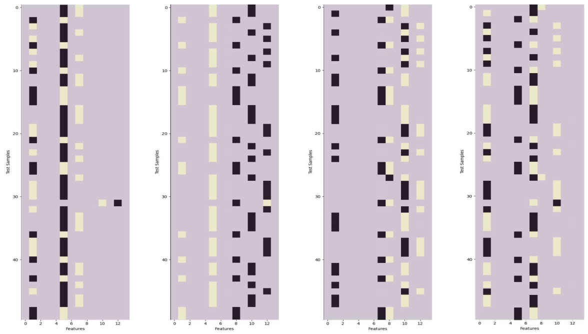
---

[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 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 4.1.2.

3.2. Mask Sampling Process

The mask sampling process for our model 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.

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

TabNet's model does not leverage its feature importance masks to make predictions. Instead, it acts as a deterministic system by producing its masks directly via its attentive transformer. On the other hand, 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)[2]

---

[2] As a small technical note, the original TabNet architecture requires a ReLU function 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 $Q(z|Y)$. The derivation can be found in Appendix 5.1.

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

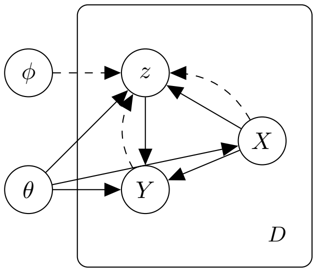


Figure 3: Graphical model of InterpreTabNet with D i.i.d samples. Solid lines denote the generative model $p_\theta(Y|z,X)p_\theta(z|X)$, dashed lines denote the variational approximation $q_\phi(z|X,Y)$ to the intractable posterior $p_\theta(z|X,Y)$. The variational parameters ϕ are learned jointly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

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

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.[3]

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 which contributes 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 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 fulfillment 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 5.2.

4. Experiments and Discussions

We evaluated the performance of InterpreTabNet on real-world classification tasks both quantitatively and qualitatively.

Datasets. The model performance was evaluated on real-world tabular datasets from UCI Machine Learning Repository (Kelly et al., 2023) and OpenML (Vanschoren et al., 2013). These datasets were selected since they were used to evaluate the existing methods (baselines). Additionally, they varied 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 was 80/10/10% apart from the Higgs dataset. Due to the inherently large Higgs dataset, we followed TabNet's method of data splitting with 500k training samples, 100k

---

[3] There are more criteria than those we enumerate here. Those enumerated here are exemplars to aid understanding.

validation samples and 100k testing samples. Details of the datasets can be found in Appendix 5.4.

Baselines: Accuracy. We compared our model against five other ML methods for tabular classification. This included 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 utilized the recommended hyperparameters mentioned by the authors of their respective papers. Furthermore, we also conducted a grid search within the range of the recommended hyperparameters to optimize the models, selecting the best-performing hyperparameter configuration.

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

Regularizer r_M Ablation Study In an ablation study on how varying r_M values affected our masks (found in Appendix 5.7), we noticed that at low r_M values, test accuracy was high but feature selection diversity was poor, and interpretability were difficult since almost all features were selected in the decision-making process. On the other hand, at high r_M values, the masks were sparse which were easily interpretable but at a cost of accuracy. Therefore, we ensured that our selected r_M using the Sparsity Regularizer Algorithm in Section 3.5 provided us with a compromise of a competitive accuracy while having the best interpretability against the baseline models.

Computational Efficiency Our model necessitated 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 incured a minimal cost, leading to a mere several-minute increase in training time. Furthermore, likewise to TabNet, our model maintained greater computational efficiency compared to other baseline models without necessitating an extensive search for fine-grained hyperparameters.

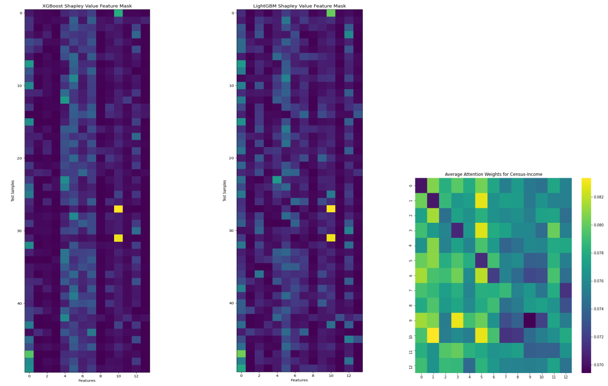
4.1. Results

In the following section, we explored the Adult Census Income dataset (Becker and Kohavi, 1996) to evaluate InterpreTabNet against other baselines.[4]

4.1.1. QUANTITATIVE ANALYSIS

The performance of our method relative to the baselines for tabular learning is shown in Table 1. As observed, we achieved the best performance in 2/7 datasets while maintaining a competitive accuracy for the remaining 5/7 datasets. Our most notable contribution was achieving a significant improvement in interpretability.

4.1.2. QUALITATIVE ANALYSIS



(a) XGBoost SHAP (b) LightGBM SHAP (c) TabTransformer 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).

Interpretability Evaluation Figure 2 highlighted 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 represented individual data samples, while the columns represented discrete features in the tabular data. Values of feature importance for each test sample/row sum up to 1. Thus, bright yellow squares indicated values close to/equal to 1, dark purple squares indicated values close to/equal to 0, and other color squares indicated values between 0 and 1.

As observed in Figure 2, our InterpreTabNet model highlighted mutually exclusive features of high importance that were easily interpretable. Practioners could easily identify

---

[4] In the remainder of our paper, we will conduct all analyses on the Adult Census Income dataset for simplicity purposes.

Table 1: Test Accuracy Scores with Optimal Mask Regularizer Values (r_M) across Different Models and Datasets. Our InterpreTabNet achieved 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**	**94.75**	99.50	96.62	72.96	55.37	53.08
Original TabNet	85.55	94.18	99.00	99.94	76.22	56.91	52.94
XGBoost	86.60	92.30	75.57	99.69	77.29	**61.44**	**72.70**
LightGBM	86.20	86.38	78.47	**100.00**	**77.86**	60.87	72.62
TabTransformer	85.09	82.55	**99.81**	**100.00**	73.17	44.45	51.97
MLP	79.76	84.89	99.70	99.82	75.16	53.99	63.17

the salient features that were contributing to the outcome prediction. On the contrary, feature masks of the Original TabNet were more difficult to interpret since each mask highlighted multiple features for a given data sample. Additionally, when compared against the other baselines, InterpreTabNet leveraged sequential decision-making that allowed users to understand how the model's focus shifted and how different features interacted over the decision process.

Figure 4 illustrated the complex pathways of model interpretation inherent in our baseline architectures like XGBoost, LightGBM, and TabTransformer. These models necessitated additional tools to render interpretative insights. Both XGBoost and LightGBM were augmented with SHAP values derived from external SHAP packages to achieve interpretability while TabTransformer relied on attention weights.

The reliance on SHAP values in XGBoost and LightGBM found in Figure 4 parts (a) and (b) distributed the contribution of the prediction across all features. This led to a less sparse representation of feature importance. In practice, this meant that while each feature's contribution to the prediction was identified, the significance of each feature was not as distinct. This led to an interpretation where barely any feature stood out, especially in models with a large number of features where many contributed incrementally to the final prediction. This lack of sparsity made it challenging for practitioners to pinpoint a concise set of features for understanding and analysis.

The attention mechanism of TabTransformer in Figure 4 part (c) provided a form of interpretability by capturing relationships between features. However, it was unable to clearly pinpoint a set of important features. Furthermore, attention weights were typically dense, meaning that most features will get some level of attention.

Human Evaluation Survey We conducted a small scale human evaluation survey on 20 Ph.D. and Masters students combined with a machine learning background. The survey was conducted in a blind format where the identities of the models were anonymous to ensure that our data was trusted and not biased. The survey asked: *"Which figure do you think is the best method to determine the important features?"*. 65% of respondents preferred InterpreTabnet as the means to highlight salient features. This validated our motivation that sparse feature selection distilled the complexity of the data into a simpler and understandable form for practical applications. InterpreTabNet provided 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. The survey's results can be found in Table 2.[5]

Table 2: Human Evaluation 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

Though the results of our survey were promising, future work could replicate these findings on a larger sample size, and progress beyond asking users for their preference by evaluating the users' ability to interpret model predictions under different interpretability schemes.

4.2. Capturing Feature Interdependencies by Prompting LLMs

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

In order to generate a precise output mapping, instructions

---

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

were provided to GPT-4 where the extracted salient features would be formatted into a dictionary. Each mask corresponded 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 in order 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, 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 improved 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 5.8 and 5.9 respectively.

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...}*

4.3. Justifying GPT-4's Analyzing Abilities of InterpreTabNet's Internal Behavior

A potential concern was whether GPT-4 actually interpreted the model's internal behavior rather than merely rephrasing the prompt input. To address this, we conducted the following three experiments to demonstrate that the model exhibited a genuine understanding of the data.

Definition Check We tested the integrity of our prompt with a definition check on 'feature mask'. Figure 12 in Appendix 5.10 aligned with our expectations, providing an accurate and detailed explanation.

GPT-4 With vs. Without InterpreTabNet We tested if GPT-4 alone without InterpreTabNet could identify the salient features and elucidated their relationships from the Adult dataset. However, as observed in Appendix 5.10 Table 8, GPT-4 was unable to determine the salient features even when the dataset information was provided. For the first prompt, it was unable to compile any aggregate analysis. In the second prompt, the important features that were extracted were not accurate as well as being dense, selecting more than 50% (8/14) of the features.

Trust in GPT-4's Analysis We tried various prompts for interpretability to strengthen the trust in the generated explanations whilst determining which prompt design yielded the most insightful and accurate explanations from GPT-4. We explored explanations in different formats, and varying levels of detail. The results can be found in Tables 9 and 10 of Appendix 5.10. Our analysis indicated that our original prompt structure led to a higher level of detail and attempted to deduce deeper meanings from the prominent features, as opposed to simply categorizing them. This suggested a level of interpretive understanding by GPT-4 that went beyond basic rephrasing, thereby reinforcing the effectiveness of our prompt design.

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 all baselines for most datasets. The salient features from our masks are channelled into GPT-4 via a carefully engineered prompt that outputs an analysis of the features' interdependencies. For practitioners, the 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 are 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.

References

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

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.

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.

C. K. D. J. Clore, John 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.

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. Hao, Y. Gu, H. Ma, J. J. Hong, Z. Wang, D. Z. Wang, and Z. Hu. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.

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.

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

OpenAI. Gpt-4 technical report, 2023.

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, Jan 2015. URL https://papers.nips.cc/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html.

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.

Appendix

5.1. Proof: cVAE Evidence Lower Bound

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

using Bayes' rule

$$\begin{aligned}
&= E[\log Q(z|Y,X) - \log\frac{P(z,Y,X)}{P(Y,X)}] \\
&= E[\log Q(z|Y,X) - \log\frac{P(Y|z,X)P(z,X)}{P(Y,X)}] \\
&= E[\log Q(z|Y,X) - \log\frac{P(Y|z,X)P(z|X)P(X)}{P(Y,X)}] \\
&= E[\log Q(z|Y,X) - \log\frac{P(Y|z,X)P(z|X)P(X)}{P(Y|X)P(X)}] \\
&= E[\log Q(z|Y,X) - \log\frac{P(Y|z,X)P(z|X)}{P(Y|X)}] \\
&= E[\log Q(z|Y,X) - (\log P(Y|z,X) + \log P(z|X) - \log P(Y|X))] \\
&= E[\log Q(z|Y,X) - \log P(Y|z,X) - \log P(z|X) + \log P(Y|X)] \\
&= E[\log Q(z|Y,X) - \log P(Y|z,X) - \log P(z|X)] + \log P(Y|X)
\end{aligned}$$

$$D_{KL}[Q(z|Y,X)||P(z|Y,X)] - \log P(Y|X) = E[\log Q(z|Y,X) - \log P(Y|z,X) - \log P(z|X)]$$

rearranging the sign to rewrite RHS as another KL Divergence

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

which is the cVAE objective function

5.2. 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: $\alpha$: Starting range (start)
Require: $\beta$: Ending range (end)
Require: $\delta$: Percentage of feature importance captured by one feature in each feature mask (col_threshold_val)
Require: $\gamma$: Number of columns that satisfies $\delta$ in each feature mask (col_threshold)
Require: $\iota$: Number of complete-feature masks that passes the algorithm's feature selection criteria (all_mask_pass)
Require: $\epsilon$: Threshold for the number of complete-feature masks that passes the algorithm's feature selection criteria
    (all_mask_pass_thresh)
Require: $\zeta$: Step size computed using a logarithmic scale at high levels (step_size)
Require: $\theta$: Dictionary storing $r_M$-accuracy pairs (reg_m_acc_dict)
Require: $\lambda$: Flag for recursion (is_recursive)
Ensure: Optimal regularization parameter $r_M^*$

Initialize $\theta$ if $\theta$ is None.
Initialize $\iota$ if $\iota$ 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 $\theta$ and $\iota$ are met then
        Update $\theta$, $\iota$
    end if
    if $\lambda$ 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
```

5.3. Reproducibility

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 5.4 of the paper.

5.4. 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 4, and download links in Table 5. 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 4: 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 5: 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`

5.5. Hyperparameters Search Space

We provided hyperparameter search spaces for all models in Table 6. 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).

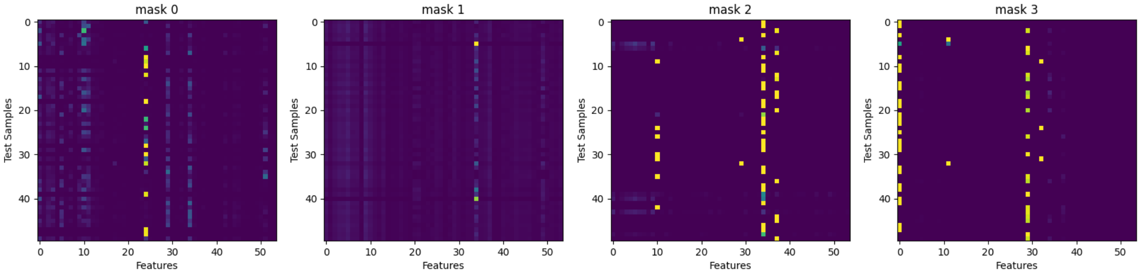
Table 6: Hyperparameter spaces for all models

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], λ: [0.001, 0.01, 0.1, 0.3], 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], λ: [0.001, 0.01, 0.1, 0.3], Learning Rate: [0.005, 0.01, 0.02, 0.025], r_M: range from [0, 1,000,000,000,000]
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

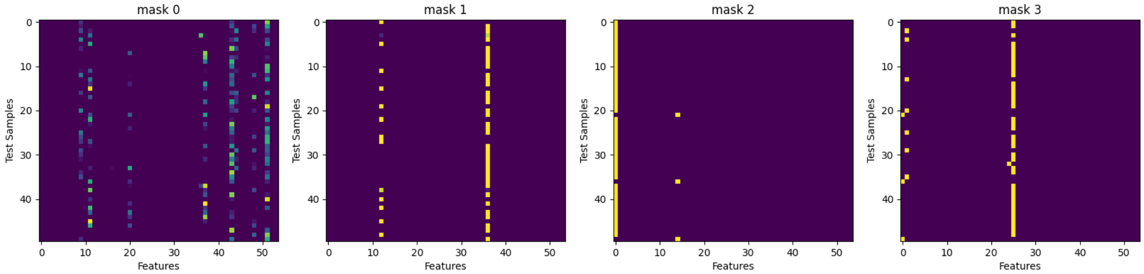
5.6. Results from other datasets (Accuracies & Masks)

FOREST COVER TYPE (DUA AND GRAFF, 2017)

Model	**Test Accuracy (%)**
XGBoost	92.30
LightGBM	86.38
TabTransformer	82.55
MLP	79.76
Original TabNet	94.18
InterpreTabNet $(r_M^* = 900)$	**94.75**



(a) InterpreTabNet

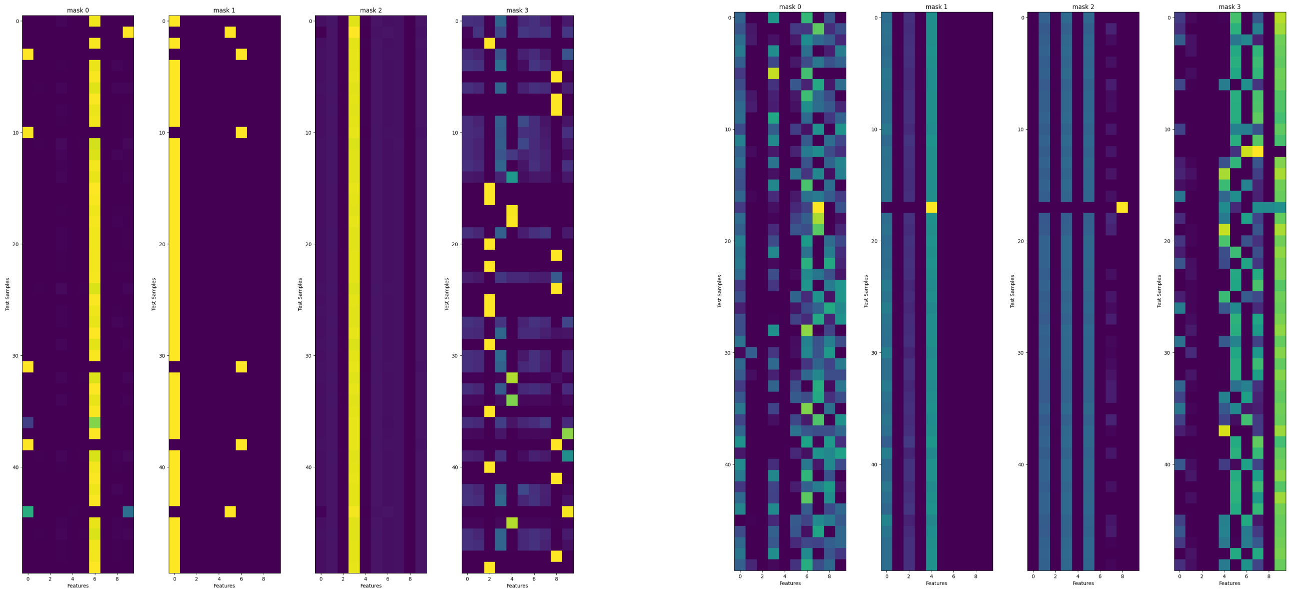


(b) Original TabNet Model

Figure 5: (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 (%)**
XGBoost	75.57
LightGBM	78.47
TabTransformer	99.81
MLP	99.70
Original TabNet	99.00
InterpreTabNet ($r^*_M = 1000$)	**99.50**



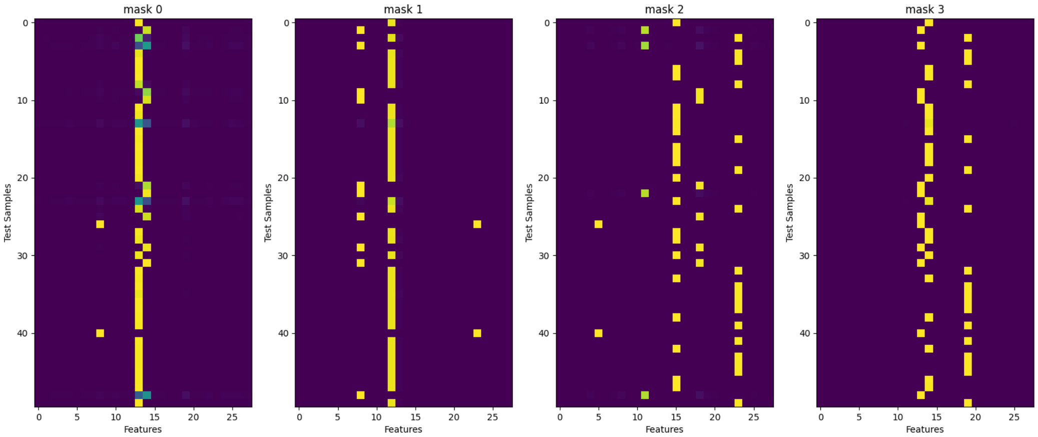
(a) InterpreTabNet

(b) Original TabNet Model

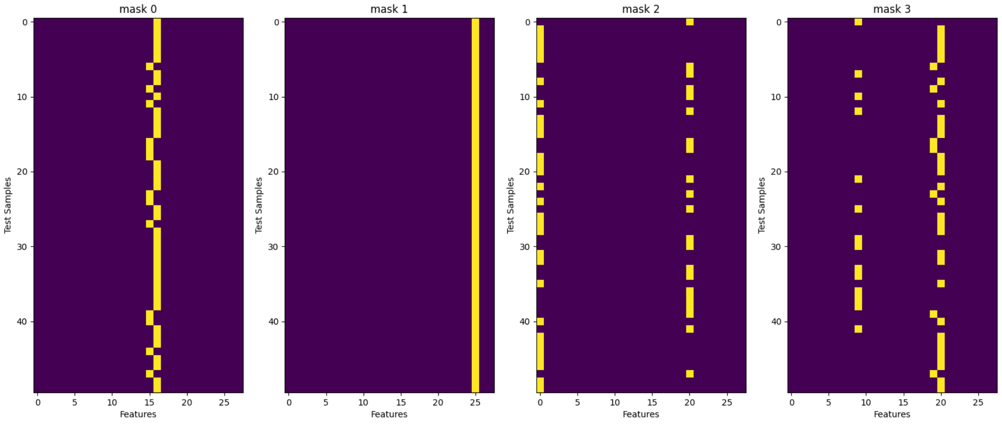
Figure 6: (a) Best performing model using InterpreTabNet $r^*_M = 1000$ with an accuracy of 99.13% on the Poker Hand Dataset. (b) The baseline performance using the Original TabNet model, attaining an accuracy of 99.23%.

Higgs (Whiteson, 2014)

Model	**Test Accuracy (%)**
XGBoost	72.91
LightGBM	72.62
TabTransformer	51.97
MLP	63.17
Original TabNet	52.94
InterpreTabNet $(r_M^* = 10000)$	**53.08**



(a) InterpreTabNet

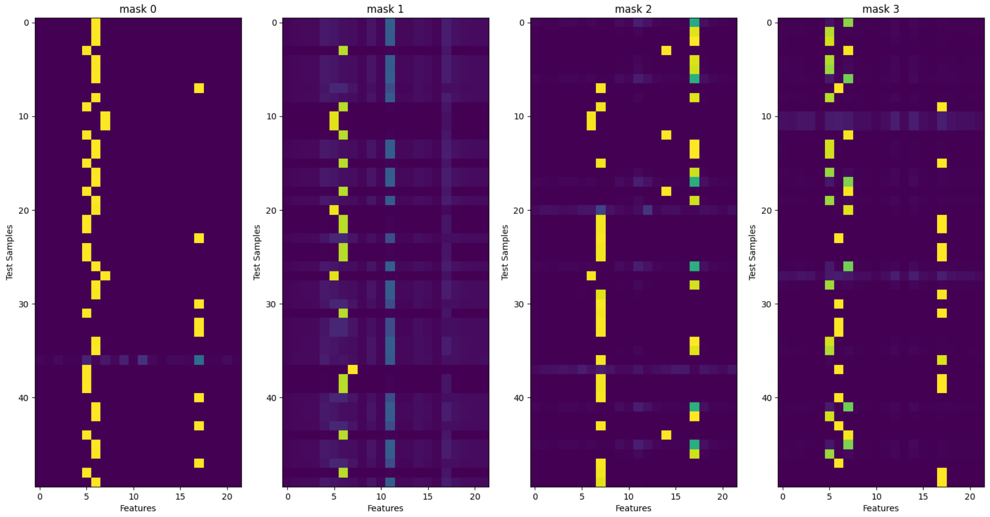


(b) Original TabNet Model

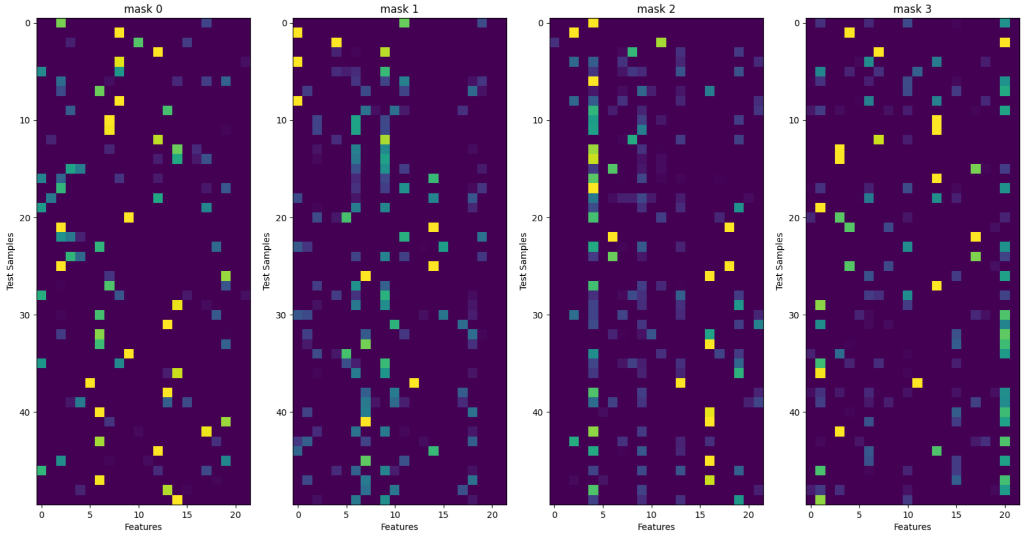
Figure 7: (a) Best performing model using InterpreTabNet $r_M^* = 10000$ with an accuracy of 53.08% on the Higgs Dataset. (b) The baseline performance using the Original TabNet model, attaining an accuracy of 60.22%.

Mushroom (mus, 1987)

Model	**Test Accuracy (%)**
XGBoost	99.69
LightGBM	100.00
TabTransformer	100.00
MLP	99.82
Original TabNet	99.94
InterpreTabNet ($r^*_M = 10,000,000,000,000$)	**96.62**



(a) InterpreTabNet

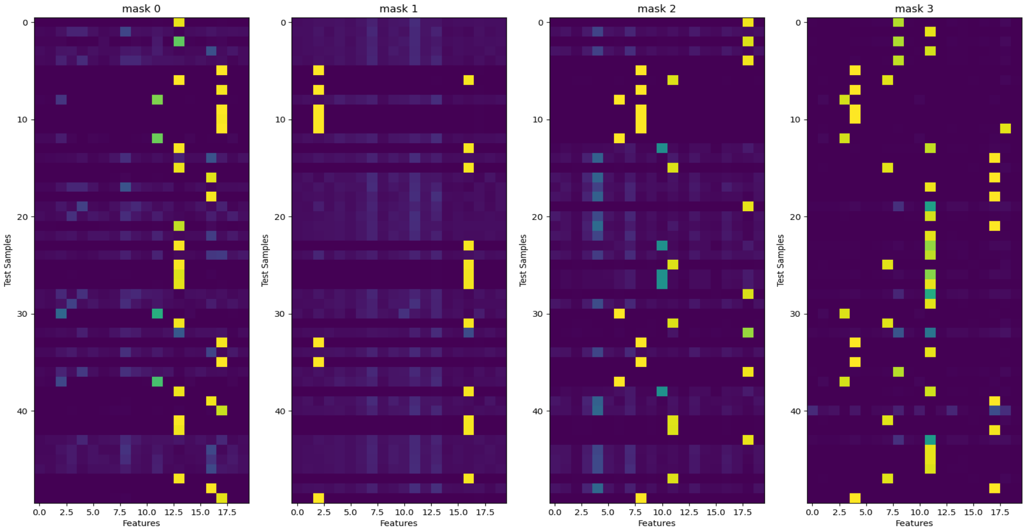


(b) Original TabNet Model

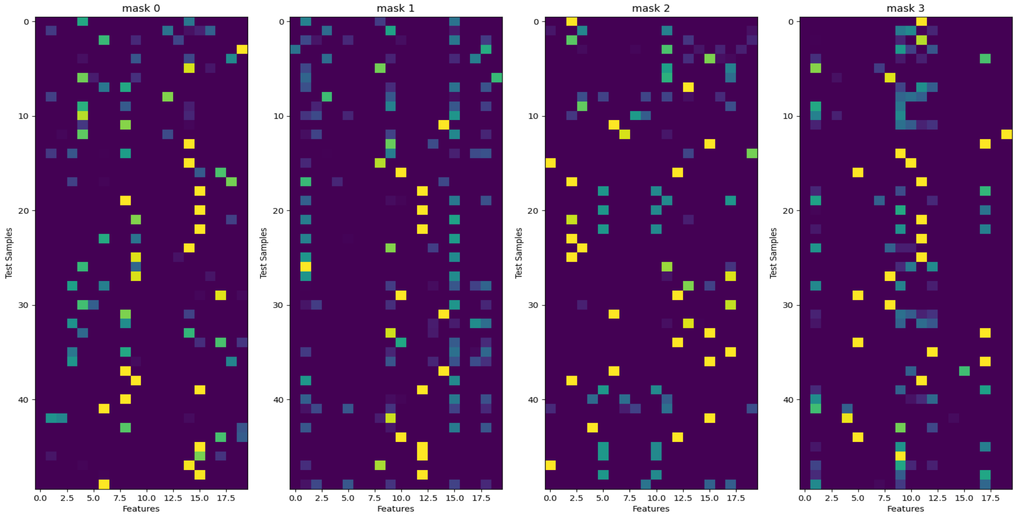
Figure 8: (a) Best performing model using InterpreTabNet $r^*_M = 1,000,000,000,0000$ with an accuracy of 96.62% on the Mushroom Dataset. (b) The baseline performance using the Original TabNet model, attaining an accuracy of 99.94%.

BLASTCHAR (BLASTCHAR, 2018)

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



(a) InterpreTabNet

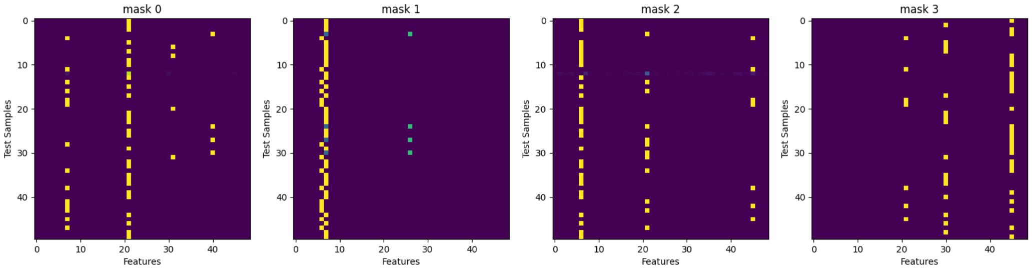


(b) Original TabNet Model

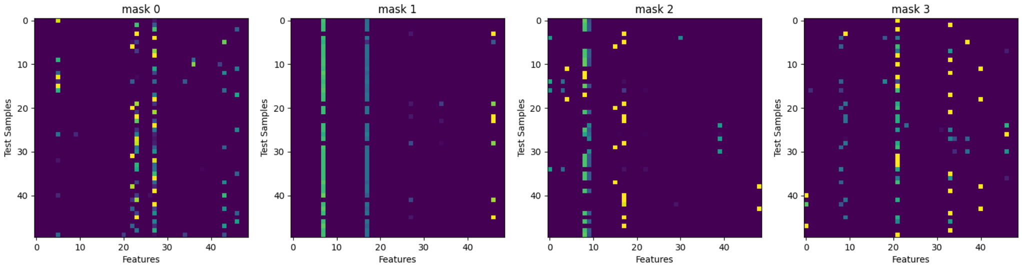
Figure 9: (a) Best performing model using InterpreTabNet $r_M^* = 10,000,000,000,000$ with an accuracy of 72.96% on the Blastchar Dataset. (b) The baseline performance using the Original TabNet model, attaining an accuracy of 76.22%.

DIABETES (CLORE AND STRACK, 2014)

Model	**Test Accuracy (%)**
XGBoost	61.44
LightGBM	60.87
TabTransformer	44.45
MLP	53.99
Original TabNet	56.91
InterpreTabNet $(r_M^* = 100,000,000)$	**55.37**



(a) InterpreTabNet

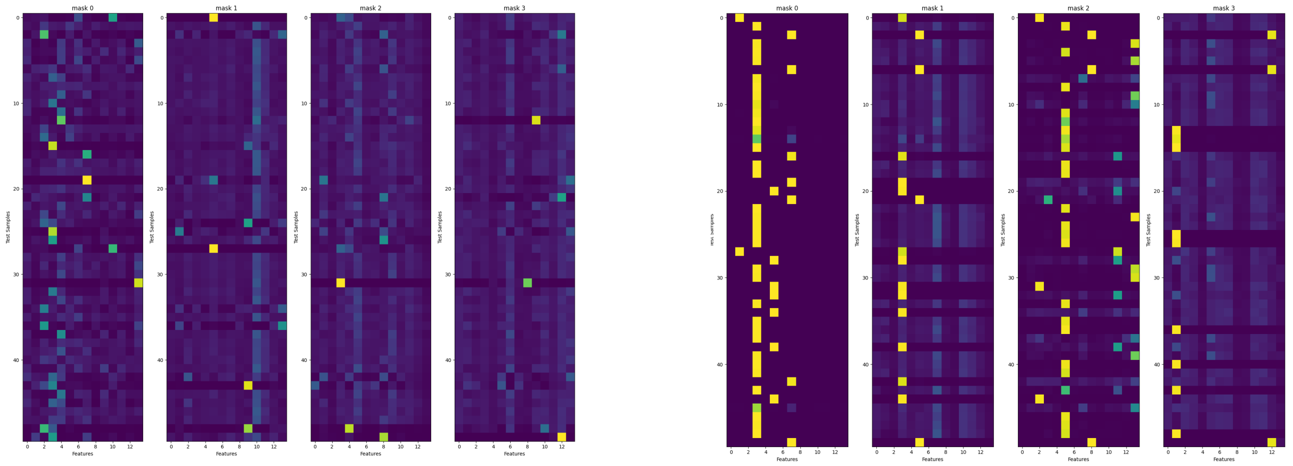


(b) Original TabNet Model

Figure 10: (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%.

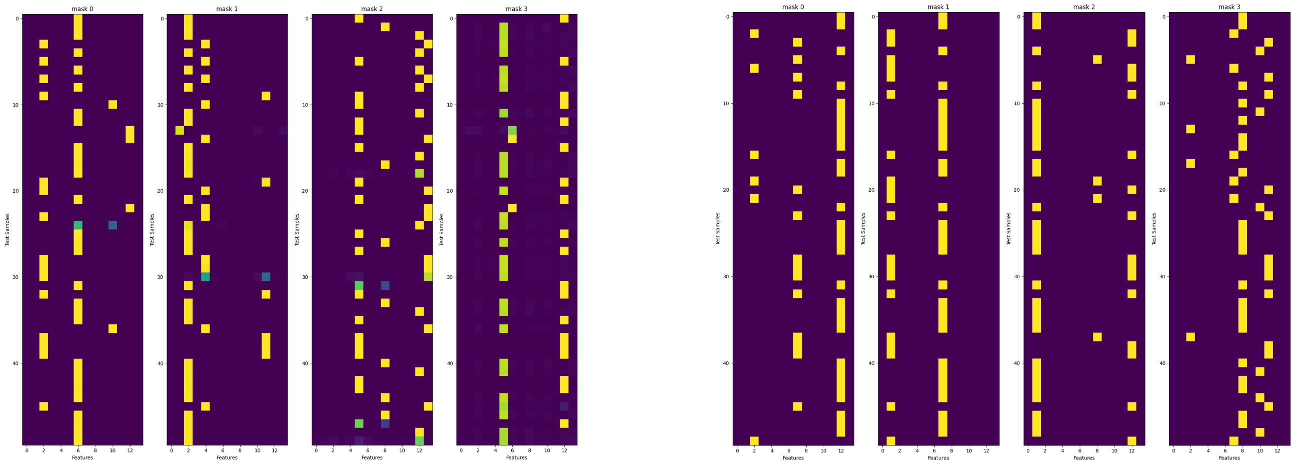
5.7. Ablation Study on InterpreTabNet's Interpretability for varying r_M values

ADULT CENSUS INCOME MASK FIGURES



(a) InterpreTabNet $(r_M = 0)$

(b) InterpreTabNet $(r_M = 1)$



(c) InterpreTabNet $(r_M = 100)$

(d) InterpreTabNet $(r^*_M = 1000)$

Figure 11: As the r_M value increases, both feature mask sparsity and feature importance increases, improving the interpretability of the masks.

5.8. 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."}

5.9. Output text from GPT-4 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."}

5.10. GPT-4 Interpretability Sanity Check

Table 8: GPT-4 Analysis without InterpreTabNet

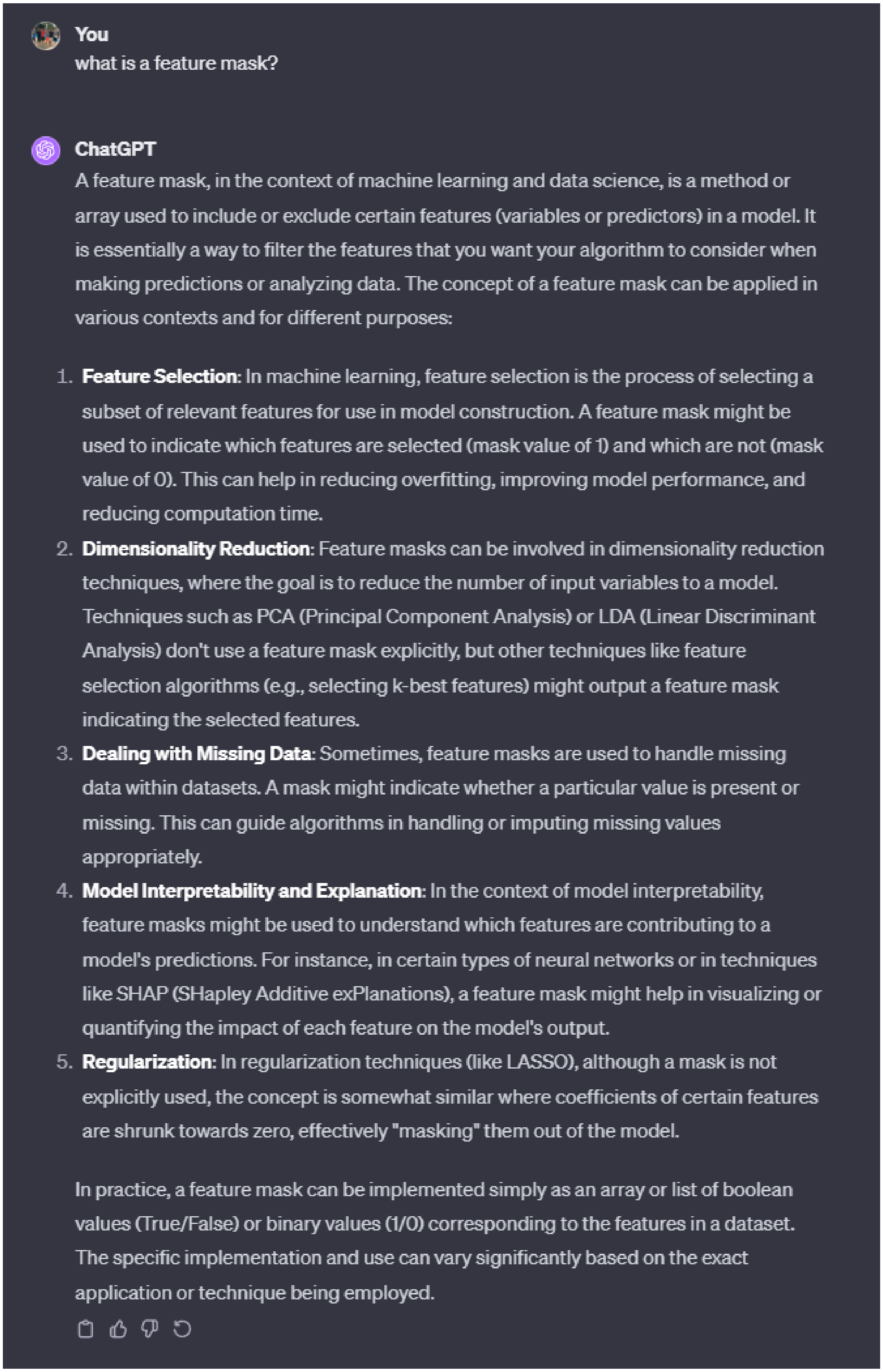
Prompts	**Outputs**
Identify the most important features of the following dataset and conduct aggregate text analysis on it. Please output ONLY a dictionary and no other natural language generation when generating the sentence. The Adult Census Income dataset is considered which predicts whether an individual's 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"] }

Table 9: 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 10: 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"]}*



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 12: Feature Mask Definition Check