ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

TabSRA: An Attention based Self-Explainable Model for Tabular Learning

Kodjo Mawuena Amekoe[1,3], Mohamed Djallel Dilmi [1,2], Hanane Azzag[1], Mustapha Lebbah[1,2], Zaineb Chelly Dagdia[2] and Gregoire Jaffre[3]

1- Sorbonne Paris Nord University - LIPN, UMR CNRS 7030, Villetaneuse, France

2- Paris-Saclay University - DAVID Lab, UVSQ, Versailles, France

3- Groupe BPCE, 7 promenade Germaine Sablon 75013 Paris, France

Abstract. We propose TabSRA, a novel self-explainable, and accurate model for tabular learning. TabSRA is based on SRA (Self-Reinforcement Attention), new attention mechanism that helps to learn an intelligible representation of the raw input data through element-wise vector multiplication. The learned representation is aggregated by a highly transparent function (e.g linear), which produces the final output. Experimental results on synthetic and real-world classification problems show that the proposed TabSRA solution outperforms existing widely used self-explainable models and performs comparably to full complexity state-of-the-art models in term of accuracy while providing a faithful feature attribution. Source code is available at `https://github.com/anselmeamekoe/TabSRA`.

1 Introduction

The success of Tree-based Boosting models such as XGBoost [3], LightGBM [7] or Attention-based solutions such as TabTransformer[5], FT-Transformer [4] on supervised learning Benckmark datasets and Kaggle competition motivates practitioners to use them as an alternative to classical statistical models (e.g. Linear Regression, Generalised Additive Models), especially when the data exhibit strong feature interactions. These full-complexity models typically use a large number of parameters (or trees), making them difficult for direct human inspection. On the other hand, interpretability is usually (i) required by regulators in real-world applications (e.g., GDPR: Article 22 in Europe), (ii) desired if the goal is to discover hidden patterns in the data (e.g., fraud detection) or to ensure that the model does not learn a bias that may lead to significant drift once in production. Therefore, recent research such as [10, 11, 9] has focused on developing post-hoc methods to explain, at least locally, the predictions of full complexity models. Unfortunately, although these methods provide some interesting properties, they are sometimes based on some computational mechanisms (e.g., exact Shapley value computation) or hypotheses (e.g., independence between features) that are difficult to achieve in practice, leading to biased explanations [1, 8]. Still discussing interpretability, [12] provides a technical reason why an interpretable model might exist among the set of accurate models in any domain and encourages researchers to move toward finding this solution, especially for high-risk domains. Convinced by this philosophy and with a special focus on tabular data modeling, we propose Self-Reinforcement Attention for

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

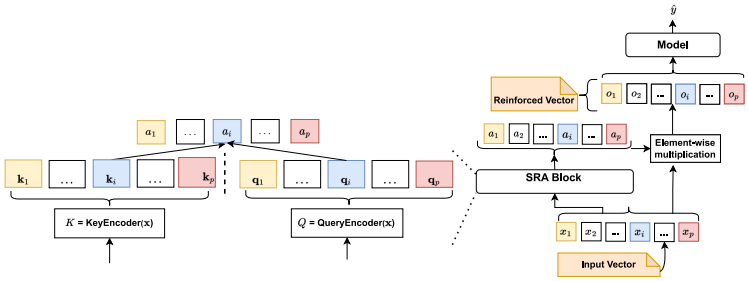


Fig. 1: TabSRA model. The attention vector $\mathbf{a} = (a_1, ..., a_p) \in \Re^p$ provided by the SRA block is used to produce a *reinforced* vector $\mathbf{o} = (o_1, ..., o_p) \in \Re^p$.

tabular data (TabSRA).
We summarize our contributions as follows: (**1**) TabSRA is an attention-based supervised model that provides an intrinsic explanation of its predictions and is trained in an end-to-end manner using back-propagation. (**2**) It contains a Self-Reinforcement Attention (SRA) block that is used to learn a *Reinforced* representation of the raw input through element-wise multiplication with the produce attention vector. This consideration allows to: (i) take into account possible interactions without unecessarily adding additional features (terms) or imposing a limit on the order of interactions between features; (ii) a global model understanding, especially using visualisation. The attention computation using SRA is completely different from that proposed by the original work [13] (or used in [5]) and preserves the input dimension. (**3**) Our experiments show that our proposed solution provides understandable representations while being accurate compare to state-of-the-art models.

2 Model Architecture

The challenge in most supervised tabular learning problems using attention mechanism [5, 4] is to estimate the output $\hat{y} = f_\theta(\mathbf{x})$ given the feature vector $\mathbf{x} = (x_1, ..., x_p) \in R^p$. The parametric model f_θ is learned using the training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $y_i \in \{0, 1\}$ for binary classification or $y_i \in \Re$ for regression tasks. Our proposed TabSRA model f_θ (Fig 1) contains a SRA block which is a novel attention mechanism layer denoted as a function $a(.)$. Given the raw input $\mathbf{x}$, the SRA block produces an attention vector $\mathbf{a} = (a_1, ..., a_i, ...a_p)$. Thereafter the attention vector is used to learn an intelligible representation $\mathbf{o} = (o_1, ..., o_i, ..., o_p)$ as $\mathbf{o} = \mathbf{a} \odot \mathbf{x}$, where $\odot$ is the element-wise multiplication.

For comprehension purposes, we propose in this paper a combination of the SRA block with a linear aggregation model (Fig 1). This instantiation of the TabSRALinear model can be formalized as follows:

$$g(\hat{y}) = \beta_1 a_1 x_1 + ... + \beta_i a_i x_i + ... + \beta_p a_p x_p \quad (1)$$

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Algorithm 1: PyTorch-style forward pass pseudocode of the TabSRA Block

```
# b is batch size, p the number of features
def forward(self, x):
   Q = self.KeyEncoder(x) # Q is (b, p, d_k)
   K = self.QueryEncoder(x) # K is (b, p, d_k)
   QK = Q*K*self.scale # scale= 1/d_k, QK is (b, p, d_k)
   a = QK.sum(axis = -1) # a is (b, p)
   return a
```

$\beta_i a_i x_i$ represents the contribution of the feature x_i to the output, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ is the linear regression coefficients and a_i is interpreted as the amplification (or the correction) that the feature x_i received from other features or itself due to the interactions. g represents the link function (e.g., usually $g(\mu) = \log(\frac{\mu}{1-\mu})$ for binary classification and $g = Identity$ for regression tasks). It is important to point out that TabSRA's attention coefficients are clearly correlated to the model's outputs (or the feature contribution measures). This is actually a desirable property for considering attention as an faithful explanation of predictions[6].

2.1 Self-Reinforcement Attention (SRA block)

Given the input vector $\mathbf{x} = (x_1, ...x_i, ..., x_p) \in \Re^p$, the SRA block encodes it into p keys in $K = [\mathbf{k}_1, \mathbf{k}_2, ..., \mathbf{k}_i, ..., \mathbf{k}_p]^T$ with $\mathbf{k}_i = (k_i^1, ..., k_i^{d_k}) \in \Re^{d_k}$ using the key encoder and queries matrix $Q = [\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_i, ..., \mathbf{q}_p]^T$ with $\mathbf{q}_i = (q_i^1, ..., q_i^{d_k}) \in \Re^{d_k}$ using the query encoder (see the pseudocode provided in Algorithm 1). The matrices of queries (Q) and keys (K) are generated by two separate fully connected feed-forward networks (FFN) namely $QueryEncoder$ and $KeyEncoder$.

The $KeyEncoder$ (resp. $QueryEncoder$) produces directly p keys (resp. queries) using a single FFN instead of using p independent FFNs per feature as in [5]. This embedding should be particularly useful for heterogeneous tabular data, especially in the presence of strong features' interactions and at the same time alleviate the need of using several attention blocks (layers) or extra processing which could affect the interpretability of the attention coefficients. Furthermore, with a Sigmoid activation function, all elements k_i^j of K (resp. q_i^j of Q) are scalar numbers bounded in $[0, 1]$. The keys in K are compared to the queries Q component by component, allowing to quantify the alignment of different transformations of the same input calculating the attention weights $\mathbf{a} = (a_1, .., a_i, ..., a_p)$ as follows :

$$a_i = \frac{\mathbf{q}_i \cdot \mathbf{k}_i}{d_k} \quad \text{for} \quad i \in 1, \cdots, p \tag{2}$$

We further use the scaling by d_k in order to reduce the magnitude of the dot-product and to get dimension-free attention coefficients $a_i \in [0, 1]$. We propose

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Table 1: Benchmark datasets

Datasets	# Samples	# Features	# Cat features	Positive rate	Metric
Bank Churn	10000	10	2	20.37 %	AUCROC
Adult Income	30162	14	8	24.89 %	AUCROC
Telco Churn	66469	63	0	20.92 %	AUCROC
Heloc Fico	10459	23	0	47.81 %	AUCROC
Credit Card Fraud	284807	29	0	0.17%	AUCPR

this attention estimation to produce a concise explanation of the decision process. Indeed, considering the potential internal conflict between the input components (due to the interactions), the attention weights vector **a** may enhance or reduce some components (of the input vector) at strategic and specific positions.

3 Experiments

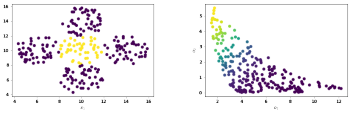
We validate our TabSRA proposition considering the: (i) **Intelligibility**: Are the reinforced representations learned using SRA block understandable? (ii) **Accuracy**: What is the accuracy of TabSRA compared well known transparent and full complexity models (Linear/Logistic Regression, TabNet[2], MultiLayer Perception, XGBoost[3])?

Model setup. We use the same architecture for the key and query encoders which is two hidden layers (ReLu fully connected neural network of dimension $\{d_1, d_2\}$ with, $d_1 = p \times (d_k/4)$ and $d_2 = p \times (d_k/2)$, $d_k \geq 4$. To increase the generalization power, we used regularization in the TabSRA block. Specifically, we used dropout in both the key and query encoders during the training. Also, we used weight decay (L_2 penalization) to empower the smoothness in the embeddings (of the key and query). As we focus particularly on finance as an application domain, we considered well-known benchmark datasets summarized in Table 1. Unless otherwise specified, all categorical inputs are one-hot encoded, and numerical inputs are scaled using the mean and standard deviation to accelerate the convergence of the algorithms.

3.1 Qualitative Results

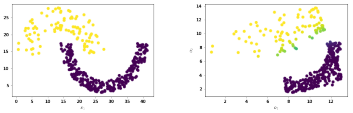
To illustrate how the raw data is reinforced using the TabSRA, we use 2D toy classification datasets where yellow color is used for the class of interest (Fig 4,6,2,7,5,3). We can notice that through multiplication with attention weights a new representation, space is learned (the input dimension is preserved) where data points are much easier to separate using post-hoc model which is a simple linear model in this work. Moreover, this representation helps to understand the global behavior of the TabSRA based model, where it is confident in predicting class 1 (in yellow color) and where it is less confident highlighted by the green color.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.



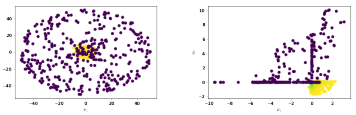
(a) $\mathbf{x}$ (b) $\mathbf{o} = \mathbf{a} \odot \mathbf{x}$

Fig. 2: Five sphere #250



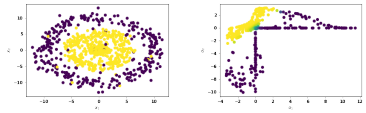
(a) $\mathbf{x}$ (b) $\mathbf{o} = \mathbf{a} \odot \mathbf{x}$

Fig. 3: Two moon with #373



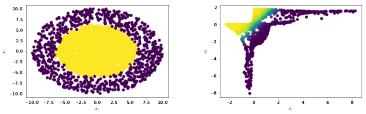
(a) $\mathbf{x}$ (b) $\mathbf{o} = \mathbf{a} \odot \mathbf{x}$

Fig. 4: Two disks #800



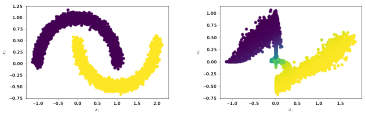
(a) $\mathbf{x}$ (b) $\mathbf{o} = \mathbf{a} \odot \mathbf{x}$

Fig. 5: Rings #1000



(a) $\mathbf{x}$ (b) $\mathbf{o} = \mathbf{a} \odot \mathbf{x}$

Fig. 6: Dense disk #3000



(a) $\mathbf{x}$ (b) $\mathbf{o} = \mathbf{a} \odot \mathbf{x}$

Fig. 7: Noisy two moon #10000

3.2 Quantitative Results

We compare the accuracy achieved by the TabSRALinear model on benchmark datasets relative to baseline models (interpretable and non-counterparts). As shown in Table 2, TabSRALinear achieved the best performance in 4/5 cases among self-explainable models (over TabNet, LR). Furthermore, the performance obtained is often close to one of the overall best-performing models, which is XGBoost (for 4/5 benchmark datasets). These results confirm the effectiveness of the SRA block particularly when observing the difference in performances between the Logistic Regression (LR) and TabSRALinear which ranges from +0.09 for the AdultIncome dataset to +10.05 AUC for the Bank Churn dataset. We recall that LR model is the resulting architecture when removing the SRA block or setting attention weights to 1 (cf. Fig 1).

4 Conclusion

In this paper, have proposed TabSRA, a new self-explainable tabular learning model based on the Reinforcement attention mechanism. The experimental studies confirm our proposition as an accurate and intelligible model. Future work will include more investigation to combine TabSRA with other models (tree based model, boosting mechanism), and on the robustness in drifting context

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Table 2: Accuracy of the TabSRALinear. Means and standard deviations are reported from 5-Fold cross-validation. Italic highlights the best performance. We consider two hidden layers MLP model of dimensions $\{4 \times p, 2 \times p\}$ as in [5].

Datasets	LR	TabNet	TabSRALinear	MLP	XGBoost
BankChurn	$76.93_{\pm 1.56}$	$\mathbf{86.99}_{\pm 0.79}$	$86.98_{\pm 0.46}$	$\textit{87.08}_{\pm 0.73}$	$86.82_{\pm 0.79}$
AdultIncome	$90.50_{\pm 0.41}$	$90.46_{\pm 0.52}$	$\mathbf{91.07}_{\pm 0.42}$	$91.45_{\pm 0.38}$	$\textit{92.63}_{\pm 0.37}$
TelcoChurn	$88.95_{\pm 0.29}$	$90.45_{\pm 0.33}$	$\mathbf{90.52}_{\pm 0.31}$	$90.54_{\pm 0.28}$	$\textit{91.13}_{\pm 0.37}$
HelocFico	$78.26_{\pm 0.52}$	$79.39_{\pm 0.57}$	$\mathbf{79.43}_{\pm 0.41}$	$79.50_{\pm 0.46}$	$\textit{79.75}_{\pm 0.74}$
CreditCardFraud	$77.08_{\pm 2.59}$	$81.09_{\pm 3.92}$	$\mathbf{86.58}_{\pm 2.81}$	$85.69_{\pm 2.53}$	$86.54_{\pm 2.19}$

particularly using a real word dataset provided by the company that supports this research work.

References

[1] Amoukou, S.I., Salaün, T., Brunel, N.: Accurate shapley values for explaining tree-based models. In: International Conference on Artificial Intelligence and Statistics. pp. 2448–2465. PMLR (2022)

[2] Arik, S.Ö., Pfister, T.: Tabnet: Attentive interpretable tabular learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 6679–6687 (2021)

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

[4] Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A.: Revisiting deep learning models for tabular data. Advances in Neural Information Processing Systems **34**, 18932–18943 (2021)

[5] Huang, X., Khetan, A., Cvitkovic, M., Karnin, Z.: Tabtransformer: Tabular data modeling using contextual embeddings. arXiv preprint arXiv:2012.06678 (2020)

[6] Jain, S., Wallace, B.C.: Attention is not explanation. arXiv preprint arXiv:1902.10186 (2019)

[7] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems **30** (2017)

[8] Kumar, I.E., Venkatasubramanian, S., Scheidegger, C., Friedler, S.: Problems with shapley-value-based explanations as feature importance measures. In: International Conference on Machine Learning. pp. 5491–5500. PMLR (2020)

[9] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: From local explanations to global understanding with explainable ai for trees. Nature machine intelligence **2**(1), 56–67 (2020)

[10] Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)

[11] Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)

[12] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)

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