MATH 6380P FINAL-PROJECT
Interpretability of Deep Learning on Home Credit Default Risk Dataset

Yipai DU (yduaz@connect.ust.hk) and Yongquan QU (yquai@connect.ust.hk)

Introduction

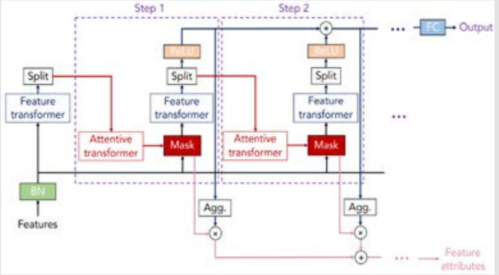
Home Credit uses as much as 221 features to predict if their clients have difficulties to repay their loan. Kagglers prefer to use efficient and interpretable tree-based models as well as complex feature engineering since it's difficult to understand how deep neural networks make predictions, especially when applied to tubular data. In this project, deep feature synthesis (DFS) is utilized to finish automated feature engineering. Based on this, we explore interpretable machine learning models with neural networks, such as TabNet (Arik and Pfister,2020) and Neural Additive Models (NAM, Agarwal et al.2020).

Deep Feature Synthesis

To efficient deal with the provided unstructured data scattered in different csv files, we utilized featuretools for extraction. We used default primitives and set the depth of feature synthesis to 2. Since our focus is not on feature engineering practice to get best possible score, we consider this default setting is sufficient. Highly correlated and non-informative (more than 95% are NAN or same value) features were removed and resulting in a 999 dimensional feature vector for network training.

TabNet

TabNet exploits spatial attention to process tabular data and uses internal information masks to focus the learning capacity of the network on the most informative features in making the decision. The sparsity of the feature masks are controlled through minimizing the mask entropy. When the training phase is done, average mask value throughout the training data can be returned to represented as importance of each feature globally. On the other hand, for each testing data, the corresponding mask value can be interpreted as local feature importance in making the decision. These two combine to give interpretability of the learned model.



Contribution

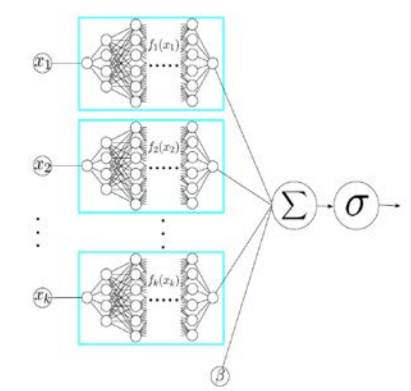
Yipai DU: Feature extraction, TabNet, MLP baseline
Yongquan QU: XGBoost, NAM

Neural Additive Models

NAM belongs to a larger model family called Generalized Additive Models (Hastie and Tibshirani, 1990), with restrictions on the structure of neural networks. GAMS has the form

$$g(E[y]) = \beta + \sum_{i=1}^{K} f_i(x_i)$$

in which shape functions f_i are smooth low-order splines or boosted decision trees. NAMs use deep neural nets to fit GAMs. While NNs with ReLUs are always over-parameterized, exp-centered units (ExU) is proposed by Agarwal et al. to overcome this failure. That is , for a input x, with an activation function f, $h(x)$ is computed by



$$h(x) = f(e^{w} * (x - b))$$

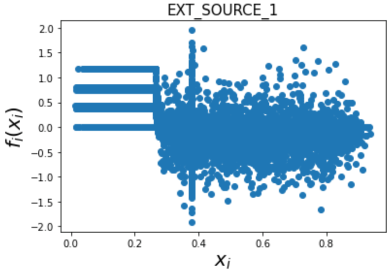
which improves the learnability of NAM when fitting jumpy functions. The interpretability of NAMs comes from the property that they can easily be visualized. In NAMs, each feature is handled independently, so we can visualize the shape functions, i.e. $f_i(x_i)$ vs x_i, to get a full view of NAMs to see how they compute a prediction.

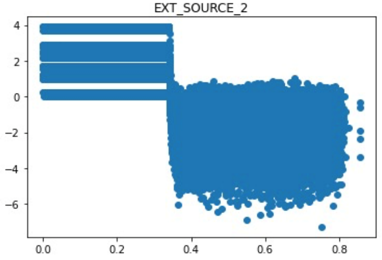
Results

Model	MLP-1	TabNet	MLP-2	NAM-1	NAM-2	XGBoost-1	XGBoost-2
AUC	0.76795	**0.76826**	0.72095	0.73351	0.70365	0.74206	0.70793

MLP-1 uses 4 FC layers to make prediction, achieving similar result with TabNet. NAM-1 uses 999 feature neural networks and one ExU layer and one ReLU-n layer in each feature neural network. TabNet outputs feature importances with 164 nonzero values and we use these features to train MLP-2 ,NAM-2 and XGBoost-2 (same structure with MLP-1 and NAM-1 XGBoost-1). The results are however downgraded.

We use TabNet and XGBoost to assign importance to features, so that we can understand how these models work globally. Top 3 features for TabNet are 'NAME_CASH_LOAN_PURPOSE_Medicine', EXT_SOURCE_1' and 'CHANNEL_TYPE)_Car dealer'. Top 3 features for XGBoost are 'EXT_SOURCE_1', 'EXT_SOURCE_2' and 'EXT_SOURCE_3'. To show the interpretability of NAMs, we plot the shape function of some important features to see exactly how these features contribute to prediction. We can see that the higher extra source score a client has, the less possible that he/she has repaying difficulty.





Conclusion

Although TabNet is able to give sparse global feature importances, we find out only taking in those important features is not a good idea. Unlike what is stated by the authors of TabNet that the model is interpretable by only looking at the importances, the information of features from other dimension in fact influence the masks and therefore is not fully interpretable. Through the visualization we observed partial interpretability of the TabNet model.

NAMs can give an exact description of how they make a prediction. From the results we can see that, NAMs combine the inherent interpretability of GAMs and advantages of deep learning, such as better expressivity. Also, NAMs can be trained on GPUs while other GAMs currently cannot. As NAMs handle features independently and no higher-order features are learned from the combination of input features, a little loss in prediction ability is understandable. Combined with other DL methods, NAMs may achieve better results.