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

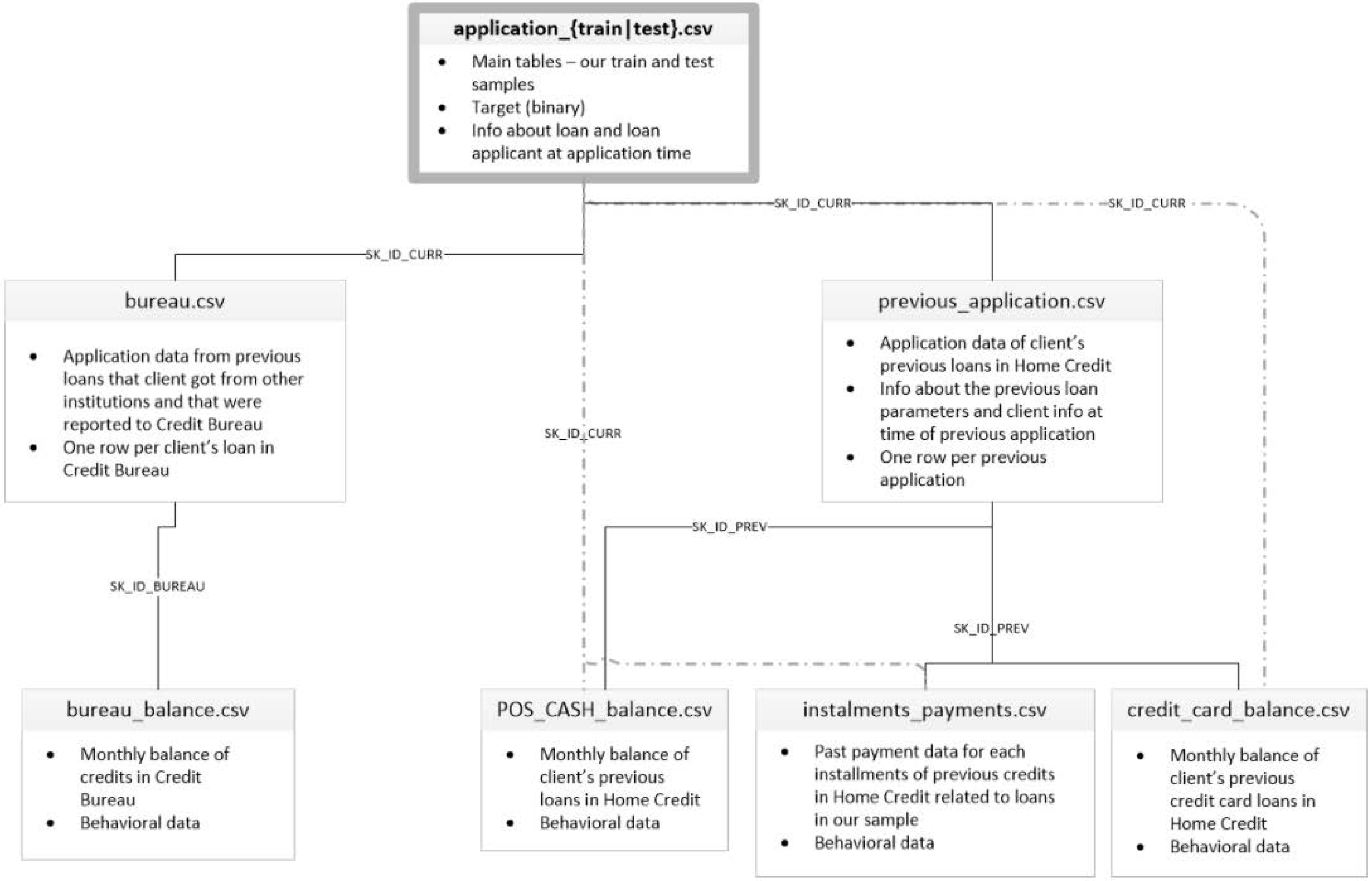
- Fin-Tech opens the future
- Credit risk assessment
- Data rich methods hard to interpret

Outline

- Data processing and feature filtering
- MLP baseline
- TabNet [1]
- XGBoost [2]
- NAM (Neural Additive Model) [3]
- Conclusion

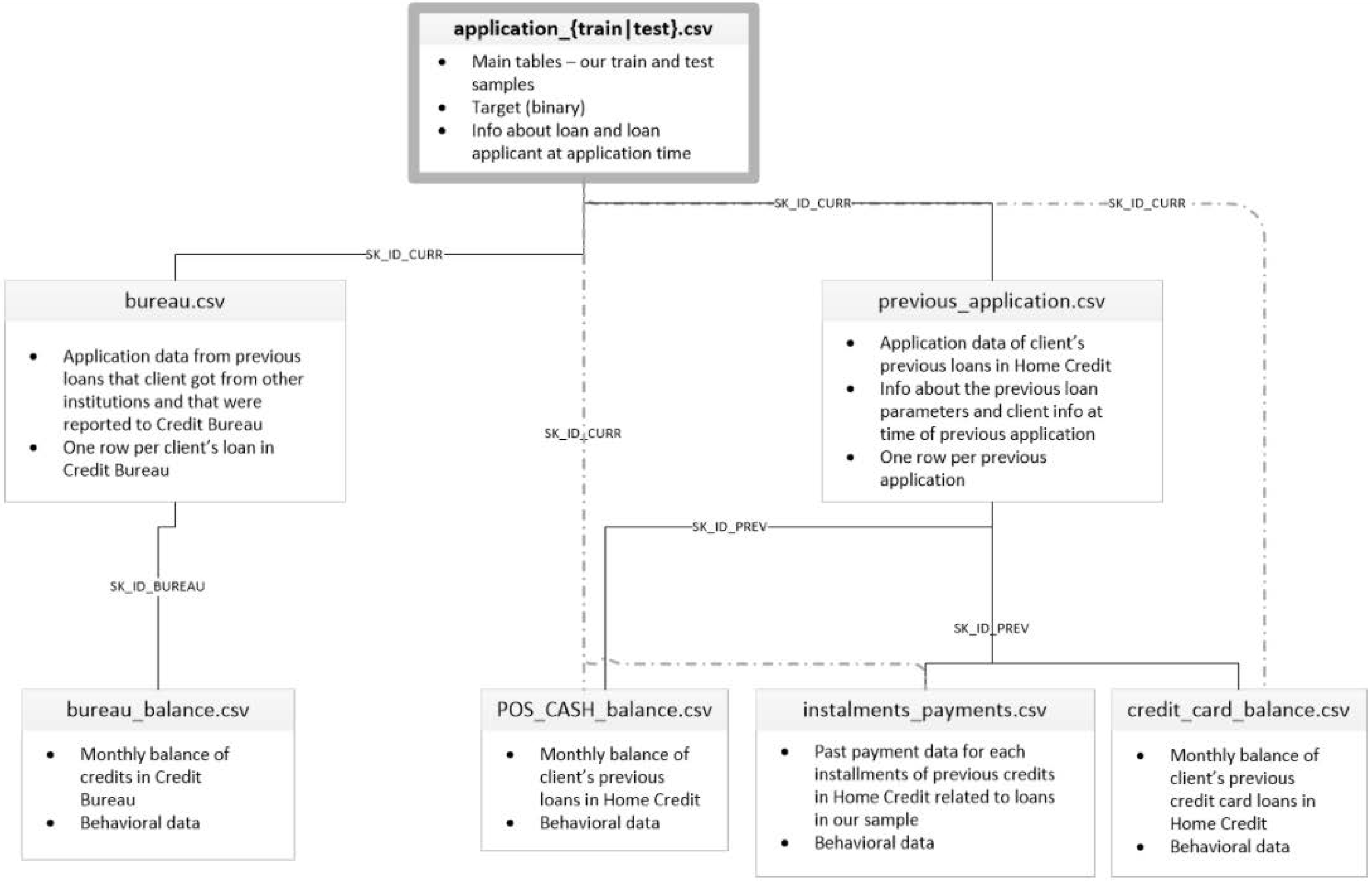
Data processing and feature filtering

- Not well structured
- Multiple files
- Different number of entries
- Use featuretools [4]



Processing with Featuretools

- Create links across files
- Synthesize features with built-in primitives
- Max depth set to 2
- Remove useless features
  - Too many NAN (>95%)
  - Correlated with other features
  - Only 1 possible value
- One-hot encoding for categories
- 999 dimensional features with interpretable meaning

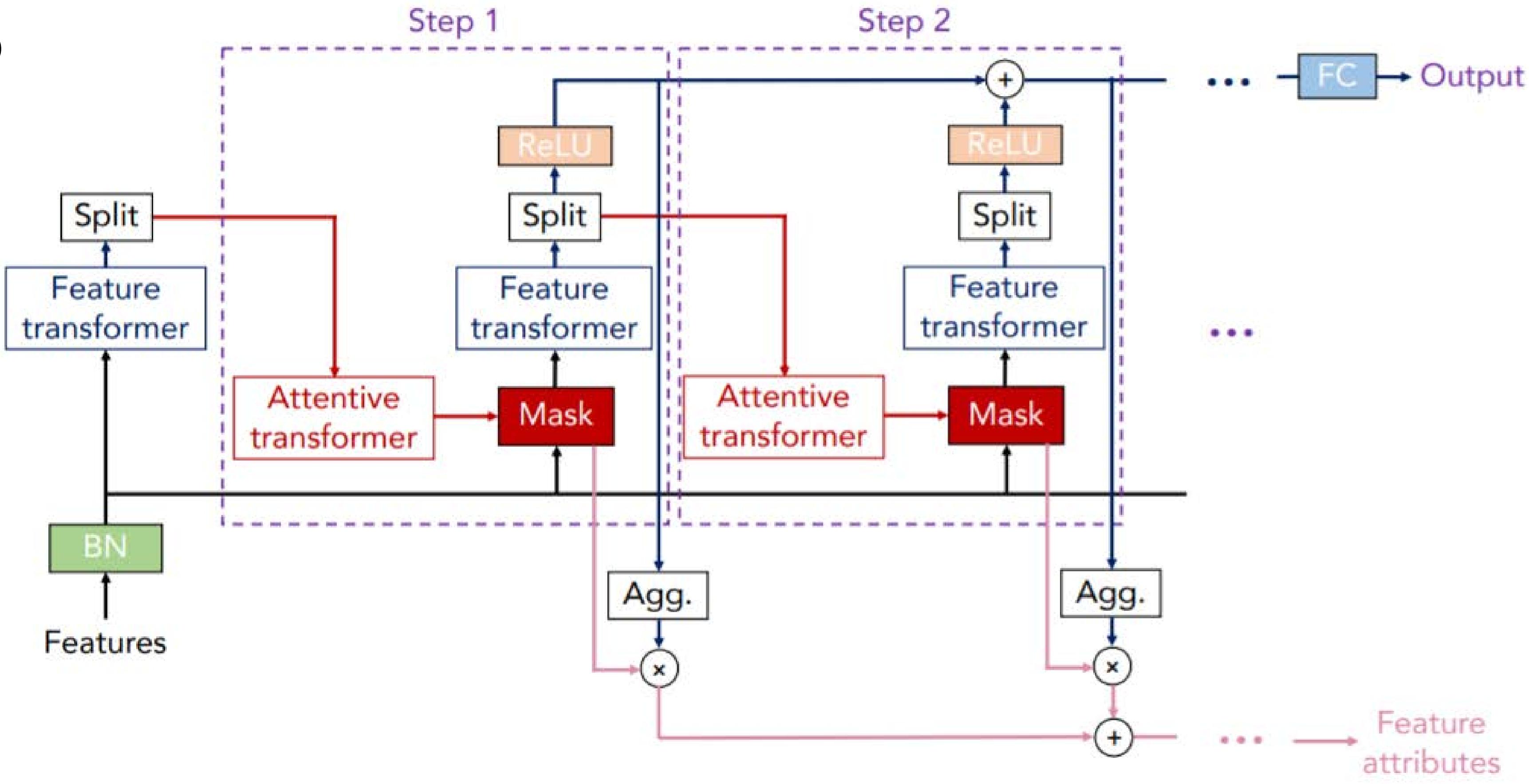


MLP baseline

- 4 layer MLP, fully connected
- 200 hidden units each
- Test AUC 0.76795
- Reference point for other models

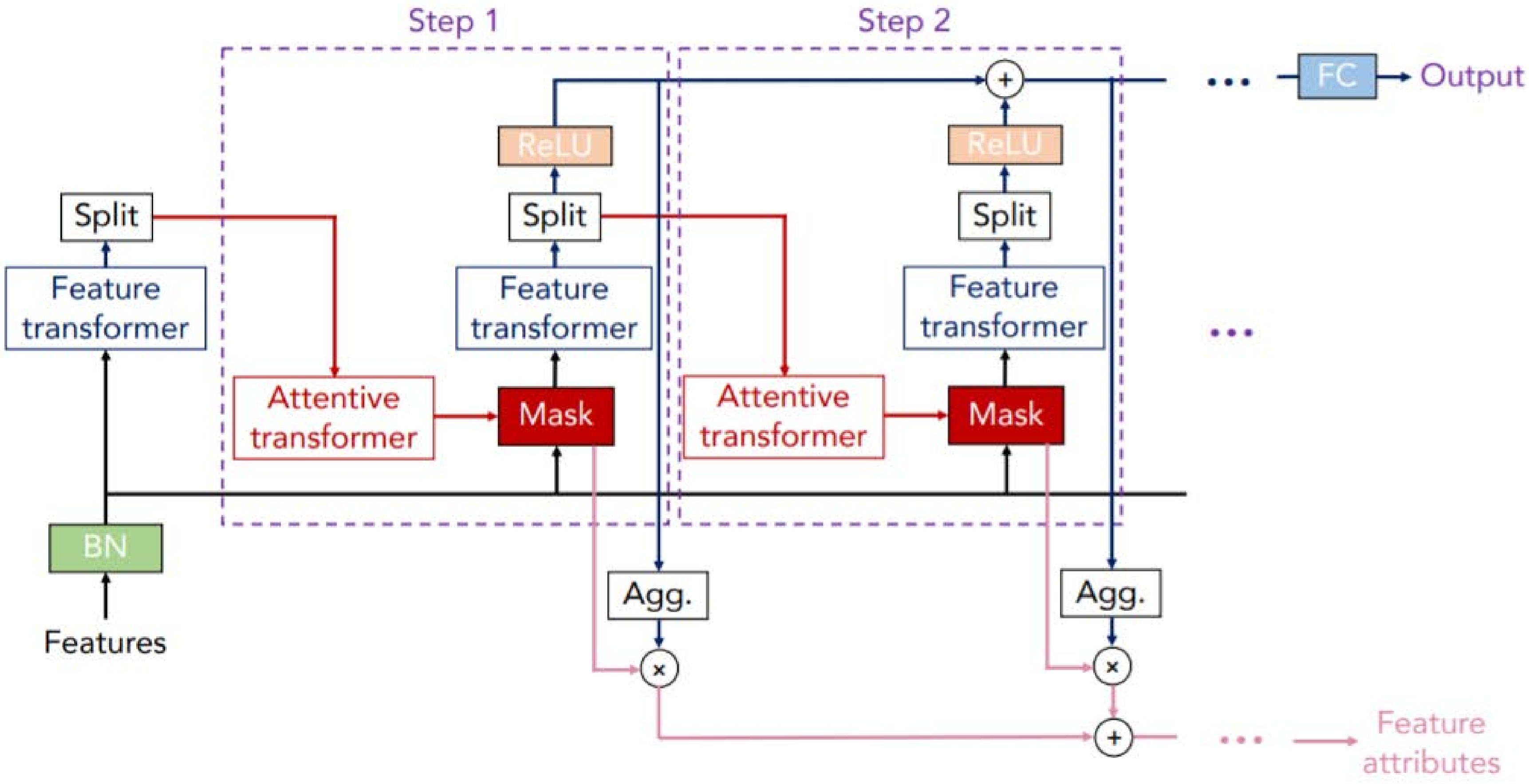
TabNet

- Multiple decision step
- Spatial attention
- Sparse mask through minimal entropy
- Global feature importance
  - Averaging mask across training data
- Local feature importance
  - Mask for new coming data



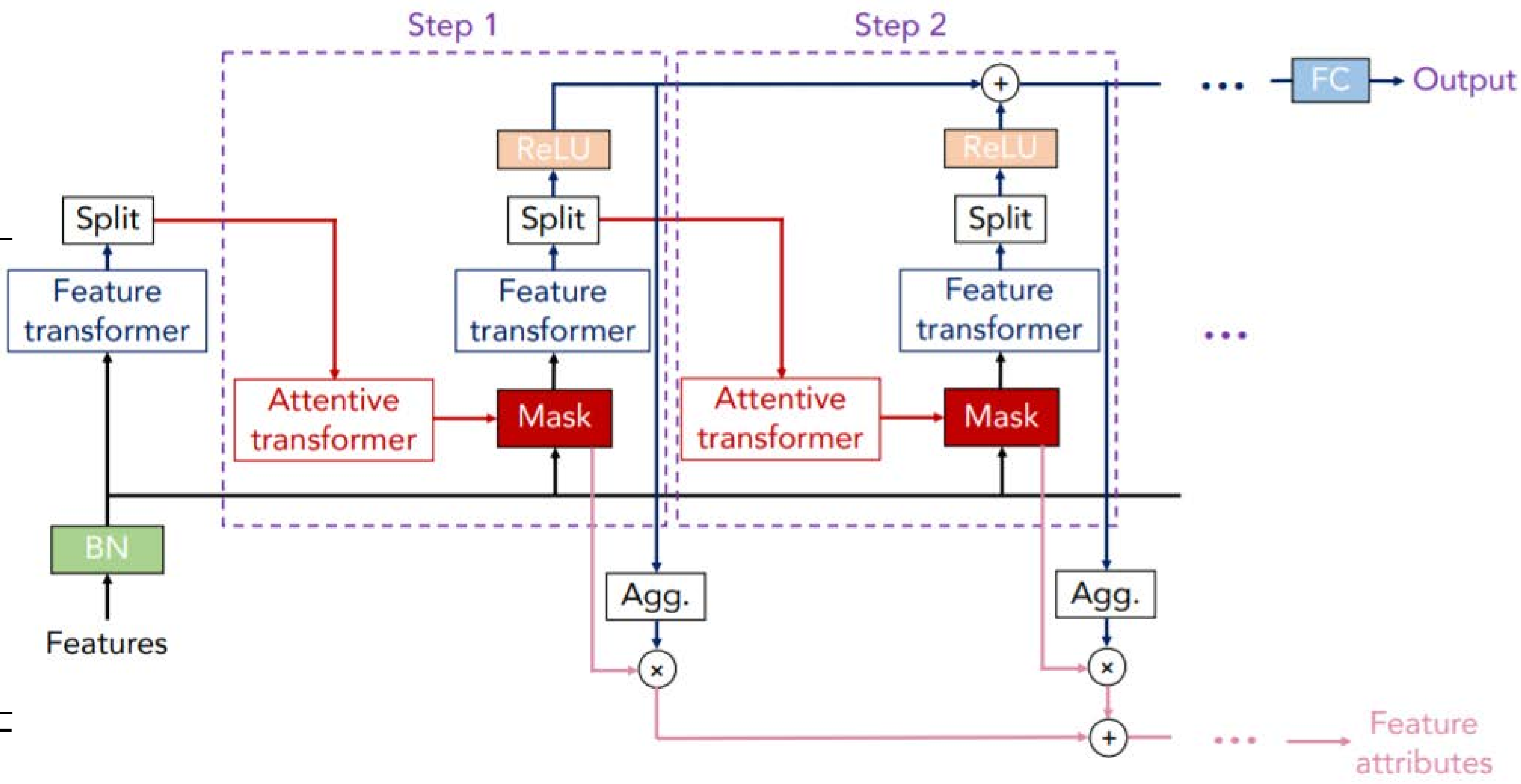
TabNet

- Test AUC 0.76826
- 164/999 nonzero importance features
- Train MLP with the 164 features
  - Test AUC 0.72095
  - Worse than baseline
- Why?
  - Mask dependent on other features
  - Not fully interpretable



TabNet

- Most important features
  - MODE(previous_app.NAME_CASH_LOAN_PURPOSE)_Medicine
  - EXT_SOURCE_1
  - MODE(previous_app.CHANNEL_TYPE)_Car dealer

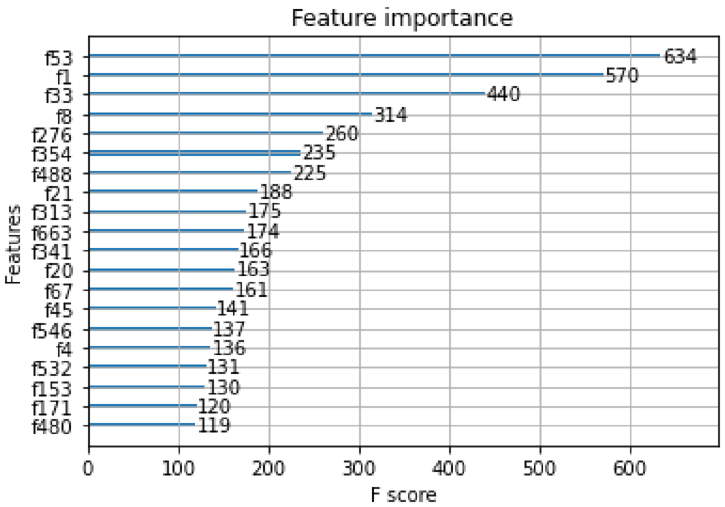


XGBoost

- eXtreme Gradient Boosting: one of the most popular machine learning algorithms on Kaggle challenges

- AUC: 0.74206 on 999 features
  0.70793 on TabNet selected features

- Classic global feature importance measurement
  Top 3: 'EXT_SOURCE_1'
  'EXT_SOURCE_2'
  'EXT_SOURCE_3'



Accuracy vs Interpretability

- Deep Neural Networks
  - Powerful function approximator
  - difficult to understand

- Generalized Additive Models

$$g(\mathbb{E}[y]) = \beta + f_1(x_1) + f_2(x_2) + \cdots + f_K(x_K)$$

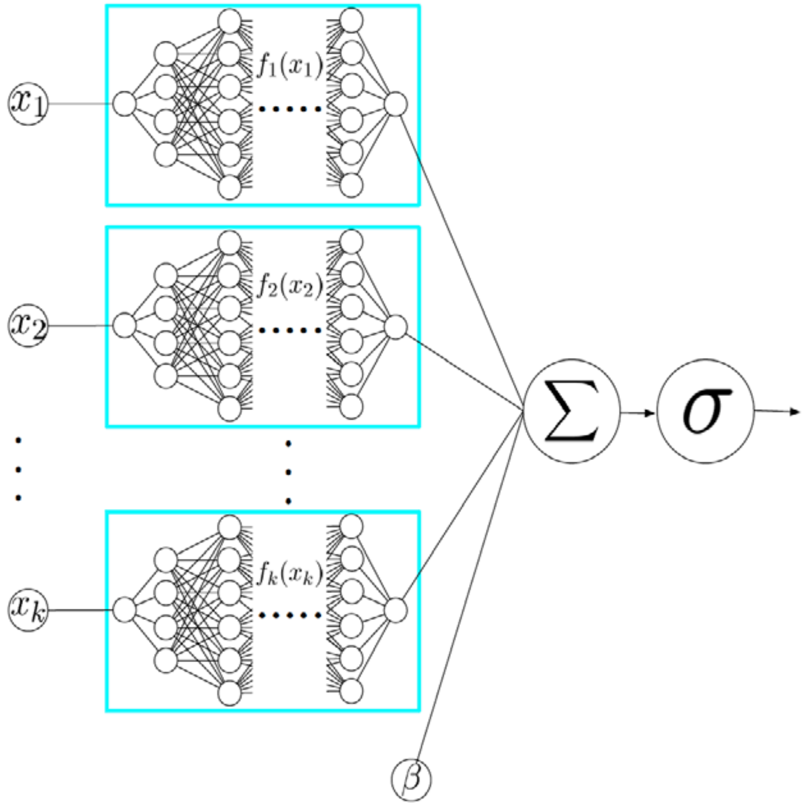
  - Interpretable
  - Splines based: underfitting
  - Tree based: computationally expensive

NAM (Neural Additive Model): Structure

- Each f_i is parameterized by a neural network
- Exp-Centered Units (ExU)

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

· Encourage learning highly jagged curves without affecting global behavior.



NAM (Neural Additive Model): Advantages

- Combined with other deep learning methods
- Easily to extended
- Can be trained on GPUs
- Less ensemble of neural nets required

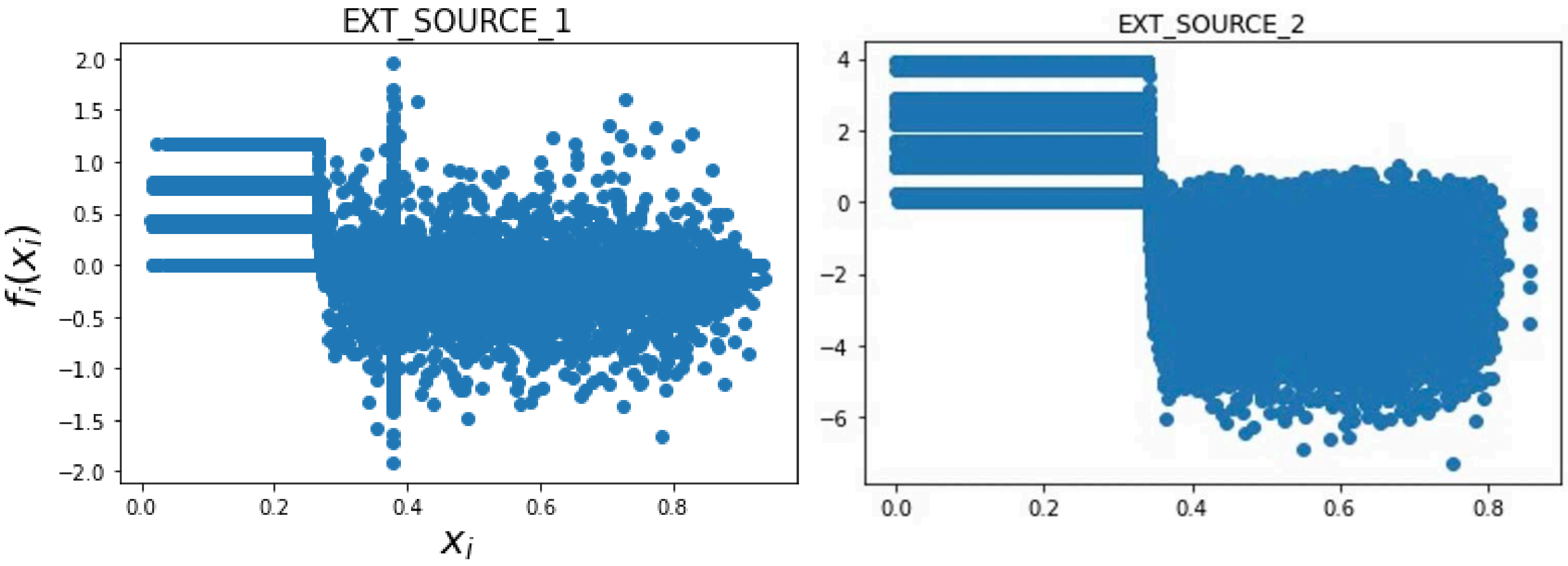
NAM (Neural Additive Model): Accuracy

- AUC Score
    - · 0.73351 on 999 features
    - · 0.70365 on TabNet selected features
- Potential Problems
    - · Tuning
    - · no higher order features are learned

NAM (Neural Additive Model): Interpretability

$$g(\mathbb{E}[y]) = \beta + f_1(x_1) + f_2(x_2) + \cdots + f_K(x_K)$$

- 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.



Conclusion

- TabNet is only partially interpretable.
- Zero importance weight for features does not mean useless information.
- Information flows in to build masks to help prediction

Conclusion

- 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.
- 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.

References

1. Arik, Sercan O., and Tomas Pfister. "Tabnet: Attentive interpretable tabular learning." arXiv preprint arXiv:1908.07442 (2019).
2. Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.
3. Agarwal, Rishabh, et al. "Neural additive models: Interpretable machine learning with neural nets." arXiv preprint arXiv:2004.13912 (2020).
4. James Max Kanter, Kalyan Veeramachaneni. Deep feature synthesis: Towards automating data science endeavors. IEEE DSAA 2015.