

F R A U D U S E R D E T E C T I O N

聚焦数创·连接未来

第七届信也科技杯算法大赛

7th Finvolution Data Science Competition



欺诈用户风险识别

2022 第七届信也科技杯图算法大赛

天才修笔匠 O 银极

主讲：修笔匠小叶

CONTENTS

01. 团队介绍

02. 任务介绍

03. 任务理解

04. 特征工程

05. 模型框架

团队介绍

程唯 (天才儿童)

南京大学硕士

何迪 (修笔匠小叶)

华南理工大学硕士

2020 芒果TV: 首届国际音视频算法大赛-冠军;

2020 招商银行: FinTech精英训练营-数据赛道-冠军;

2021 第四范式: ECAA首届电子商务AI算法大赛-亚军;

2020 图灵联邦: 视频点击预测大赛-亚军;

2019 华为: DIGIX极客算法精英大赛-亚军;

2020 天池: 数字中国创新大赛算法赛-智慧海洋建设-季军;

2021 Peking University: Causal Inference and Recommendation-冠军;

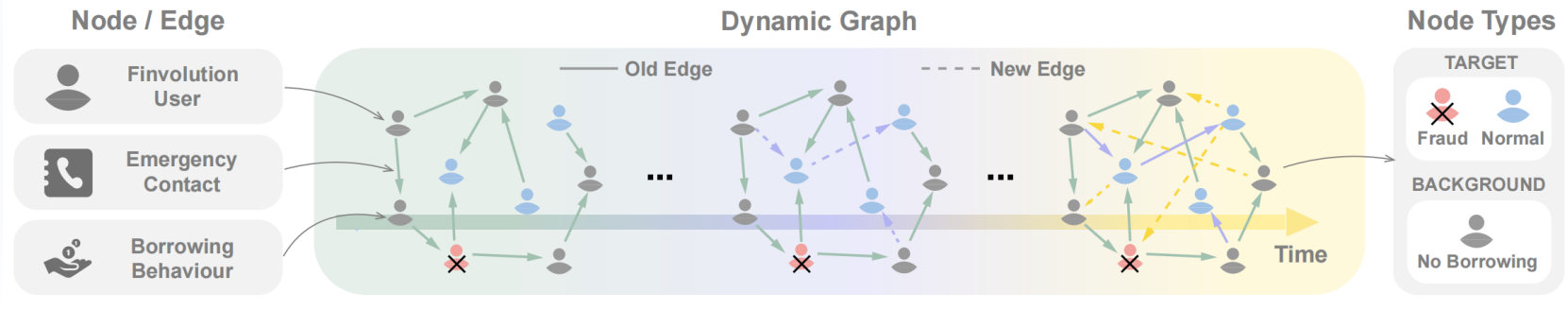
2020 华为: GDE全球开发者大赛-核心网KPI异常检测-冠军;

2021 芒果TV: 第二届国际音视频算法邀请赛-防盗链-亚军;

2020 华为: HDC2020网络人工智能黑客松大赛-硬盘异常检-亚军;

2022 IEEE: CVPR 2022 NAS Track2-Performance Estimation-季军;

任务介绍



任务说明:

反欺诈是金融风控永恒的主题，本次任务为欺诈用户风险识别，评价指标为AUC。

数据说明:

大赛提供2个脱敏的全连通社交网络有向动态图，1个用于训练，1个用于预测（选手不可见），其中节点代表信也科技的注册用户，从节点A指向节点B的有向边代表用户A将用户B填为他的紧急联系人。数据提供了节点的属性，边的属性以及创建日期。

任务理解

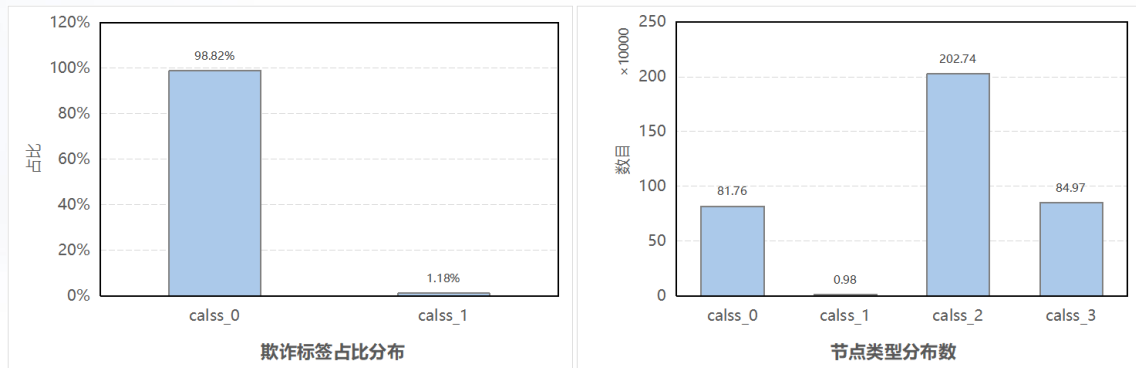


图1. 节点类型分布 (class_0: 正常, class_1: 欺诈, class_2: 背景1, class_3: 背景2)

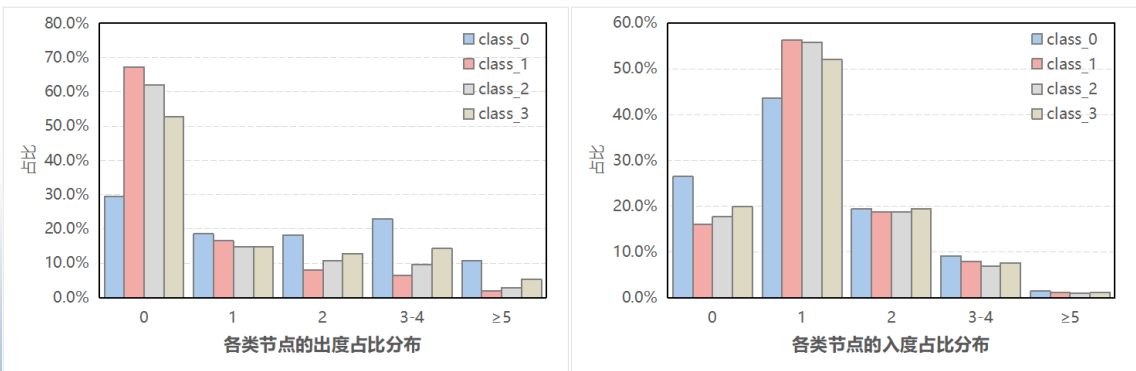


图2. 各类节点的边分布 (class_0: 正常, class_1: 欺诈, class_2: 背景1, class_3: 背景2)

本任务是一个典型的二分类风控反欺诈任务，在借贷用户中欺诈占比 1.18%。然而，不同于传统的风控场景，该任务基于有向图数据。除普通借贷节点的自身特征外，还包含了大量的背景节点，以及海量的关联信息。

且欺诈节点与普通节点的边分布存在显著差异，因此可考虑构建图特征，图算法模型并尝试与风控行业成功经验相结合，充分利用图数据中的信息进行反欺诈节点识别。

特征工程 特征开发

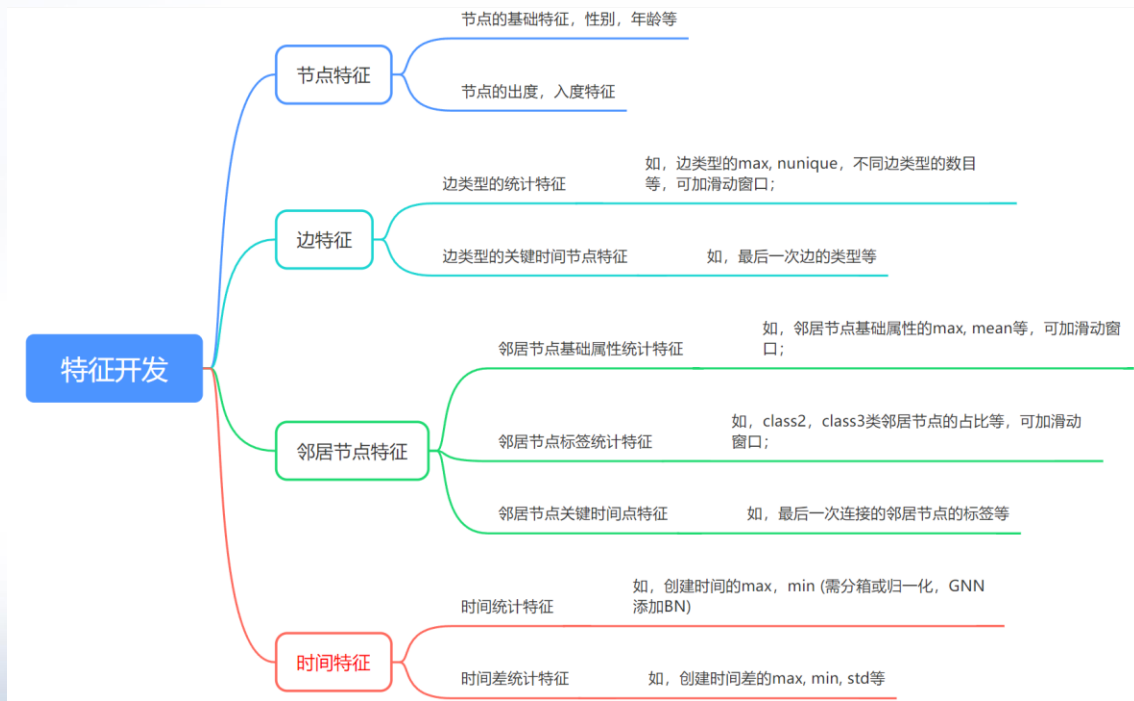


图3. 特征开发

无论是传统机器学习方法还是GNN, 变量开发对于模型最终效果均能带来显著的提升。本任务的图数据中包含[节点, 边, 邻居节点, 创建时间]四方面的信息, 因此, 变量开发从这四个角度入手, 具体思路见图3。需要注意的是, 由于test是一张完全不同的图, 图中时间范围不同, 因此在时间特征构造时, 需要对时间做分箱或归一化处理。

特征工程

特征开发

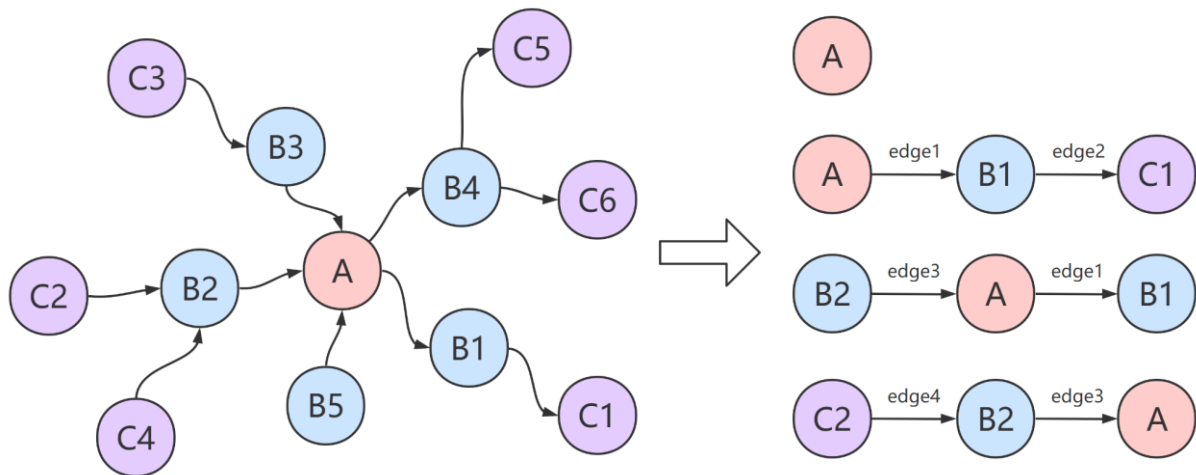


图4. 特征构造时的节点位置 (以节点A为例)

在特征的开发过程中图特征的思路可以直接向高阶拓展（直接将高阶邻居看作邻居或层层拓展），为获得更多的关联信息，我们将二阶关联以内的图特征均进行了开发。图4左侧是以节点A为例的一个局部二阶关联图，在每3个节点构成的二阶关联中，A所处的位置可以简化为图四右侧三种状态，在建模的过程中，三种状态的图特征均被构造，最终所有图信息以特征的形式给到节点A。

特征工程 特征筛选

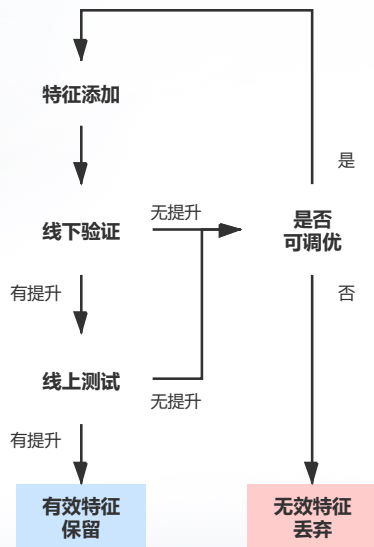


图5. 建模过程中的特征筛选

建模过程中的特征筛选：

在模型开发的过程中以特征组（一个小方向的特征集合为一组）为单位进行特征构造及建模实验不仅能保证模型的稳定性，而且能够明确不同特征组带来的提升大小并针对提升大的特征组做进一步的挖掘，这对于模型的持续提升方具有极大的指示作用。在这个过程中无效或过拟合特征会被剔除，因此建模过程自带特征筛选，其具体流程见图5。这个过程最终有210个特征。

建模末期的特征筛选：

特征开发基本完成后，整个特征集合仍可能存在一些冗余或噪音特征，噪音的剔除能带来模型的进一步提升，我们在建模末期尝试了如下两个特征筛选方案。

1. Null Importance [3]：将特征重要性与随机标签建模的特征重要性进行对比，剔除特征重要性低于随机标签模型特征重要性的变量；
 2. TOP N Feats：仅保留特征重要性前N的特征，其余特征删除，N由线下实验确定；
- 两种特征筛选方案在线下线上均能带来1-2k的提升，其中TOP N变量分数略高。

最终在树模型中我们保留了TOP100的特征。在图神经网络模型中仅使用建模过程中自带的特征筛选，使用了210个特征。

特征工程 特征一览

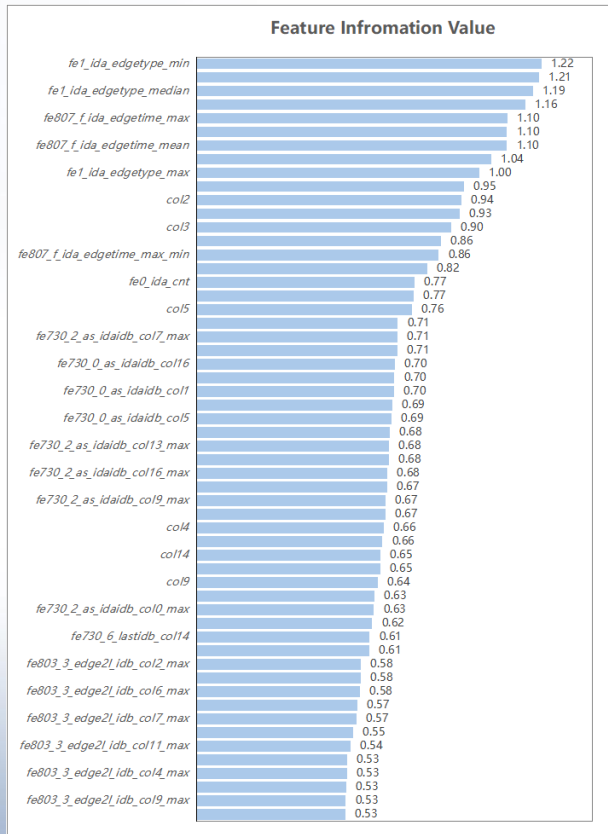


图6. TOP特征IV值

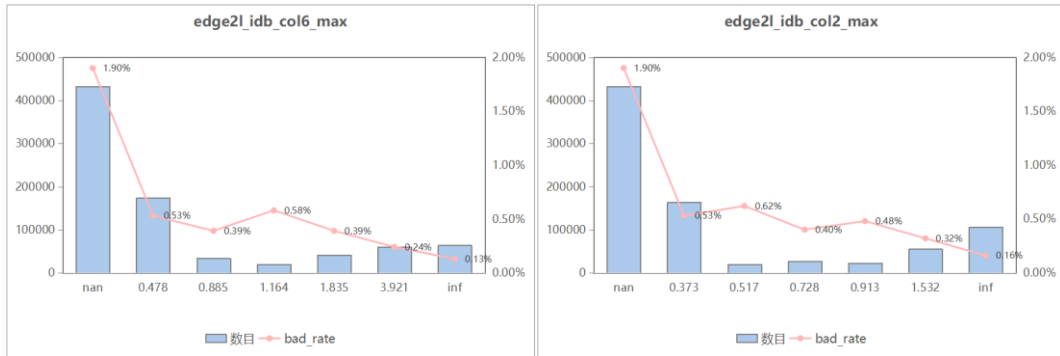


图7. 二阶特征示例

从图6的TOP特征的IV值一览中可以看出：

图特征在欺诈识别任务中有着极强的区分能力；

2阶图特征仍有很强的区分能力；

模型框架 机器学习模型

在机器学习模型方向，我们尝试了多种不同的算法：

1. 在风控行业有大量成功经验的树模型：LightGBM, XGBoost;
2. 匿名结构化任务中展露头角的神经网络新星：TabNet [4];
3. 结构化数据中CNN的奇妙应用：1DCNN [5];
4. 传统神经网络模型：MLP;

最终实验结果表明，除LightGBM, XGBoost外，其他模型线上都难以突破0.8+。

因此我们最终选择LightGBM作为机器学习主模型，LightGBM单模线上0.8107。

模型框架 图神经网络

在图神经网络架构设计的过程中，我们尝试过很多结构，如GraphSAGE, GATv2等等，并进一步尝试将图神经网络层与MLP, CNN等传统神经网络结构进行组合，最终基于GraphSAGE的单结构和组合结构效果均明显优于其他图卷积，因此最终的图神经网络架构基于GraphSAGE。

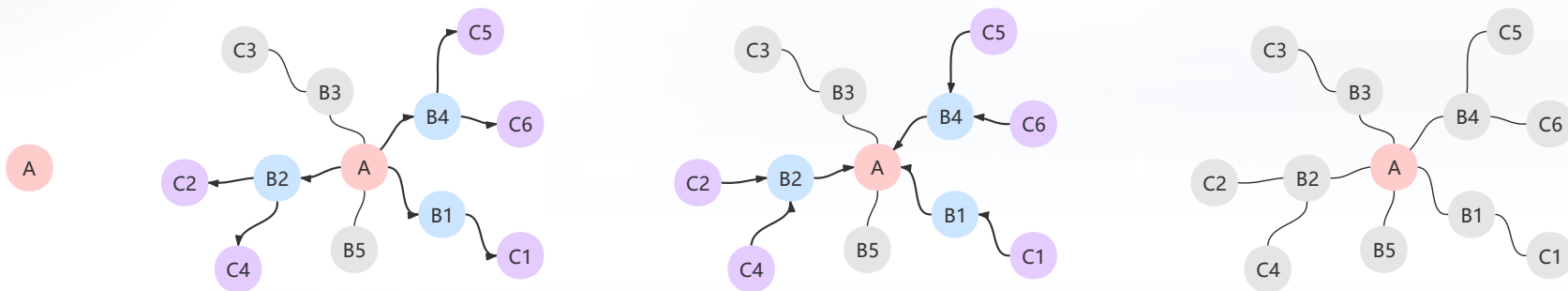


图8. GraphSAGE

本任务中将有向图改为无向图能显著提升图神经网络模型的效果，因此使用无向图作为模型的输入；

NeighborSampler使用三阶邻居节点，每层的所有邻居均被保留；

210个入模变量使用均值进行缺失值填充并添加变量是否为空的辅助变量，变量及辅助变量合计403维；

模型框架 图神经网络

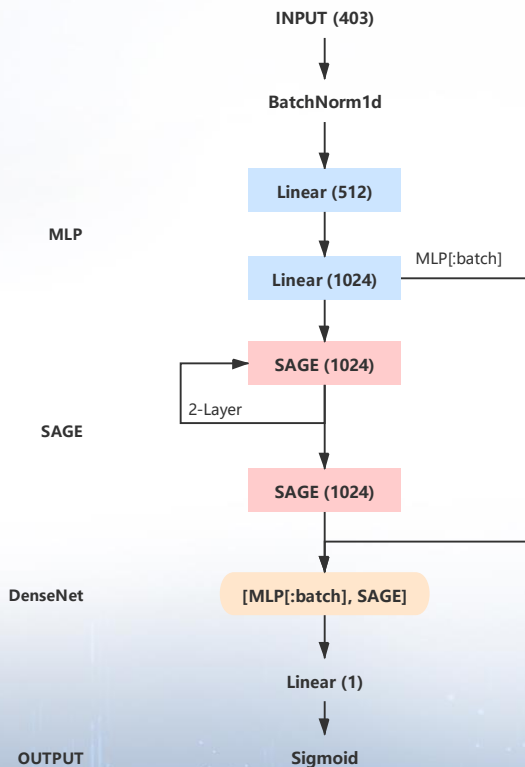


图9. GNN网络架构

1. 最终使用的图神经网络架构如图9所示，每个batch三阶内的所有节点均先做BatchNorm，再经过一个2层的MLP，之后NeighborSampler的前两层使用同一个SAGE，第3层单独使用一个SAGE，最终将MLP块中心节点的输出与SAGE的输出做稠密连接，再FC为1维变量，做Sigmoid后成为最终的输出；
2. 使用BCELoss作为损失函数进行模型训练；
3. 我们的GNN框架在不做调优的情况下能够轻松到达0.804，成绩远超普通神经网络模型，且与树模型的oof相关性仅0.87，能够与树模型进行很好的互补；

模型框架

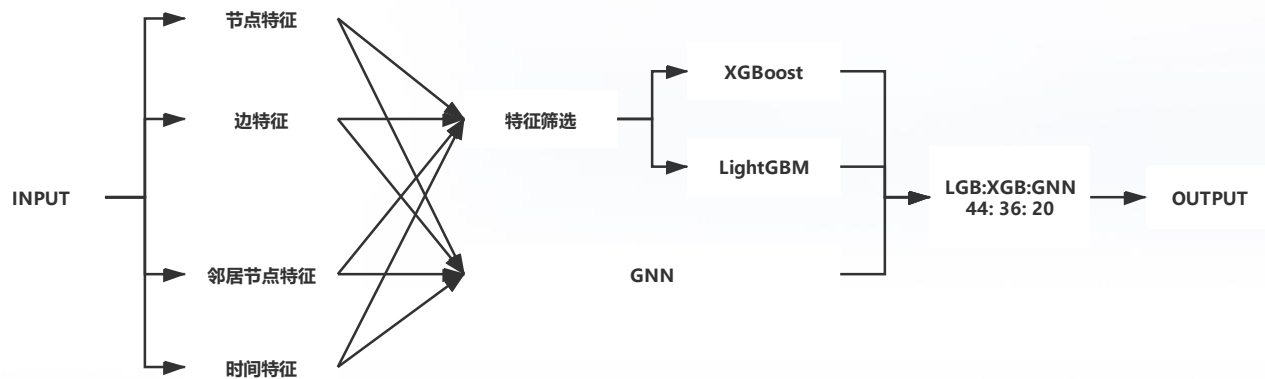


图10. 模型框架

总结

1. 图特征对于树模型和图神经网络模型的提升均有着至关重要的作用，图特征的思路可以用高阶邻居或层层传递的方式向任意阶拓展；
2. 在图数据中树模型结合图特征仍然能取得很好的效果；
3. 特征筛选能带来模型的进一步提升；
4. 图神经网络效果明显优于其他神经网络方法如MLP, TabNet, 1DCNN；
5. 将有向图转为无向图能显著提升图神经网络模型的效果；
6. 图神经网络前加入MLP对所有输入节点预处理能明显提升模型效果；
7. 图神经网络与树模型相关性不高，能够与树模型相辅相成，提高欺诈识别上限；

Reference

- [1] https://github.com/DGraphXinye/2022_finvcup_baseline
- [2] Huang X, Yang Y, Wang Y, et al. DGraph: A Large-Scale Financial Dataset for Graph Anomaly Detection[J]. 2022.
- [3] <https://www.kaggle.com/code/ogrellier/feature-selection-with-null-importances>
- [4] <https://github.com/dreamquark-ai/tabnet>
- [5] <https://github.com/YeYeYetta/Kaggle-MoA-2nd-Place-Solution/blob/main/training/1d-cnn-train.ipynb>
- [6] Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs[J]. Advances in neural information processing systems, 2017, 30.

THANK YOU

