1. 老师好，我是GPU8.0队的汇报人邓语苏，我们队的方案是：基于Stacking的GBDT多模型融合Voting方案 用来解决点击反诈预测。
2. 该赛题的数据集提供了一共20个维度的特征，训练集为50万条，测试集为15万条。是属于较大的数据集
3. 首先介绍一下我们的成果，我们队在A题中取得了第一的好成绩。
4. 我们爬取了全飞桨平台同名比赛的榜单，可以看到89.4以后，每提升0.01都是极其艰难的事。**而我们的成绩达到了惊人的89.4847，这也同时是2020-2023年，比赛创办以来，全平台断层式的第一。**
5. **接下来我将从：模型考察与选择、数据预处理、框架探索与改进、测试与细节精进四个方面介绍我们的方案**
6. **关于模型选择，我们的愿景是尽可能的准确、稳健和灵活。**
7. **我们紧接着对比了目前主流的两种模型：GBDT、NN网络。**
8. **查阅文献可知，表格数据中GBDT效果仍然优于深度学习。**其原因如下：表格数据的特征往往有较为明确的含义，NN对表格做特征提取，可能会导致信息损耗。而GBDT仅做特征空间的划分，因此更优。
9. **在得到验证后，最终敲定了GBDT作为主要模型。我们选择catboost模型作为我们的起点，它有很好的处理大量类别特征的能力，正与本题一拍即合**
10. **选定模型之后，我们对数据集做预处理。包含了特征清洗、特征增广、特征筛选三个方面。**我们对version列进行特征清洗、对已有特征列根据频率编码、nunique编码、交叉类别构造新的特征列。接着我们使用GBDT自带的模块查看特征重要性，将重要性极低的特征剔除。
11. 训练catboost模型可以得到初步结果--89.424。**接着我们进行框架上的探索与改进。**
12. **初次探索，很容易想到的是，将多个不同的模型使用Voting进行融合。**
13. **事实证明，这确实是一个简单有效的思路。**我们分别对xgboost、catboost、LGBM做训练，将得到预测结果做硬投票之后，得到了历史新高的准确率--89.45。但**Voting没有复杂的机制，实现简单的同时也带来了许多局限性，例如稳健性很低，上限不高。让我们接着向机器学习的深处探索.....**
14. **怎样改？通过Stacking模块增加层次结构**
15. Stacking通过将多个基础模型的预测结果作为输入来训练元模型，从而捕捉到不同模型之间的潜在关系。//但是天下没有免费的午餐，任何选择都伴随着代价。那么我们的选择伴随着怎样的代价呢？//由于涉及多个模型多层次的训练，Stacking计算复杂度高、容易过拟合。可以说是一个牺牲时间，换取准确率的策略
16. 我们选择randomforest等五个模型作为基模型做训练，即可获得对train、test集label值的各五列预测。// 将其与原数据集拼接，即可获得新的数据集。
17. 将新的数据集放入xgboost、catboost、lgbm三个元模型进行训练，把得到的预测结果做硬投票即可得到最终结果，由原先的89.45提升到了89.4807。
18. **先通过框架的探索与改进，保证大方向上的正确。紧接着，我们进行测试与细节精进，把控细节，精益求精。**
19. 在数据预处理模块，需要对时间戳进行处理，有两个方案。方案一是计算时间差后做分桶，方案二是做时间频率编码。经过3次测试，取平均准确率，可知分桶处理更好
20. 由于GBDT本质为树模型，对缺失值的填补和类别编码是不必要的。我们好奇的是，手动处理和树模型内置处理哪个效果更好？

为此我们尝试了随机森林填补缺失值和独热编码，准确率不升反降。封装的越好的模型，里面值得挖掘的点就越多，好奇心虽好，但也要对其抱有敬意。

1. 在训练模型的部分，我们在Stacking中//使用了五折交叉验证//增加模型稳健性。在模型训练前//使用随机搜索//优化模型的参数选择，提高了模型准确率。
2. 至此，我们队的核心技术路线建立完毕。如果让我用两个词形容我们的工作：
3. **——勇气与严谨。与其他队伍不同的、勇于做框架级别的改进；与其他队伍不同的、严谨的实验验证； 与其他队不同的，对细节的把控。**
4. **使得我们能够取得A题第一，同时也是比赛创办以来，全平台断层式的第一。**
5. **感谢老师的聆听，希望能将我们队的工作清楚的展现给大家。也感谢队友们对我的信任和支持，陪我一遍遍的练习。**