人工智能导论小组作业

小组成员：121072021031 许季韬（组长）、121072021020 缪与凡

一、任务背景

在2022 世界互联网大会乌镇峰会期间发布的《世界互联网发展报告2022》显示，2022 年第一季度，全球5G 用户数增加7000 万人，总数达到6.2 亿人左右，5G人口覆盖率超25%。据爱立信及全球移动通信系统协会(GSMA)预测，到2022年底，全球5G用户数量将突破10亿人。对于通信运营商来说，面对如此庞大的5G 市场，如何基于一些用户侧的信息进行用户画像，再进一步对于潜在的5G使用者进行精准的推销也是非常有帮助的。

二、任务目标

根据用户基本信息和通信相关数据、比如用户话费信息、流量、活跃行为、套餐类型、区域信息等特征字段，然后通过训练数据训练模型，预测测试集中每个样本是否属于5G用户。

三、评估指标

评价标准采用AUC，即分数越高，效果越好。

四、数据分析与预处理

1、读入数据并检查是否有缺失值

#	Column	Non-Null Count	Dtype
0	id	800000 non-null	int64
1	cat_0	800000 non-null	int64
2	cat_1	800000 non-null	int64
3	cat_2	800000 non-null	int64
4	cat_3	800000 non-null	int64
5	cat_4	800000 non-null	int64
6	cat_5	800000 non-null	int64
7	cat_6	800000 non-null	int64
8	cat_7	800000 non-null	int64

结果分析：没有缺失值

2、转换数据格式，离散型数据全部转为 int，连续性数据全部转为 float

RangeIndex: 800000 entries, 0 to 799999
Data columns (total 60 columns):

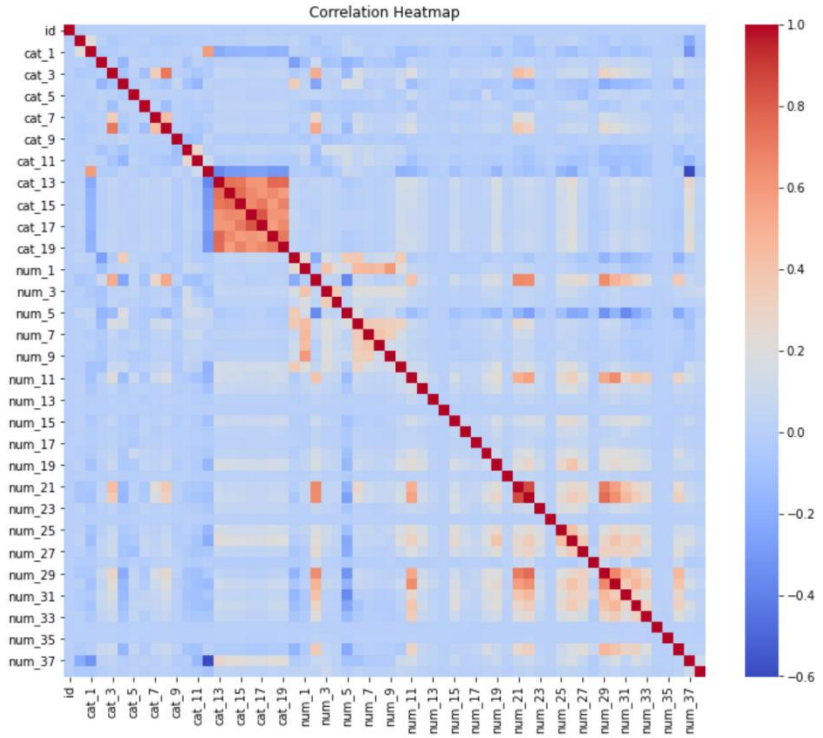
#	Column	Non-Null Count	Dtype
0	id	800000 non-null	int64
1	cat_0	800000 non-null	int64
2	cat_1	800000 non-null	int64
3	cat_2	800000 non-null	int64
4	cat_3	800000 non-null	int64
5	cat_4	800000 non-null	int64
6	cat_5	800000 non-null	int64
7	cat_6	800000 non-null	int64
8	cat_7	800000 non-null	int64
9	cat_8	800000 non-null	int64
10	cat_9	800000 non-null	int64
11	cat_10	800000 non-null	int64
12	cat_11	800000 non-null	int64
13	cat_12	800000 non-null	int32

结果分析：数据格式转换成功

3、对数据进行标准化处理

	id	cat_0	cat_1	cat_2	cat_3	cat_4	cat_5
0	0	-1.238127	0.114263	0.750939	0.211040	-0.623082	0.786152
1	1	-1.503629	0.992248	-0.456075	-4.497053	-1.488844	0.786152
2	2	-0.574371	-0.277270	-0.053737	0.211040	1.108443	-1.272019
3	3	-0.043366	0.861737	-0.053737	-4.497053	-0.623082	0.786152
4	4	-1.503629	1.004113	1.153276	0.211040	-0.623082	-1.272019
...	...	...	...	...	...	...	...
799995	799995	0.753142	1.549887	-0.053737	0.211040	1.108443	0.786152
799996	799996	-0.308868	-0.372188	0.348601	0.211040	0.242681	-1.272019
799997	799997	2.346157	0.209181	-1.260751	-6.851100	1.108443	0.786152
799998	799998	0.885893	2.748217	-0.053737	0.211040	-0.623082	0.786152
799999	799999	-0.972624	0.363421	-2.467765	0.211040	-1.488844	0.786152

4、用热力图的深浅来观察特征之间的相关性



结果分析：如上热力图显示特征 cat_12~cat_19 的颜色集中较深，相关性较强，num_21 和 num_22 的相关性较强。

5、找出相关性超过 0.7 的列

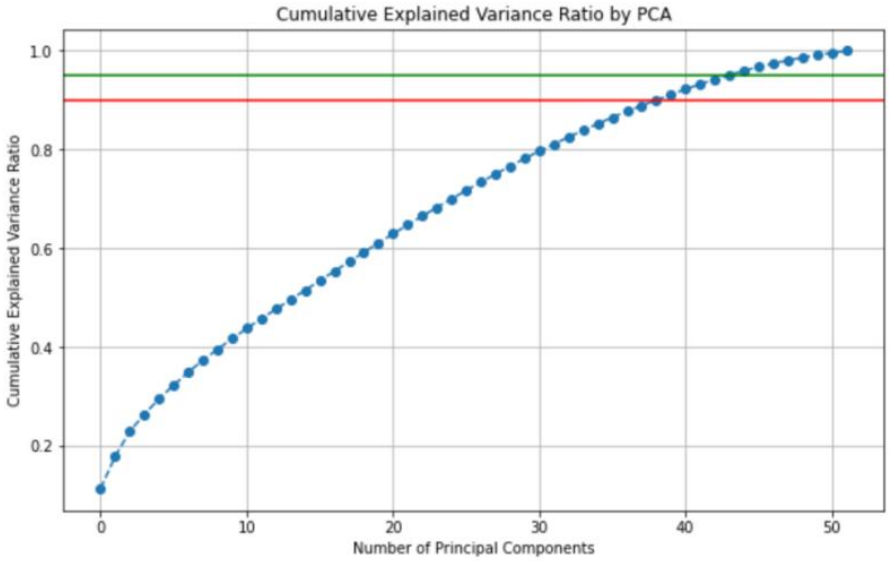
6、只保留特定列，删去一部分列

删除的列为['cat_3','cat_13','cat_15', 'cat_19','cat_17','num_22','num_29']

7、尝试使用 pca 对数据进行降维

（1）使用累积解释方差比率：选择解释方差累积达到一定百分比（例如 90%或 95%）的主成分数量。

（2）绘制碎石图：每个主成分的解释方差并寻找拐点（“肘部”），在此点之后主成分的边际增益变得不明显。




```
Number of components to explain 90% variance: 39
Number of components to explain 95% variance: 44
```

从图上来看，没有拐点，故不需要 pca 处理

8、划分训练集与测试集

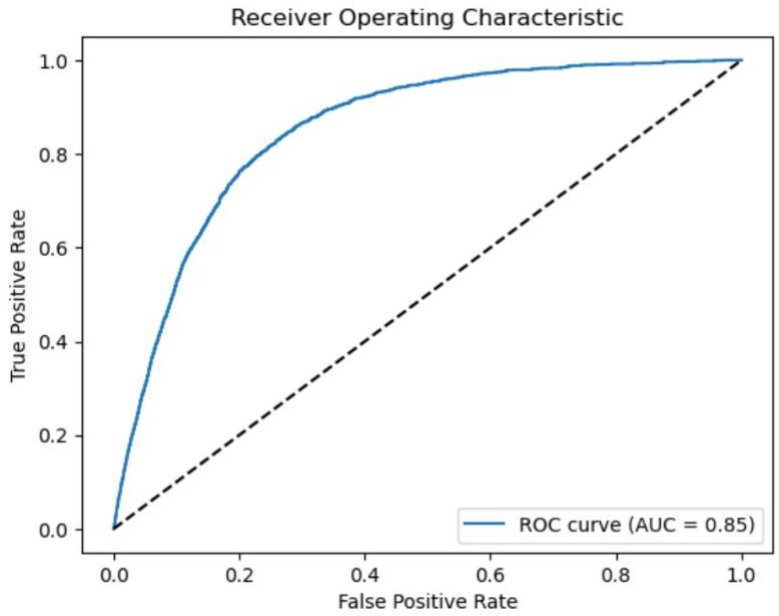
观察数据集发现，正负样本比例过于悬殊（78:1），故不能直接利用 scikit-learn 中的采样方法。

这里从 780000 的负样本中随机选择 9000 个负样本，从 10000 个正样本中随机选择 9000 个正样本组成训练集，剩余的作为测试集用户后续的模型训练。

9、 继续处理数据，注意将前面采样出的作为训练集的样本删去

五、模型建立与结果

1、使用逻辑回归建立预测模型并计算 AUC 的值




```
AUC：  0.8522570845079446
```

2、使用决策树建立预测模型并计算 AUC 的值

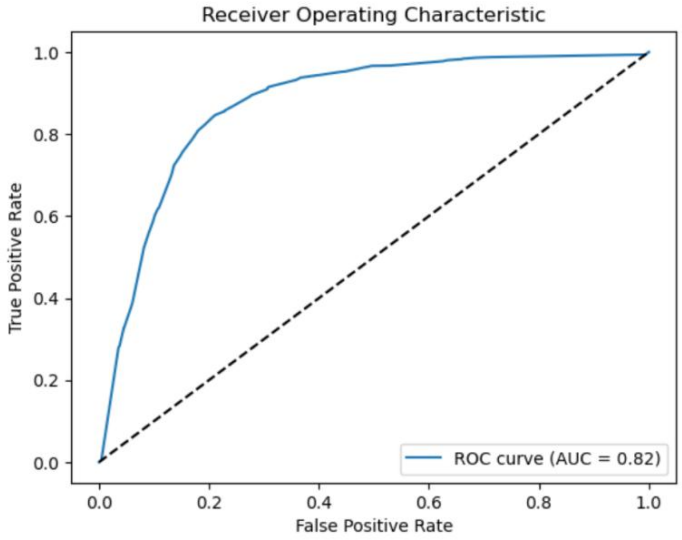
（1）定义超参数的候选值

```
param_grid = {
    'criterion': ['gini', 'entropy'],
    'max_depth': [3, 5, 7],
    'min_samples_split': [2, 5, 10],
}
```

（2） 使用网格搜索进行超参数调优

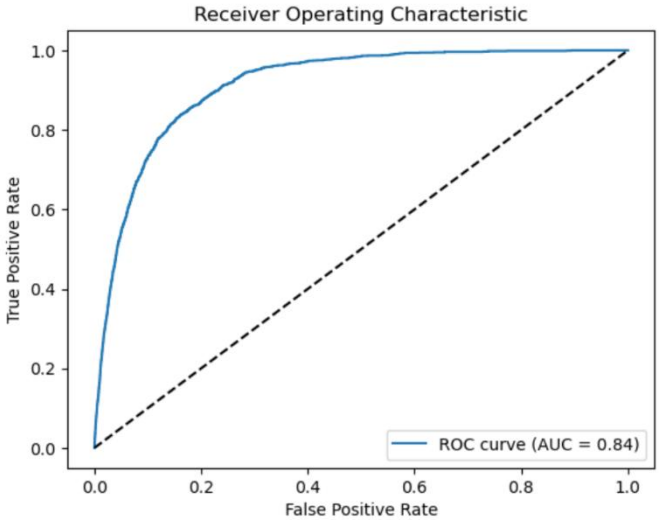
```
grid_search=GridSearchCV(estimator=clf,param_grid=param_grid, scoring='accuracy', cv=5)
grid_search.fit(X_train, y_train)
```

（3） 输出最佳超参数组合，使用该组合在测试集上进行预测并评估模型性能



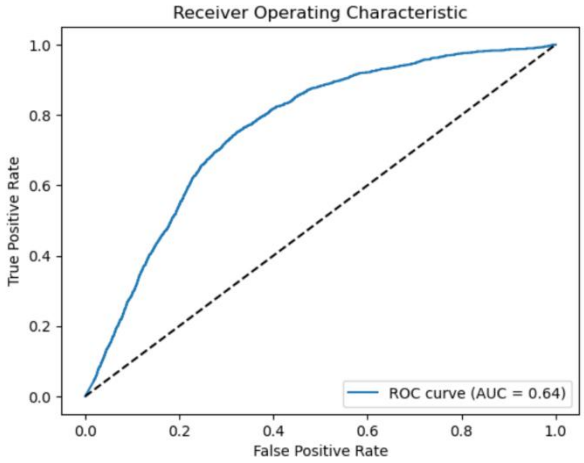

```
最佳超参数组合： {'criterion': 'gini', 'max_depth': 7, 'min_samples_split': 5}
AUC： 0.8175201819579703
```

3、使用随机森林建立预测模型并计算 AUC 的值



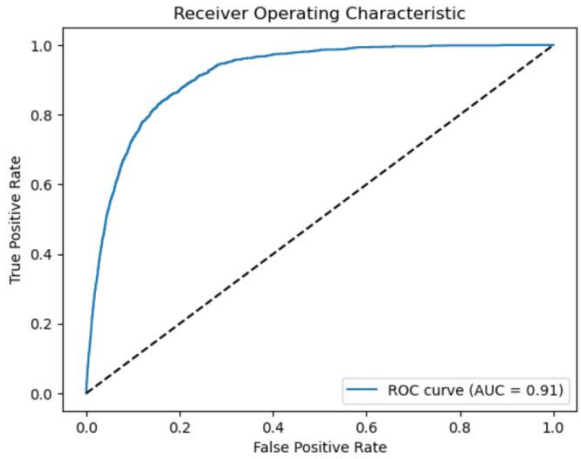

```
AUC： 0.8353434136340339
```

4、使用朴素贝叶斯建立预测模型并计算 AUC 的值



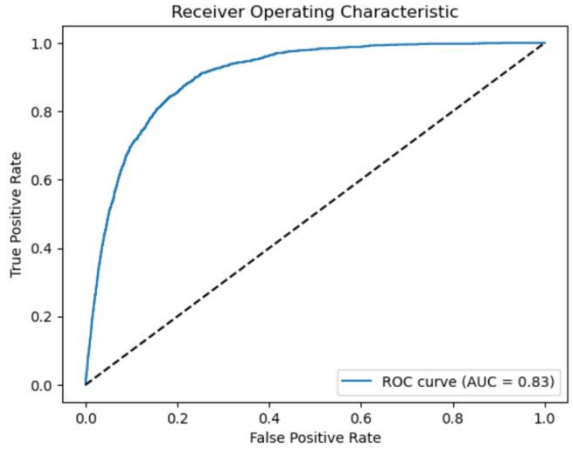
AUC： 0.6365383136852896

5、使用 Xgboost 建立预测模型并计算 AUC 的值



AUC： 0.9102812748270117

6、利用 gbdt 建立预测模型并计算 AUC 的值



AUC: 0.828997148898001

六、模型解释与思考

1、逻辑回归、随机森林、xgboost 三个模型都是属于分类模型。

2、逻辑回归模型由两部分组成：输入和激活函数 。在输入中给予不同特征不同的权重，以此得到线性回归的结果。激活函数 sigmoid 会将输入转变为[0, 1]区间中的一个概率值，默认 0.5 为阈值。当大于阈值时认为是正样本，反之为负样本。

3、在说明随机森林和 xgboost 之前，这里先介绍集成学习的两个重要思想：Bagging 和 Boosting。Bagging 是一种基于自助采样的集成学习方法。它通过有放回地随机抽样生成多个独立的子训练集，然后在每个子训练集上训练一个基本模型。最终的预测结果通过对各个基本模型的预测结果进行投票或平均得到。

4、随机森林是 Bagging 的一个变体。随机森林的做法是在 Bagging 采样得到的样本集合的基础上，随机选出 k 个属性再组成新的数据集后再训练决策树。最后训练 T 棵树进行集成。随机森林算法广泛运用于分类问题和回归问题。

5、Boosting 是一种迭代的集成学习方法，它通过顺序训练多个基本模型来逐步提升整体模型的性能。在每次迭代中，Boosting 会调整样本的权重，使得之前模型预测错误的样本在后续迭代中得到更多的关注。一般将以基模型为决策树的模型称为提升树。常见的 Boosting 算法包括 AdaBoost 和梯度提升树(Gradient Boosting Tree)。最终的预测结果是通过对所有基本模型的加权组合得到，权重通常与基本模型的性能相关。

6、xgboost 是一种梯度提升框架，它基于决策树集成的算法，被广泛应用于分类和回归问题。xgboost 结合了梯度提升算法和正则化技术，通过迭代地训练决策树模型，并使用梯度提升来提高模型的性能。其中，梯度提升算法可以减小残差，提高模型的预测性能；正则化技术可以防止过拟合和提高模型的泛化能力。

七、github 仓库地址

Github 仓库地址：https://github.com/eclisep/-.git