**EDA部分answer**

1. **Which features are most likely to be predictive of each target class?**
2. 均值法
3. t-test法
4. features\_importance\_
5. **How does the class imbalance affect the learning process, and what methods can mitigate this issue?**

2.1 怎么影响：

2.1.1介绍：class imbalance是几个class数量特别多(多数类)，其余class数量特别少（少数类）。写论文的时候可以举例介绍（病人分类问题）

2.1.2影响：使模型更集中于学习数量占多数的那几个class，因为即使模型只预测这几个class，也能得出一个不错的准确率

2.1.3

所有的evaluation metrics

Accuracy = (TP + TN) / (TP + TN + FP + FN) 衡量预测正确的样本占总样本的比例

Precision = TP / (TP + FP) 预测为正的样本中，有多少是真的正例

Recall = TP / (TP + FN) 真实为正的样本中，有多少被正确预测为正

F1-score = 2TP / (2TP + FP + FN) 精确率和召回率的调和平均

Specificity = TN / (TN + FP) 真实为负的样本中，被正确预测为负的比例

|  |  |  |
| --- | --- | --- |
| 预测  真实 | Positive（少数类） | Negative（多数类） |
| Positive（少数类） | TP | FN |
| Negative（多数类） | FP | TN |

所以这就是一个真实少数类和误判为少数类的选择问题，我们要做的是平衡所有的少数类，而不是忽略它们

TP 是真实的少数类

# 因为这些类的数量少，所以TP is small

FP 是多数类被误判为少数类

## 但是，因为多数类数量很多，所以它们被误判为少数类的概率会很大。即FP is large

TN is large，本身多数类，被预测为多数类的概率会很大

FN 少数类被预测为多数类

### 这种情况基本不会发生，可以理解为FN is so small

综上，如果使用Precision和Recall会导致它们的值偏小

2.2 方法：

2.2.1Resampling methods

e.g., SMOTE

#SMOTE应在划分数据集之后使用

找一个少数类的点，找它的邻居，然后连线。从连的线上面进行采样数据

假设A类很多，B类很少

那我只对A类的部分数据进行采样(Undersampling)，对B类的所有数据进行重复采样(Oversampling)

2.2.2 Model方面

cost-sensitive method

当多数类预测错误的时候，将其权重设置为1/2

当少数类预测错误的时候，将其权重设置为2

class-weighted method

注：报告里面所有用到的方法如SMOTE要解释原理，模型用到的所有参数如max\_iter要解释为什么这么选

3. What are the appropriate evaluation metrics for this task, and why do accuracy-based metrics fail in imbalanced classification?

3.1

F1 - score

AUC- PR

AUC- ROC

3.2

模型更集中于学习数量占多数的那几个class，因为即使模型只预测这几个class，也能得出一个不错的准确率。所以我们需要的是既可以反应少数类，也能反应多数类的评估指标。