**恶意加密流量检测模型文档**

1. **数据分析：**

通过对训练集和测试集的分析，两者类别特征的分布不均匀，测试集中出现了很多训练集中没有出现的结果，因而，将训练集和测试集分开做特征工程，而不是对训练集做特征工程，再将结果映射到测试集中。

1. **特征工程：**

**变量衍生：**

1. 删除唯一值的变量
2. 来源地址与来源端口拼接，目的地址和目的端口拼接，组合成地址+端口变量
3. 地址与端口之间做count和nunique统计特征，再做一个nunique/count的特征
4. 数值特征做log处理，使其符合正态分布
5. 类别特征做频次编码
6. 数值特征与类别特征之间做交叉统计特征
7. 数值特征之间做互相做算术计算

最终共生成499个特征

**变量筛选：**

1. PSI筛选，考虑到训练集和测试集分布不均匀的情况，将PSI值大于0.2的删除
2. 单特征AUC筛选，将训练集划分成两部分，其中25%用作验证集，对每个特征用lightgbm建模，评估指标选AUC，将验证集的结果AUC<0.52（区分能力较差）的删除
3. 相关性分析，将特征之间的相关性大于0.98的变量保留其中一个，保留上一步中AUC值较大的那个

最终选中了86个特征入模

1. 模型训练

使用LightGBM对train进行建模，模型保存为train\_code/lgb.txt。