**AI大数据安全分析赛-恶意加密流量检测方向赛题说明**

**1 赛题背景**

2020年将超过60%的企业将无法有效解密HTTPS流量，从而“无法有效检测出具有针对性的网络恶意软件。”届时加密的流量中将隐藏超过70%的网络恶意软件，而对抗这些威胁的手段将会受制于反解密系统，即便是最大的IT团队也无法忽视这一问题。

**2 赛题任务**

本赛题提供的日志数据为在某安全产品中产生加密流量数据，其中部分数据根据是否是恶意流量进行了标注。请根据提供的已标注的训练样本进行分析和模型训练，使得模型可从测试样本中区分正常加密流量和恶意加密流量，并根据要求提交预测结果、相关代码及说明文档。

**3 数据说明**

本赛题提供的数据的字段说明详见**数据字段说明.xlsx**文档。具体数据存放在train.csv作为训练数据，带有label标签，test\_1.csv作为发放给参赛者的测试数据，格式除了没有label字段，其他数据规格与训练数据一致。注意，主办方会保留额外同格式测试数据test\_2.csv，用以测试参赛者提交模型的泛化能力。

**4 排名&晋级规则**

排名规则：排名分A榜和B榜，其中A榜每周一至周四19点前根据当日提交有效内容情况更新；B榜每周五19点前根据当日提交有效内容情况更新。

注意：A榜基于测试数据test\_1.csv的结果给出，B榜基于测试数据test\_1.csv+test\_2.csv给出。

晋级规则：每周五（10.16日和10.23日）B榜前2直接晋级决赛，重复队伍不额外占决赛名额，初赛结束后根据最后一周周五（10.30）B榜情况并复核参赛者提交代码质量和文档质量确定剩余6支队伍，即最终10支队伍晋级决赛。

**5 评价准则说明**

本赛题的评价准则使用F1得分，计算如下：



其中P为查准率，R为查全率。本题将F1得分转化为百分制，作为排名的依据。

**6 提交内容说明**

A榜提交内容的一个样例请见文件：大数据队\_eta\_submission\_1011.csv，其中eventId列同test\_1.csv中的该列，此列作为样本的标识符，label作为参赛者预测的标签。

B榜提交内容的一个样例请见压缩文件：大数据队\_eta\_submission\_1011.zip，其中data文件夹下存放训练和测试数据，train\_code文件夹下存放训练代码和其它必要文件，predict\_code文件夹下存放预测代码和其他必要文件，result文件夹下存放预测的结果文件。请按照样例的格式完善和补充如下内容：

1. 完善训练代码（train\_code文件夹下的main\_train.py中的train\_func函数），保证可一键运行，并在同文件夹下保留生成的训练好的模型文件（例如model.pkl）。
2. 完善预测代码（predict\_code文件夹下的main\_predict.py中的test\_func函数），保证可一键运行，并在同文件夹下保留预测使用的模型文件（例如model.pkl，此文件应与train\_code文件夹下的model.pkl相同）。
3. 根据test\_1.csv预测的结果文件：大数据队\_eta\_submission\_1011.csv（同A榜提交文件）存放至result目录下。
4. 根据代码和解题思路提供说明文档：恶意加密流量检测提交结果说明文档.docx，其中可包括对数据的认识、特征工程、使用的模型、训练和调优过程以及在训练集上的表现等内容。
5. 其它可保证代码正常运行的必要文件。

注意：

1. 主办方会使用参赛者按要求封装好的main\_predict.py对test\_2.csv进行预测，并将输出结果作为基于此给出参赛者B榜成绩，请确保该脚本的质量。
2. 本赛题提交文件命名方式为：队名\_eta\_submission\_日期，例如队名为“大数据队”，提交时间为10月11日，则该队B榜提交的压缩文件命名方式为：大数据队\_eta\_submission\_1011.zip。
3. 请勿修改文件结构和文件名称，请勿修改代码的结构和函数名称。
4. 主办方会对提交的代码进行审核，对重复度高于一定值且有疑似开小号情况的参赛者取消参赛资格。