**AI大数据安全分析线下赛-恶意加密流量检测方向赛题说明**

AI大数据安全分析线下赛分为决赛赛题和附加赛题部分。以下将具体介绍赛题说明、评分标准，并以附录的形式给出其它注意点。

**1 决赛赛题**

**1.1 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。

**1.2 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**。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。

**1.3 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。

**1.4 提交内容说明**

决赛的成绩包括决赛成绩A和决赛成绩B。其中决赛成绩A根据test\_1.csv的预测结果计算得出，决赛成绩B根据参赛队伍提交模型在保留数据test\_2.csv上的预测结果计算得出。

决赛成绩A所需提交内容的一个样例请见文件：样例/finalA/大数据队\_finalA.csv，其中eventId作为样本的标识符，label作为参赛者预测的标签。

决赛成绩B所需提交内容的一个样例请见文件夹：样例/finalB/predict\_code，其中需存放预测代码和其它非数据必要文件。请完善预测代码（predict\_code文件夹下的main\_predict.py中的test\_func函数），保证其在传入参数后可一键运行。

其它需提交内容为训练代码以及根据代码和解题思路提供说明文档。其中训练代码的一个样例请见文件夹：样例/processCode/train\_code，其中需存放训练代码和其它非数据必要文件，请完善预测代码（train\_code文件夹下的main\_train.py中的train\_func函数），保证其在传入参数后可一键运行；说明文档请完善文件样例：样例/ documents/大数据队\_恶意加密流量检测\_结果说明文档.docx并完成文件名修改。

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（主办方的运行环境相关信息请参考附录3.1 Python库及版本），并将输出结果计算决赛成绩B的依据，请确保该脚本的质量，并且保证该函数预测与test\_1.csv中同等数量数据所消耗时间不超过5分钟。

2) 选手不要改变main\_predict.py中的test\_func的结构和路径，主办方会按照约定的路径调用选手的test\_func函数，然后会传入test\_path,save\_path，其中test\_path传入test\_2.csv的路径，save\_path是计算结果保存的路径，请确保在代码中test\_path,save\_path这两个参数是有效的。

3）以上说明是基于样例在本地的情况，请根据主办方分发的纸质文档（或参考附录3.2文件目录及提交内容）将指定内容通过FTP提交至指定路径下，并按照规定命名。

4）主办方会在运行预测模型代码时，仅会使用finalB下的整个predict\_code文件夹，请注意代码调用及路径问题。

5）预测代码应仅为预测功能，切勿在预测代码中包含训练过程，否则判定为成绩无效。

**2 附加赛题**

**2.1 赛题任务**

本赛题提供的日志数据为在某安全产品中产生恶意加密流量数据，这些流量日志归属于不同的恶意家族（这里可以假定同一目的IP地址仅归属于一个恶意家族）。请根据各自在test\_1.csv中预测为正样本（恶意加密流量）的日志通过算法预测出每条日志所属的恶意家族类别，并根据要求提交预测结果、相关代码及说明文档。

**2.2 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。具体数据为在test\_1.csv中预测为正样本的日志。

**2.3 评价准则说明**

本赛题的评价准则加权的FM指数：



其中，F1同1.3章节评价准则，首先根据附加题结果队名\_addition.csv文件和test\_1.csv标签计算F1 score。

然后根据队名\_addition.csv文件和test\_1.csv标签中正样本的交集eventId和label作为FM指数的数据输入，即成功召回的样本。FM指数（Fowlkes–Mallows index, FMI）说明如下：

给定数据集共有*n*个样本，将样本两两组成样本对，则可组成共计个不重复的样本对。若参赛者提交的类别划分集合为（共计*m*类划分），真实类别划分集合为（共计*k*类划分），则FM指数计算如下：



其中：

*TP*为在*A*中属于相同类别且在*B*中也属于相同类别的样本对的数目；

*FP*为在*A*中属于相同类别且在*B*中属于不同类别的样本对的数目；

*FN*为在*A*中属于不同类别且在*B*中属于相同类别的样本对的数目。

本题将WFMI转化为满分20分，作为排名的依据。

**2.4 提交内容说明**

附加题成绩根据选手提交的预测结果计算得出。

附加题成绩所需提交内容的一个样例请见文件：样例/addition/大数据队\_addtion.csv，其中eventId作为样本的标识符，label作为参赛者预测的标签。

其它需提交内容为预测代码以及根据代码和解题思路提供说明文档。其中预测代码的一个样例请见文件夹：样例/processCode/addition\_code，其中需存放代码和其它非数据必要文件，请完善预测代码（addition\_code文件夹下的main\_ addition.py中的addition\_func函数），保证其在传入参数后可一键运行；说明文档请完善文件样例：样例/ documents/大数据队\_恶意加密流量检测\_结果说明文档.docx并完成文件名修改。

注意：

1) label作为参赛者预测的标签，无需给出具体的标签，仅需给出可区分的类别标识即可（例如0，1，2，3，4，5，6用来标识7种类别）。

2）以上说明是基于样例在本地的情况，请根据主办方分发的纸质文档（或参考附录3.2文件目录及提交内容）将指定内容通过FTP提交至指定路径下，并按照规定命名。

**3 评分标准**

**3.1 评分标准**

比赛的总成绩由初赛成绩、决赛成绩A、决赛成绩B和附加题成绩组成，其中初赛成绩以公示的为准，决赛成绩A和决赛成绩B满分为100分，附加题成绩满分为20分。总成绩计算方式：

总成绩=初赛成绩\*0.2+0.3\*决赛成绩A+0.5\*决赛成绩B+附加题成绩

此外，参赛队伍需按照各赛题方向的要求提交必要的代码和说明文档。对于在比赛提交截止时无法提交材料的，主办方会根据情况直接判定该队伍成绩无效；对于提交材料不全的或者代码和文档出现颠三倒四、与成绩表现不符合的情况，主办方会根据情况直接判定该队伍成绩无效或在总得分的基础上进行5分扣除。

**3 附录**

**3.1 Python及库版本**

由于主办方需测试参赛者提交模型的泛化能力，故需要使用参赛者提供的预测代码对主办方保留的测试数据进行测试。为了保证代码的顺利运行，现给出主办方的运行环境供参赛者参考：

平台环境：Linux

Python版本：3.8.3

Python主要依赖库及版本：

Package Version

------------------------- -------------------

blis 0.4.1

bokeh 2.0.2

Bottleneck 1.3.2

catboost 0.24.2

cloudpickle 1.4.1

combo 0.1.1

cufflinks 0.17.3

cycler 0.10.0

cymem 2.0.3

Cython 0.29.20

cytoolz 0.10.1

deepctr 0.8.2

dill 0.3.2

gensim 3.8.3

gmpy2 2.1.0b1

HeapDict 1.0.1

ImageHash 4.1.0

imageio 2.8.0

imbalanced-learn 0.7.0

importlib-metadata 1.7.0

jieba 0.42.1

joblib 0.16.0

Keras 2.4.3

lightgbm 3.0.0

matplotlib 3.2.2

missingno 0.4.2

mlflow 1.11.0

mlxtend 0.17.3

mpmath 1.1.0

nltk 3.5

numba 0.48.0

numexpr 2.7.1

numpy 1.18.5

pandas 1.0.5

pandas-profiling 2.9.0

patsy 0.5.1

pickleshare 0.7.5

Pillow 7.2.0

plotly 4.11.0

pyLDAvis 2.1.2

pyod 0.8.3

PyWavelets 1.1.1

querystring-parser 1.2.4

regex 2020.9.27

scikit-image 0.16.2

scikit-learn 0.23.1

scipy 1.4.1

seaborn 0.10.1

spacy 2.3.2

sqlparse 0.4.1

srsly 1.0.2

statsmodels 0.11.1

suod 0.0.4

sympy 1.5.1

tables 3.6.1

tblib 1.6.0

tensorflow 2.3.1

textblob 0.15.3

thinc 7.4.1

torch 1.6.0+cpu

tqdm 4.47.0

umap-learn 0.4.6

wordcloud 1.8.0

xgboost 1.2.0

xlrd 1.2.0

yellowbrick 1.1

**3.2文件目录及提交内容**

参赛选手被分配的路径下共有文件夹7个，具体提交内容、命名方式等信息如下表所示：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **文件夹名称** | **初始状态** | **需提交内容** | **备注** | **截止提交时间** |
| data | 存放赛题压缩文件 | 无 |  | -- |
| finalA | 空 | 测试集预测结果的csv文件 | 命名方式：队名\_ finalA.csv | 16时30分 |
| finalB | 空 | 存放预测模型代码的文件夹predict\_code | 请注意提交文件夹；请勿修改main\_predict.py的结构和函数名称 | 16时30分 |
| addition | 空 | 附加题结果csv文件 | 命名方式：队名\_addition.csv | 16时30分 |
| documents | 空 | 提交结果说明文档 | 命名方式：队名\_赛题\_结果说明文档.docx | 17时 |
| processCode | 空 | 存放训练代码的文件夹train\_code；存放附加题代码的文件夹addition\_code | 请注意提交文件夹；请勿修改main\_train.py及main\_addition.py的结构和函数名称 | 17时 |
| errorMsg | 空 | 无 | 此文件夹为主办方返回的finalB中代码运行的报错信息和其它导致成绩无效的信息。 | -- |

注意：errorMsg文件夹为主办方返回的finalB中代码运行的报错信息和其它导致成绩无效的信息。关于finalB中代码运行的报错信息：在12时和16时30分，以及增设的10时30分和14时30分两个时刻作为代码检查点，根据截止至这四个时刻的finalB的最新提交内容进行代码检查，届时会在此文件夹中返回参赛队伍代码运行的报错信息。