代码目录说明：

- data

  存放原始数据文件，生成的特征文件也在此。

- utils

  存放一些常用的代码文件。

  Cal_mic.py 文件是计算不同模型结果的 mic 值，并可视化为混淆矩阵形式

- feature_engineering

  特征工程，包括生成排序特征，离散特征，计数特征和缺失值处理的代码，该目录下含五份代码文件，文件名就是对应生成的特征。

- feature_select

  特征选择，该目录下有三个文件夹，分别是原始特征，排序特征，离散特征的特征选择的代码。

- M1

  模型 1 的代码，包括 4 个单模型，每个模型对应一个文件夹。

- M2

  模型 2 的代码文件，以及原始特征，排序特征，离散特征经过特征选择后所产生的特征重要性文件。

- M3

  模型融合，Ensemble.py 文件是模型加权融合的代码。

- M4

  m4.py 是迭代半监督代码

-M5

  Gen_samples.py 文件用于生成半监督所用的训练数据和无标签数据。

  Label.py 文件用于给无标签数据打标签。

  Select.py 文件选择样本添加到训练集。

运行环境要求：

-- 运行 Python 代码需要：Ubuntu，Python2.7，scikit-learn，pandas，xgboost, minepy，numpy， matplotlib

-- 运行 R 代码需要：Windows,RStudio,xgboost

-- 运行 Java 代码需要:Windows,eclipse,jdk1.7,Maven,Xgboost

代码运行步骤：

1. 将原始数据解压到 data 目录下
2. 数据预处理和特征工程

   运行 null.py 对缺失值进行处理，运行 visualize_null.py 对缺失值可视化分析。

   运行 rank.py 生成排序特征。

   运行 discretization.py 生成离散特征。

   运行 n_discretization.py 生成基于离散特征的计数特征。

   注：生成的文件都在 data 目录下

3. 特征选择

运行 feature_select/rank_feature/use_rank_feature.py，训练多个 xgb 模型并输出每个模型对特征的排序文件，再运行同目录下的 avg_featurescore.py 得到排序特征重要性的最终输出。

对于离散特征和原始特征，其代码在文件夹 feature_select/raw_feature 和 feature_select/discret_feature 下，同样地运行方法。

4. 模型 M1

运行 svm/svm_use_rank_feature.py，训练 SVM 模型，线上 auc 为 0.6938

运行 xgb717.py，训练 xgboost 模型，线上 auc 为 0.717

运行 R_7199.R，训练 xgboost 模型，线上 auc 为 0.7199

运行 M1\Java_7218\xgboost4j-demo\src\main\java\edu\cqupt\xdata\model、\Java_7218.java，训练 xgboost 模型，线上为 0.7218

5. 模型 M2

运行 solution_725.py，训练多个 xgboost 模型，再运行 avg_preds.py 对多份结果取平均，线上 auc 为 0.725 左右

6. 模型 M3

运行 ensemble.py，对模型进行加权融合，线上 auc0.7279 左右。

7.模型 M4

将 unlabel 数据作为 test 数据，运行 M3，得到 M3 对无标签数据的预测值，将该份文件置于 M4 文件夹下。

运行 m4.py，将 socre 低于阈值 a 的样本作为负样本，将 socre 高于阈值 b 的样本作为正样本，添加到训练集，再运行 M3, 如果线上得分提高，则保留这部分样本，否则继续改变阈值 ab 重复实验。

最后将保留的样本全部添加到训练集，运行 M3 得到最终结果。

线上 auc0.73 左右。

8.模型 M5

运行 Gen_samples.py 生成训练数据和无标签数据（做了特征选择）

运行 label.py 给无标签数据打标签

运行 select.py 选择线下 auc 提升最大的 top5000 无标签样本，然后从这部分样本中每次随机选择 50 个样本

将这 50 个样本添加到训练集运行 M3，如果线上得分有提升则保留样本，然后再选择 50 个样本添加到训练集，重复上述过程直到样本被选择完。

最终将保留的样本全部添加到训练集，运行 M3 得到最终结果。

线上 auc0.734 左右，因为提交次数限制，我们只添加了部分样本，如果全部添加，线上 auc 还能提升。