**财经数据各模型使用接口说明**

为方便业务人员使用机器学习模型分析财经数据，各模型代码开放了部分变量修改的接口，业务人员可以按需求修改训练参数，选取总数据集中的部分进行训练，并对新的测试数据进行验证或预测。本次提供的代码为GBRT/RF模型，暂不包含时间序列模型。以下是各接口的说明以及变量修改操作步骤。请严格按照以下说明进行操作，否则程序运行结果会出错。

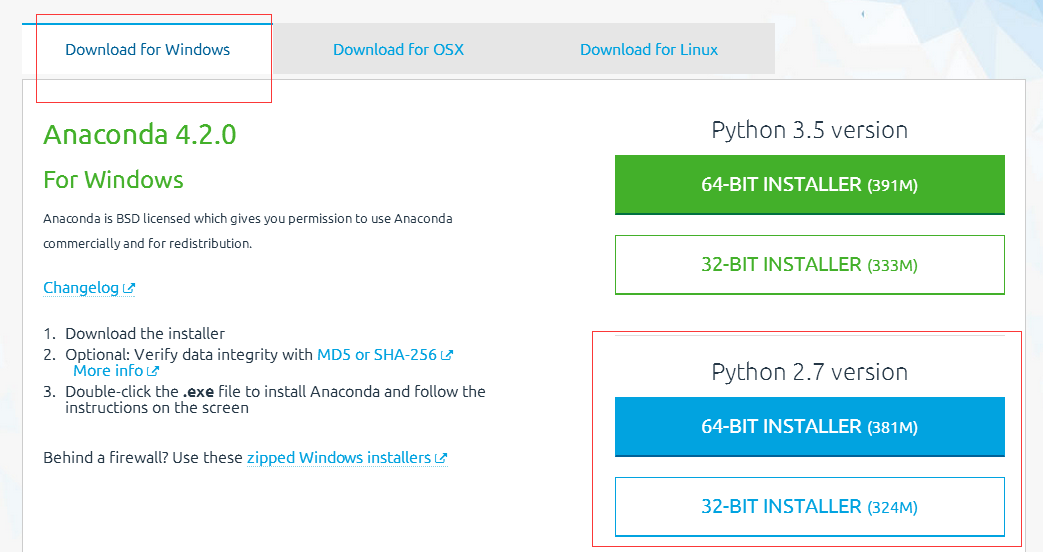
1. **接口说明**

因变量可选的GBRT/RF模型的代码，共10个训练参数可以修改

1. Data\_start: data文件中所有数据的起始年份，默认1997，当更换总数据集时才需修改。
2. Data\_end: data文件中所有数据的终止年份，默认2016，当更换总数据集时才需修改。
3. Year\_start：本次模型训练用到的训练数据的起始年份，设置时应大于等于Data\_start。
4. Year\_end: 本次模型训练用到的训练数据的终止年份，设置时应小于等于Data\_end。
5. Var\_name: 本次实验的预测目标：可选输入值：'Y1'，'Y2'，'Y3', 'Y4'，'Y5', 'Y6', 'Y7'
6. OtherFeatureList：预测该目标用到的非时变因变量的列号，以列表形式输入。具体输入格式为：OtherFeatureList = [12, 23]，其中变量为非时变因变量，也就是在1997-2016之间不随时间变化的量，例如人数，准入资本市场难度等等。里面的数字为变量在excel中的列号，例如准入资本市场难度这一项在excel中为23列，所以填入23。
7. BaseFeatureList：预测该目标用到的时变因变量的列号，以列表形式输入。格式同上，这里的变量在1997-2016中每年都有一个不同的值，所以填入这个列表中。例如Total Revenue在excel中从36列开始，到55列结束，这里填入第一次出现的列号36.具体列好如果不方便查阅可以参考excel附件，里面有详细的列号对应关系。
8. Industry\_labels：限定本次模型训练所用到公司的Industry Dummy属性，提供四种可输入值：[]，0，1，2， 其中[]代表训练数据不限定公司的Industry Dummy属性，即使用全部公司数据训练，0代表只使用Industry Dummy=0的公司的数据，1和2以此类推。
9. Dev\_label：限定本次模型训练所用到公司的发展周期属性,可输入值：[],1, 2，3。其中[]代表训练数据不限定公司的发展周期属性，即使用全部公司数据训练，0代表只使用发展周期=1的公司的数据，2和3以此类推。
10. Algo：本次训练使用的模型算法，可输入值：'gbrt'或'rf'。

Delta后GBRT/RF模型的代码，共6个训练参数可以修改

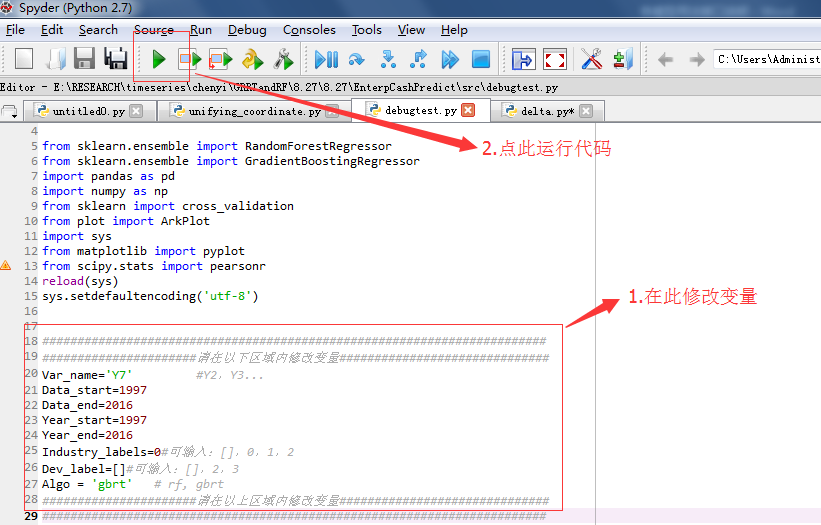
1. Var\_name: 可选输入值：'Y8'，'Y9'。
2. Data\_start: 默认1998，当更换data数据时才需修改。
3. Data\_end：默认2015，当更换data数据时才需修改。
4. Year\_start：设置时应大于等于Data\_start。
5. Year\_end：设置时应小于等于Data\_end。
6. Industry\_labels：可输入：[]或2。
7. Algo：可输入值：'gbrt'或'rf'。
8. **具体操作步骤**
9. 安装anaconda
10. 进入anaconda官方网站：<https://www.continuum.io/downloads>
11. 在主页下方找到软件下载链接，选择windows, python2.7版本，见下图。



1. 按照提示顺序安装软件。软件安装完毕以后，我们主要使用其中的Spyder工作环境来编辑运行代码。
2. 准备数据：对全量模型除华为以外的全部公司数据都在data文件夹下的data.xlsx中，华为数据在test\_data.xlsx中。增量模型除华为外所有数据都在delta\_data.xlsx中，华为数据在delta\_test\_data.xlsx中。另外，如果要新加预测数据，可以按照一下规则编写excel，继续运行即可：
   1. 按照OtherFeatureList和BaseFeatureList中的顺序组织数据，例如OtherFeatureList为[12,25],BaseFeatureList为[36],所以组织数据的顺序为[12,25,36]。
   2. 按顺序组织数据填入excel，格式为：每一列为一个变量的数据，每一行为一年的数据，其中列的顺序必须与a)中的顺序一致。
   3. 命名excel，存入到data文件夹中。并将文件名写入pred\_file\_name这个变量中。如果这个变量为[]则表示不进行预测。
   4. 具体格式样例可参考future\_Y2\_1110.xlsx
3. 打开代码文件: 找到开始菜单>所有程序>Spyder，打开Spyder环境。

在目录**\8.27\EnterpCashPredict\src**中找到代码文件，若使用变量可选的GBRT或RF模型，将文件 **’1110\_Y2-Y7削减变量——devlabel\_new.py’** 拖至Spyder工作环境中；若使用增量GBRT或RF模型，将**delta.py**拖至Spyder工作环境中。

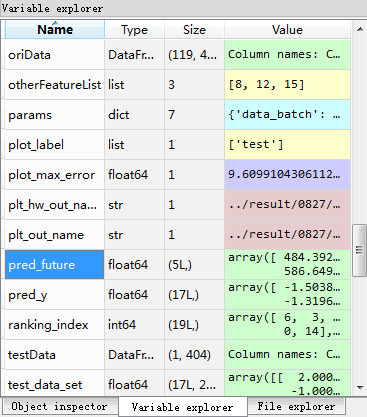
1. 修改代码：找到变量修改区，按照需求修改变量，运行代码，见下图。



1. 查看结果：

全量模型实验结果在 **8.27\EnterpCashPredict\result\1021** 中查看；增量模型实验结果在**8.27\EnterpCashPredict\result\0827** 中查看。实验结果包含以下几部分：

1. 交叉验证结果：结果文件的命名规则示例：：cv\_Y7\_industry\_label\_0\_dev\_label\_[]\_gbrt\_from\_1997\_to\_2016.txt，表示使用1997-2016年间、industry\_label =0、不限定dev\_label的公司数据、 以GBRT算法进行训练、对变量Y7进行预测的模型的交叉验证结果，具体包含四种评价误差。
2. 华为数据的测试结果：结果文件的命名规则示例：test\_Y7\_industry\_label\_0\_dev\_label\_[]\_gbrt\_from\_1997\_to\_2016.txt，表示使用1997-2016年间、industry\_label =0、不限定dev\_label的公司数据、 以GBRT算法进行训练、以华为数据为验证数据，对变量Y7进行预测的模型的验证结果，具体包括四种评价误差，特征排序。
3. 预测结果：预测结果没有输出到文件，可以在变量浏览器里面查看，找到pred\_future,双击打开即可，见下图。



**若有任何疑问，可联系：**

**陈易(外网邮箱：)**

**李小光（外网邮箱：[z5394550@163.com](mailto:z5394550@163.com), 工号：00381354）**