2019“雅典娜杯”数据挖掘邀请赛

参赛手册

“雅典娜杯”数据挖掘大赛组委会 编写

2019 年 8 月

目 录

一、赛事介绍

为响应国家大数据战略，以比赛交流和技术实战为契机，进一步加强校企合作，加快大数据人才的发掘与培养，推动金融科技发展水平，中国农业银行联手《金融电子化》杂志社，拟于 2019 年 9 月 9 日-28 日联合举办“2019 雅典娜杯数据挖掘大赛”，并邀请金融机构、科技企业、知名高校参与本次竞赛。

（一）比赛方式

比赛以线上自由组队参赛形式，每队上限 3 人，每位选手仅能加入一支队伍，通过运用数据挖掘技术和机器学习算法，针对赛题场景构建预测模型，输出预测结果，挖掘数据潜在价值。

本次大赛所有环节均在网站上进行,选手在比赛专用平台上提交模型代码并得到分数后实时显示比赛排名，最终依据有效队伍的模型结果分数进行排名，确定奖项。

网站主页地址为 https://aicontest.abchina.com（建议使用 Chrome64、IE10 以上版本浏览器），若有疑问可发送邮件至大赛邮箱（aicontest@abchina.com）。

（二）比赛平台介绍

本次比赛需使用基于 Jupyter Notebook 搭建的大赛专用平台参赛，每个队伍平台相互隔离。平台支持 Python2.7、Python3.6、R 语言，建议使用 Python3.6（Python2.7、R 语言赛事组委会不单独提供技术支持），用户可自行安装所需程序包。为了给参赛队伍一个良好的体验，平台为每支队伍分配内存 100G，CPU24 核，硬盘空间 100G。

比赛开始后，登录大赛网站点击左侧“建模平台”菜单即可进入比赛平台。**主页 Token 有效时间为两个小时，超过 Token 时间，用户提交答案前需重新登录，获取最新 Token。如果出现 403 Forbidden，说明 Token 过期，用户需要重新登录。**

登录比赛平台后，可在数据源中查看赛题数据，具体位置为项目目录 problems 文件夹下，此为共享数据集，建议拷贝到本地，请勿修改或重命名，请勿在 problems 目录下建文件！部分文件较大，请勿直接打开，可用程序载入预览。

（三）赛题介绍

本次比赛题目为贷款风险预测，基于用户基本信息、收入记录、支出记录、借贷信息、信用卡还款记录、浏览产品记录、产品基本信息等，解决用户的征信预测问题。比赛开始后，登录大赛网站点击左侧“赛题详情”即可查看数据表结构及预测结果文件格式。

（四）评选标准及规则

本次大赛采用网站自动测评的方式，根据参赛队伍提交的结果，后台自动计算分数，并实时更新大赛排行榜。

比赛结束后进行代码审核，实现模型发布并通过重复率检测。若没通过代码审核，则后面的队伍依次递补获奖。

（五）大赛流程

时　间	流　程

发布通知 （6.15 - 8.11）	7月下旬发布大赛启动宣传通知，并征集有意向参赛的单位报名登记
发送邀请信息 （8.10 - 8.25）	根据应邀单位，发送邀请函、邀请码及选手报名链接
选手报名 （8.26 - 9.8）	选手填写邀请码完成报名
分配比赛平台资源 （9.6 - 9.8）	根据报名情况分配比赛平台资源
邀 请 赛 （9.9 - 9.28）	进行邀请赛
颁 奖 （10.15 - 11.30）	大赛闭幕式，现场颁奖

二、比赛规则

1.大赛以团队形式报名，参赛团队成员人数为1-3名，报名时所有成员需提供个人基本信息。每支队伍需指定一名队长，队伍名称不超过15个字符，每名选手只能参加一支队伍，参赛团队报名之后，若有人员变动，团队可在截止日前发送回执更新参赛人员信息，报名截止后不允许更改队员名单。

2.邀请赛定向邀请部分金融机构科技部门、科技企业、知名高校。

3.禁止在指定考核技术能力的范围外利用规则漏洞或技术漏洞等不良途径提高成绩排名，禁止在比赛中抄袭他人作品，经发现将取

消比赛成绩并严肃处理。

4.不允许使用主办方提供的数据集之外的任何外部数据。

5.参赛队伍可在参赛期间随时提交代码得到A榜验证集的预测结果，一天不能超过3次，官方网站会实时更新各队伍的最新排名情况。

6.B榜测试集发布后，只有2天的时间计算预测结果，各参赛队伍要确保算法能及时计算出结果。

7.组委会保留对比赛规则进行调整修改的权利、比赛作弊行为的判定权利和处置权利、收回或拒绝授予影响组织及公平性的参赛团队奖项的权利。

三、常见问题解答

1.是否可以使用其他工具参赛，是否可以下载数据？

本次比赛必须在大赛专用平台上参赛，不能下载数据。

2.如何参加邀请赛？

- 受邀机构的成员凭邀请码登录大赛网站报名参赛。
- 有意向参赛的非受邀机构，可向组委会申请参赛，通过审核后，其成员可凭邀请码登录大赛网站报名参赛。

3.大赛报名网站无法登录怎么办？

8月1日-8月25日为邀请阶段，网站不可访问；8月26日大赛网站正式开放报名，所有参赛选手均需在网站注册登录，组队成功即为报名成功。

若8月26日之后仍不可访问，请更换chrome、火狐等高版本浏

览器。

4.若密码找回的答案忘记了怎么办？

请发送密码找回申请至大赛邮箱,并抄送至组委会网站管理员邮箱（guiyihui@abchina.com）。

5.有选手交流群吗？有问题在哪里咨询呢？

邀请赛将设立队长交流群，若有疑问请发送邮件至大赛邮箱。

6.从哪里能获得比赛通知信息呢？

选手获取通知信息渠道：a.大赛网站公告；b.“数说吧”公众号。

参赛队伍队长获取通知信息渠道：a.选手交流群；b.队长邮箱。

7.AB 榜有什么区别？

A 榜数据集优先开放，用来调算法和参数;为防止有些算法只对特定数据集做调整,影响算法正确性,所以后开放 B 榜数据集做验证。类似于 A 榜开发测试，B 榜投产上线。

8.这个比赛是必须得用 Python 编码吗，支持 sas 吗？

目前平台主要支持 Python3，包含有常见的统计分析包 numpy、pandas 等，若有额外的安装包需求，可自行安装。若有疑问可联系大赛邮箱，并抄送至组委会平台管理员邮箱（hesong@abchina.com）。

四、正式报名步骤

参赛选手必须在大赛网站注册登录用户并加入队伍才算成功报名参加雅典娜杯数据挖掘大赛，具体需要 4 步。

（一）访问比赛网站

通过 https://aicontest.abchina.com 访问比赛网站。

（二）注册参赛账号

1．填写注册信息

注意验证参赛邀请码有效。

2．接受参赛协议

阅读并接受参赛协议，完成用户注册。

（三）登录比赛网站

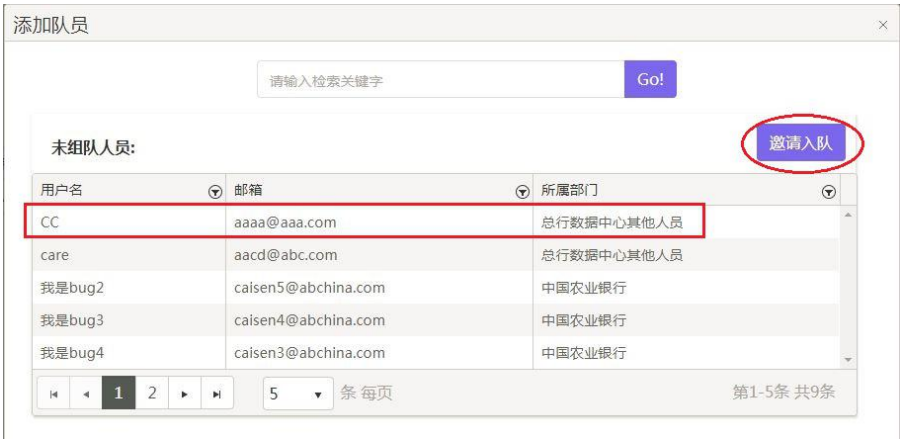
使用用户名及密码，登录比赛网站。

（四）组队参赛

参赛选手可以输入队名，创建一支新队伍，也可以选择加入已有队伍。

1. 创建队伍

如果您选择创建队伍，请输入队伍名称，点击“创建队伍”按钮完成队伍创建。之后进入“我的队伍”，点击“招募队员”按钮，给未组队人员发送入队邀请。



2．加入队伍

若您选择加入已有队伍，则点击“不！我想加入已有队伍”按钮。在可加入队伍列表中选择队伍，点击“申请加入”按钮发送入队申请。

3．确认组队信息

确认队伍名称及队员，开赛后不可变更队伍及队员信息。

- **队员接受邀请。**若您才华横溢，收到了入队邀请，请尽快处理。若同意邀请，则可顺利加入该队伍。

• **队长接受入队申请。**若您是德高望重的队长，收到了入队申请，也请尽快处理。若同意申请，则该用户顺利加入队伍成为您的队友。

恭喜您组队成功，预祝比赛取得好成绩！

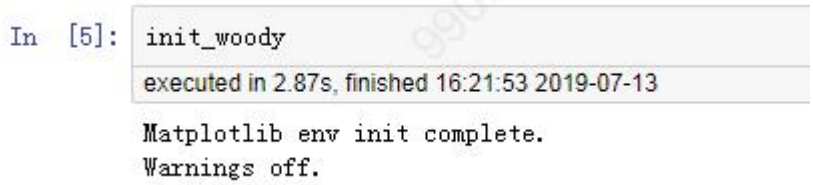
五、平台操作指南

（一）简明操作指南

简明操作指南为之前使用过数据分析挖掘平台或者 Notebook 的选手提供关键的操作说明，包括数据拷贝、Python 包安装、提交比赛结果、终止异常作业。

1. INIT_WOODY

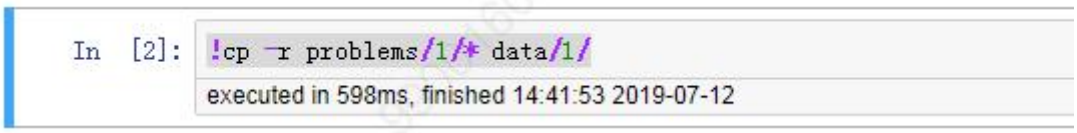
比赛平台做了一些个性化设置，用户打开一个新的页面或者在 shutdown 某个页面重新打开之后，建议先执行 INIT_WOODY 命令，如下：



2. 数据拷贝

登录比赛平台后，可在数据源中查看赛题数据，具体位置为项目目录 problems 文件夹下，此为共享数据集，建议拷贝到本地。在 Notebook Cell 里面输入命令前面加“!”可以执行 Shell 命令。

可在本地新建一个 data 文件夹，如果选手选择第一题，可以再建一个 1 的子文件夹，执行：!cp -r problems/1/* data/1/ 命令进行文件拷贝（注意文件路径，建议使用相对路径），如下：



3. 安装 Python 包

比赛平台提供了常用的分析挖掘算法包，选手可以执行!pip list 命令查看已安装的包，如下：

```
In [1]: !pip list
executed in 2.65s, finished 16:21:21 2019-07-13
DEPRECATION: The default format will switch to columns in the future.
isable this warning.
absl-py (0.1.10)
alabaster (0.7.10)
anaconda-client (1.6.9)
anaconda-navigator (1.7.0)
anaconda-project (0.8.2)
asn1crypto (0.24.0)
astor (0.6.2)
astroid (1.6.1)
astropy (2.0.3)
attrs (17.4.0)
Automat (0.6.0)
Babel (2.5.3)
backports.shutil-get-terminal-size (1.0.0)
beautifulsoup4 (4.6.0)
bitarray (0.8.1)
bkcharts (0.2)
blaze (0.11.3)
```

如果选手需要单独安装某个包，可以点击右上角“Upload”图标手动上传到比赛平台：

上传之后通过执行!pip install **** 命令进行 Python 包安装，如下：

```
In [3]: !pip install install/tqdm-4.31.1-py2.py3-none-any.whl
executed in 7.06s, finished 16:21:39 2019-07-13

Processing ./install/tqdm-4.31.1-py2.py3-none-any.whl
Installing collected packages: tqdm
Successfully installed tqdm-4.31.1
```

4. 提交比赛结果

比赛平台通过 Magic 命令进行比赛结果提交，大赛平台 Token 有效时间为两个小时，超过 Token 时间，用户提交比赛结果将出现 Token 验证失败错误。建议用户在提交比赛结果之前重新登录主页，从主页进入平台新建 Notebook 进行提交。提交之前用户需要执行 INIT_WOODY 命令，如下：

```
In [5]: init_woody
executed in 2.87s, finished 16:21:53 2019-07-13

Matplotlib env init complete.
Warnings off.
```

之后用户通过 %predict 命令进行比赛结果提交，可以输入%predict 命令查询 Magic 命令具体用法：

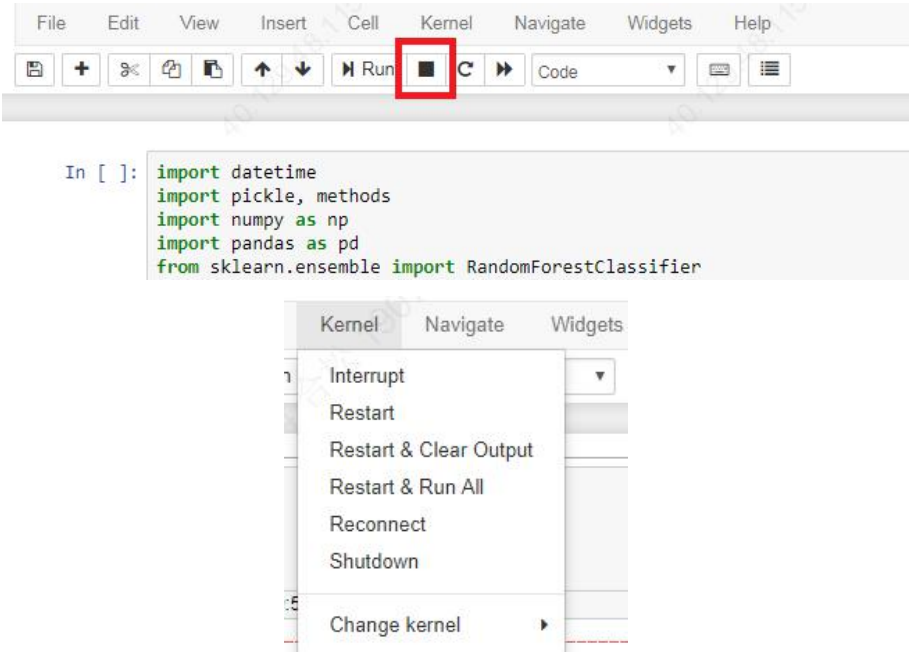
```
In [21]: %predict
executed in 5ms, finished 14:51:14 2019-07-12

Usage:
提交评分（predict）：
%predict problem_id file_name
其中problem_id为题目序号，取值为1,2,3
file_name为模型结果csv文件，建议用相对路径
示例：
%predict 1 ./out/result.csv
```

提交之后比赛结果之后，后台会自动执行算分程序，并返回算分结果，如下：

5. 终止异常作业

在实际操作过程中，时长出现某个计算作业异常（如陷入死循环），可以通过方法进行异常作业终止。

（1）选中正在执行的作业，点击菜单栏黑色方块进行作业停止，或者通过菜单栏里面相应的菜单，进行停止或者重启操作，如下：



（2）在主页点击 Running，可以看到目前正在运行的 Notebook，点击 shutdown 强制关闭（如果作业比较大，会需要一些时间）。**选手关闭前台页面，Notebook 会在后台保持运行，不会真正关闭，用户可以手动关闭不需要的 Notebook 以节约计算资源**。如下：

6. 私服使用

在 B 榜开放期间，比赛平台搭建了私有仓库，并将常用的包入库，如果代码提交阶段有新的入库需求，可以联系平台管理员进行入库操作。使用!pip install 包名 命令可以进行包安装，其余高级命令可以查看 pip 使用手册。

```
!pip install lightgbm
executed in 7.80s, finished 21:17:07 2019-08-11
Collecting lightgbm
  Downloading http://nexus.acp.paas.test.abc/platform/repository/pypi/packages/lightgbm/2.2.3/lightgbm-2.2.3-py
    100% |████████████████████████████████| 1.2MB 69.3MB/s ta 0:00:01
Requirement already satisfied: numpy in /root/anaconda3/lib/python3.6/site-packages (from lightgbm)
Requirement already satisfied: scipy in /root/anaconda3/lib/python3.6/site-packages (from lightgbm)
Requirement already satisfied: scikit-learn in /root/anaconda3/lib/python3.6/site-packages (from lightgbm)
Installing collected packages: lightgbm
Successfully installed lightgbm-2.2.3
```

7. 提交代码审核

根据组委会的安排，比赛结束后将对每题前十名进行代码审核，实现模型发布并通过重复率检测。若没通过代码审核，则后面的队伍依次递补获奖。在 B 榜开放期间到比赛结果后十天内均可以进行代码提交（B 榜开放期间可以只提交结果，待比赛结果后再提交代码）。为保证比赛公平，B 榜开放期间代码提交成功占用正式提交次数，比赛结束后不限制提交次数。

提交的代码总入口是 predict.py 文件，里面需要实现一个

predict(wfp,test_dir,train_dir)的方法，用来对 B 榜数据进行预测。其中 wfp 是结果文件写句柄，用户将预测结果写入该文件中（相当于 wfp = open('result.csv','w')）。用户直接用句柄写结果就行，写完无需关闭。test_dir 为 B 榜数据集目录（第一题示例 test_dir='./data/problems/1/B'），train_dir 为训练数据集目录（第一题示例 train_dir='./data/problems/1/train'），在实现程序中不建议使用 train_dir，邀请赛阶段将会关闭该接口，建议将训练好的模型保存下来，用模型直接来进行预测。Predict.py 示例如下：

```python
import os
import numpy as np
import pandas as pd

def predict(wfp,test_dir,train_dir):
    pass

```

另外还定义一个 requirements.txt 来定义程序依赖包，不在平台上自己安装的包需要在该文件里面申明。提交后后台将会根据 requirements.txt 里面申明依赖去私服里拉取依赖包，如果包不在私服里面，需要联系平台管理员入库。requirements.txt 文件定义如下：

```
pandas
numpy
pyarrow
lightgbm==2.1.1
graphviz>=0.10.1
```

除了这两个预定义的文件外,用户可以在同一个文件夹下面添加

自己的程序或者文件，供 predict.py 文件调用。

用户将 predict.py，requirements.txt 及自定义的文件放在一个文件夹下面（比如 mymodel 文件夹），然后通过 Magic 命令提交代码。

发布模型（release）：

%release problem_id workdir

其中 problem_id 为题目序号，取值为 1,2,3

workdir 为模型文件夹，建议用相对路径

示例：

%predict 1 ./mymodel

查看最近一次模型结果（releasestate）：

%releasestate

用户提交代码审核之后，后台将会根据 requirements.txt 里面申明依赖去私服里拉取依赖包，之后调度会向 predict.py 文件传入 wfp,test_dir,train_dir 三个参数进行模型预测，并将预测结果提交。

如果代码有错误或者依赖包不存在，后台将会报错，并保存完整的报错信息。如果执行成功，后台把 wfp 的结果提交给算分程序，并提交算分结果。由于执行时间会比较长，用户可以通过执行%releasestate 命令查询当前执行状态。

用户也可以通过自己程序先验证需要提交的代码是否正确,以减

少提交报错的概率，示例如下：

```
from model import predict
import os
try:
    fp = open('result.csv','w')
    test_dir = './data/1/A'
    train_dir=''
    predict.predict(fp,test_dir,train_dir)
except Exception as e:
    raise e
finally:
    fp.close()
```

executed in 10ms, finished 16:05:06 2019-08-02

此外，由于用户大部分是通过交互式的方式进行建模，用户可以用 ipython nbconvert 命令将 ipynb 文件转化为 py 文件，再在这个基础上进行修改，示例如下：

```
!ipython nbconvert model.ipynb --to script
```

executed in 2.27s, finished 15:57:36 2019-08-02

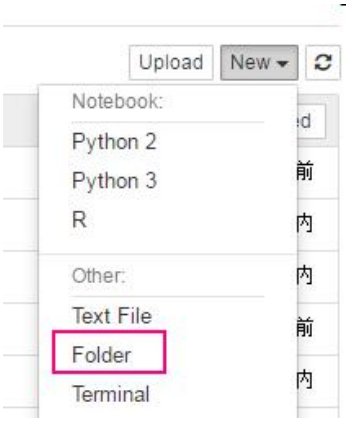
```
[TerminalIPythonApp] WARNING | Subcommand `ipython nbconvert` is deprec
[TerminalIPythonApp] WARNING | You likely want to use `jupyter nbconver
[NbConvertApp] Converting notebook model.ipynb to script
[NbConvertApp] Writing 2187 bytes to model.py
```

（二）基础操作指南

基础操作指南为没有使用过挖掘平台或者 Notebook 的选手提供一些基础的操作说明，包括新建目录、新建文件、菜单说明等。

1．新建文件夹

选择页面右上角 New→Folder，即可建立个人文件夹。



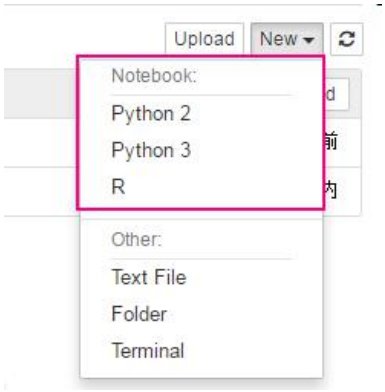
2．文件夹重命名及删除文件夹

新建个人文件夹默认名为 Untitled Folder。

勾选文件后点击页面上方的重命名与删除选项,即可对文件夹进行操作。

3．新建文件

点击文件夹即可进入文件夹进行操作，选择右上角 New→Notebook，即可新建 Python 或 R 脚本。

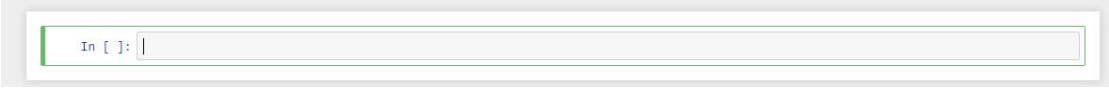


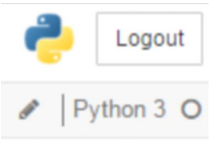
4. 修改文件名

进入文件后，点击上方文件名即可在弹出窗口修改文件名称。

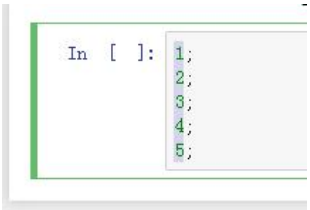
5. 编辑

进入文件后可在代码单元内编辑代码，进入编辑状态后，右上角会出现铅笔图样。





按下 Alt 键并拖拽鼠标选中列，即可实现列编辑功能。

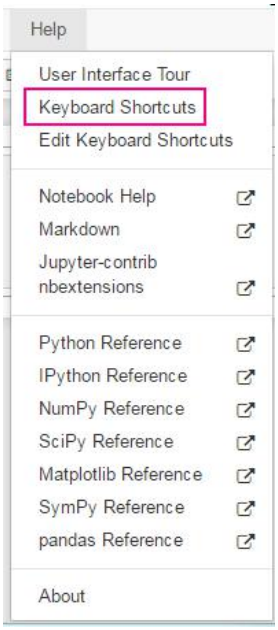


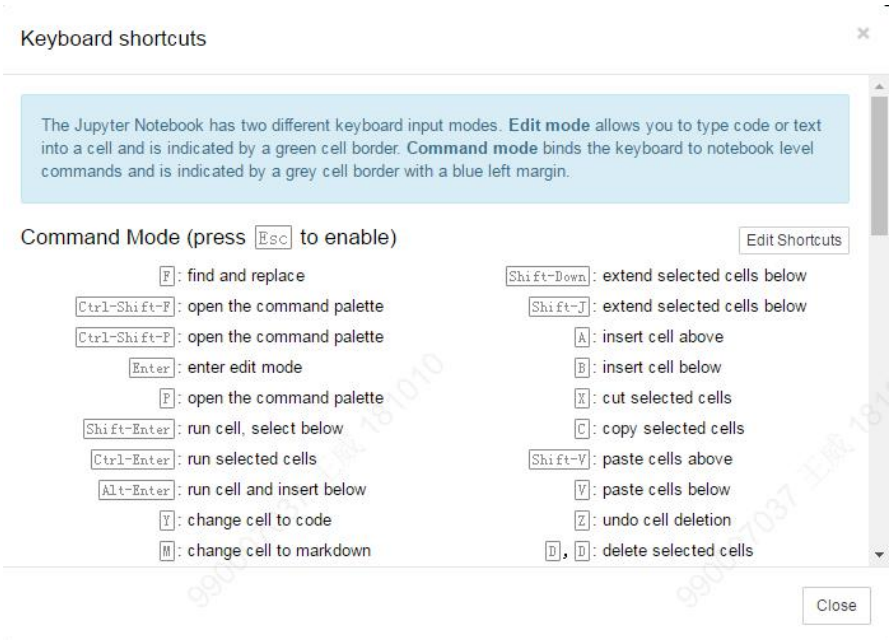
6．菜单

菜单栏内包含对文件、代码单元及内核的各种编辑选项。

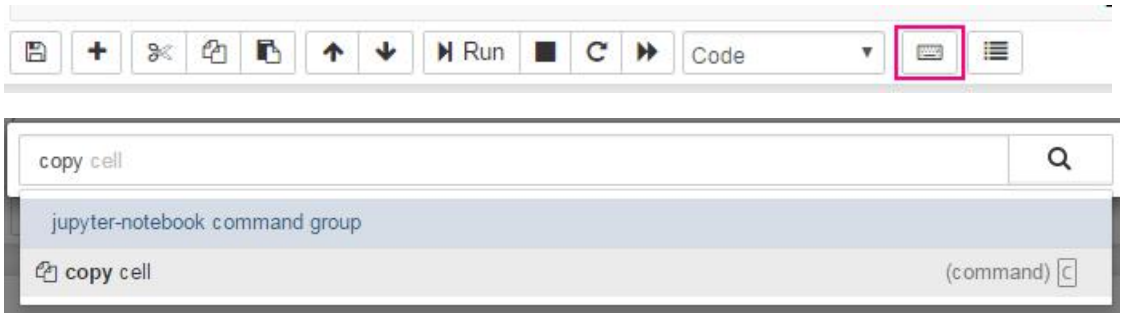
7．快捷操作

帮助下的键盘快捷键列出多种编辑快捷键，也可通过键盘编辑。





通过菜单栏或 Ctrl-Shift-P 调出命令面板，可以通过名称来运行命令。



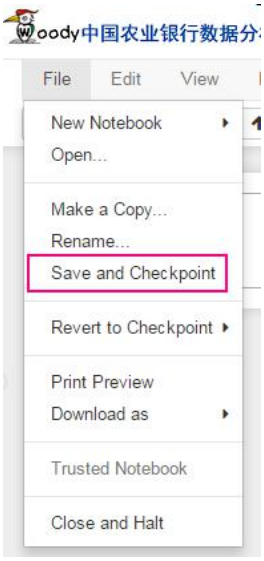
8．运行

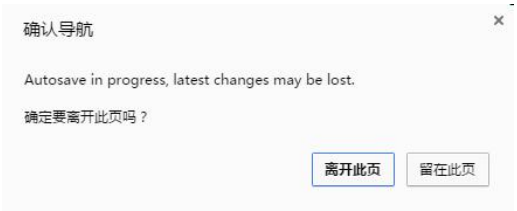
编辑后的代码可以通过菜单栏的运行键或使用快捷键 Ctrl+Enter 运行，运行结果会显示在编辑栏下方。

9．保存

比赛平台会自动保存运行后的代码及结果,对尚未保存的会在上方显示 unsaved changes。

此时若退出页面会弹出提示窗口,可通过菜单中的文件→保存并创建检查点手动保存。





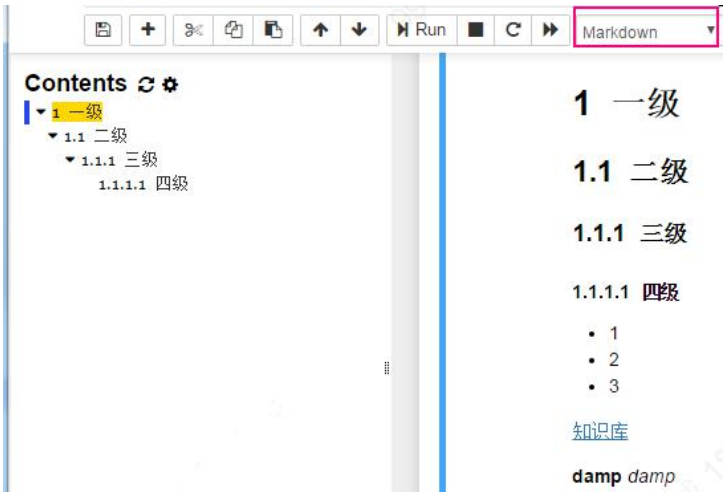
10. 上传

将文件拖拽到文件夹下。

点击 upload 即可实现上传（非 UTF-8 编码无法正常读取，文档可选择 UTF-8 编码的 txt 文件或 PDF 文件，数据可选择 txt 文件或 csv 文件）。

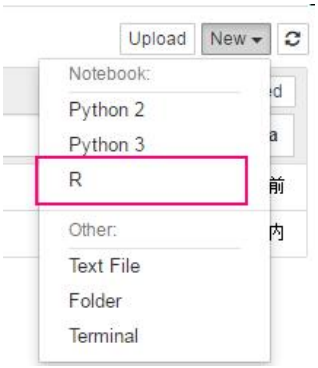
11. 使用 Markdown

选择 markdown 语言可以对文档格式进行编辑\定义目录级别、编辑表格、公式、拖拽上传图片，同时打开侧边栏可看到目录级别。



12. 使用 R 语言

新建选择 R 语言文件。



比赛平台已安装常用的 R 语言包，使用 library()即可调用。

13. Python 与 R 切换

可以在文件内进行 Python 与 R 之间的切换,如在 Python 文件中选择 Kernel→Change kernel→R，但切换后 R 中读入的数据会丢失。