

# 数据科学理论与实践课程

## 综合实战项目——A股上市公司营收大数据分析

在金融领域，每 24 小时都会产生大约 2.5 亿字节的数据，早已超过人脑处理的极限，面对全球百万亿美元的资产管理规模，行业迫切需要人工智能的加入，提升行业运行效率，让投资变得更加智能。在价值投资成为股票市场主流的背景下，准确预测公司营业收入成为投资制胜的重要法宝。买入盈利超预期的公司，避开盈利能力差的公司，就能获得超额收益。但基金经理和研究员面临的挑战是，难以高效跟踪数据，对众多上市公司营收进行准确预测。因此，我们将尝试借助算法的力量，解决这一难题。

本实验源自阿里云等主办的《FDDC2018 金融算法挑战赛-A 股上市公司季度营收预》项目，<https://tianchi.aliyun.com/competition/entrance/231660/introduction>。“FDDC2018 金融算法挑战赛”是全球首场专注金融领域，深入投资实战的技术大赛，本场大赛提供了海量金融数据、真实业务场景，世界级 AI 及金融导师指导，面向全球征集金融及算法精英，以科技摘取投资圣杯。



### 一、已知数据

【1】资产负债表：GB\_Balance\_Sheet.xlsx，分为四个 sheets：General Business,

Bank, Securities, Insurance

【2】利润表：GB\_Income\_Statement.xlsx，分为四个 sheets：General Business, Bank, Securities, Insurance

【3】现金流量表：GB\_Cash\_Flow\_Statement.xlsx，分为四个 sheets：General Business, Bank, Securities, Insurance

【4】宏观经济指标表：MacroIndustry.xlsx

【5】公司运营数据表：CompanyOperating.xlsx

【6】股票行情数据表：MarketData.xlsx

注：

1. 以上数据文件的下载为 <https://tianchi.aliyun.com/ailab/course/detail/433>（课程：数据科学——综合项目 上市公司营收大数据分析实践），由于文件太大，本文件夹不带原始文件，请自行下载。
2. 本次比赛若使用外部数据，必须向其他参赛队公开。
3. 每个数据表的字段及其含义，见数据字典文件夹——<https://tianchi.aliyun.com/ailab/course/detail/433>（课程：数据科学——综合项目 上市公司营收大数据分析实践）的学习资料。

## 二、分析任务

银行类和保险类企业的 2018 年二季度的营业收入。

注：

1. 教师提供的示例代码中为银行类企业的数据预测。
2. 学生实验代码中需要自行编写保险类企业的数据预测（可参照-1）。

## 三、分析结果的提交要求

提交格式参考示例文件 FDDC\_financial\_submit.csv，包含两列数据，分别是公司代码和二季度预测营收；预测值以百万为单位，保留两位小数。

## 四、分析结果的评价指标

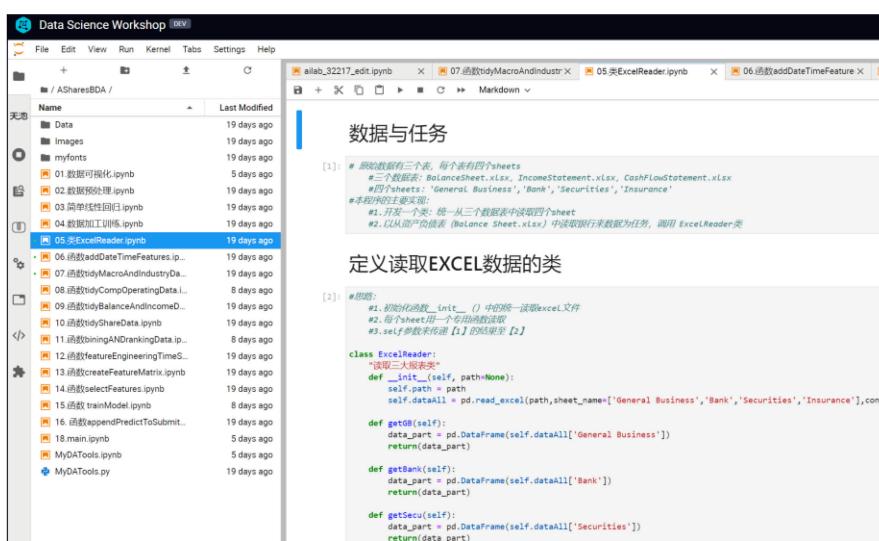
参赛队伍需要提交指定公司的二季度营收数据，以百万为单位，保留两位小数。该结果将与真实财报发布的数值进行对比。计算各个公司的相对预测误差，并进行对数市值加权，计算公式如下：

$$E_j = \frac{1}{n} \sum_{i=1}^n [\min \left( \left| \frac{y'_{ij}}{y_i} - 1.0 \right|, 0.8 \right) * \log_2^{\max(S_i, 2)}]$$

其中， $n$  为截止时间之前发布财报的公司数， $y'_{ij}$  为第  $j$  个参赛队对第  $i$  个公司的预测营收， $y_i$  为第  $i$  个公司的财报发布的营收数值， $S_i$  为第  $i$  个公司 5 月 31 号的收盘市值(精确到亿，最低两亿)。为防止某些不可控的异常事项导致财务数据变化太大，将误差上限定在 0.8。

## 五、参考解决方案及源代码

为了确保学习效果，任课教师提供了参考解决方案，解决方案及全套源代码从 <https://tianchi.aliyun.com/ailab/course/detail/433> (课程：数据科学——综合项目 上市公司营收大数据分析实践) 下载。



The screenshot shows a Jupyter Notebook interface within the Data Science Workshop. The left sidebar lists files in the 'AsharesBDA' directory, including various Python scripts and notebooks. The main area displays two code snippets:

**数据与任务**

```
[1]: # 读取前三张表，每个表有四张sheet
# 三个sheet: 'BalanceSheet', 'IncomeStatement', 'CashFlowStatement'
#sheet: 'General Business', 'Bank', 'Securities', 'Insurance'
#本节代码的主要实现:
#1. 开发一个类，统一从三个数据表中读取四张sheet
#2. 以从资产负债表(Balance Sheet.xlsx)中读取银行余额为任务，调用 ExcelReader类
```

**定义读取EXCEL数据的类**

```
[2]: #思路:
#1. 初始化函数__init__( ) 中的参数读取excel文件
#2. 每个sheet用一个专门的数据类
#3. self参数来存储 [1] 的结果里 [2]
class ExcelReader:
    "读取三大报表类"
    def __init__(self, path=None):
        self.path = path
        self.dataAll = pd.read_excel(path,sheet_name=['General Business','Bank','Securities','Insurance'])

    def getGB(self):
        data_part = pd.DataFrame(self.dataAll['General Business'])
        return(data_part)

    def getBank(self):
        data_part = pd.DataFrame(self.dataAll['Bank'])
        return(data_part)

    def getSecu(self):
        data_part = pd.DataFrame(self.dataAll['Securities'])
        return(data_part)
```

## **六、课程作业的提交要求**

1. 提交实验报告（有模板）
2. 准备演示 PPT（有模板）
3. 程序代码（两个 ipynb 文件，含教师版本和学生版本两个文件，需要提供必要的注释）
4. 课堂演示及现场答辩