课程概述

您好！欢迎来到数据分析中级课程，本课程是中级部分的第一个案例，**新三板：公司交易预测。**

在熟悉了数据分析的流程后，中级课程将把重点放在数据处理和模型的构建上，主要使用 scikit-learn 机器学习库来实现数据分析过程。跟初级课程相比，案例的难度有了非常大的提升，需要付出更多的努力，当然也会收获更多。希望您可以带着探索精神与敏捷的数据感知力开始本节课程的学习。

1. 内容简介

案例的大体结构不变，仍然秉承着问题界定、数据准备、数据建模及数据可视化的课程结构展开。

1）问题界定

通过对“新三板”背景知识和业务内容的了解，构造主业务问题，同时结合字段检视提取与本案例相关的数据。

2）数据准备

通过对数据数量、字段情况和数据分布情况的进一步了解，提出符合本案例完整的数据改造方案，进而对数据进行预处理、分割与重构，最终获得满足建模需要的数据。

跟初级课程相比，本案例涉及的建模字段有几十个之多，在具体代码实现方面引入了很多数据预处理方法，包括 SimpleImputer、OneHotEncoder、PowerTransformer、StandardScaler、FunctionTransformer，并使用 DataFrameMapper 直接将这些方法应用到案例数据集上。另外，数据重构也是本案例的重点内容之一。

3）数据建模与数据可视化

构建逻辑归回模型，通过网格搜索超参数寻找最优参数组合，并通过混淆矩阵和 ROC 曲线对模型进行评估，最终得到影响新三板公司交易是否发生的公司特征及其权重。

2. 学习目标

学习完本次案例，你能够达到以下目标：

1）能够根据业务问题和数据情况提出合理的数据改造方案；

2）能够根据数据改造方案，使用 sklearn 库中的方法进行数据处理；

3）能够使用逻辑回归模型进行建模和分析，并结合业务知识合理解释得到的结果。

业务现状描述

1. 新三板简介

“新三板”的前身“老三板”成立于 2001 年 7 月，也叫“代办股份转证系统”，最初专为承接 A 股退市公司所设。在将中关村科技园区纳入其版图后，为方便区分，特将这些新兴企业称为“**新三板**”，它是继上交所、深交所后的第三个全国性股权交易市场，主要为创新型、创业型、成长型的中小微企业发展提供服务。

截至 2015 年底挂牌公司总计已达 4514 家，远超沪、深两市总和。最新统计，截至 2018 年 11 月 16 日，新三板挂牌公司累计成交金额 706.80 亿元，挂牌公司总计 10828 家，共有 1229 家公司完成股票发行融资，累计发行股票金额达 540.59 亿元。

2. 业务内容

间接融资、定向增发、股票交易是新三板市场中中小企业获得资金很重要的手段，间接融资往往是公司向银行机构或者个人贷款，负债能增加公司利用资金获利的能力；定向增发是指投资人与公司达成协议之后进行大金额投资；交易即公司股票在市场上随机流动，类似于 A 股股票。间接融资、股票交易两项指标对于判别公司资金流动性和融资成果具有重要意义，能够指导投资人参与公司的定向增发活动。

对于企业来说，在新三板挂牌，不但门槛低（两年的股份制公司即可），而且耗时少（通常申报后 42 天可完成审批），更有政府补贴拿；然而对于投资者来说，新三板尚属于“奢侈品”，个人投资者想要入手新三板，不仅要有 2 年以上的证券投资经验，本人名下最近 10 个转证日的日均金融资产①还要求在 500 万元人民币以上。

3. 困境

新三板市场定位不清，市场流动性不足，许多中小企业仍然很难在新三板市场中成

①金融资产是指银行存款、股票、债券、基金份额、资产管理计划、银行理财产品、信托计划、

保险产品、期货及其他衍生产品等。

功融资，公司鱼龙混杂质量参差不齐；对于投资人而言，没有明确有效的盈利模式，现在常见的有包括股票交易赚取短期差价、扶持 IPO②上市退出股份。但是在新三板市场中，投资人较少、资金不足，股票交易十分有限，投资人很难通过快速买入卖出获利；另一方面仅有极少量的公司能够 IPO 上市成功，大部分企业不仅没有前景上市，甚至会因为违规违纪等因素无法完成业务目标退市。

但总体而言，新三板完善了我国多层次证券市场，随着衡策制度的不断改进，“新三板”未来的表现值得期待。

②首次公开募股（Initial Public Offerings，简称 IPO）是指一家企业或公司（股份有限公司）第一次将它的股份向公众出售(首次公开发行,指股份公司首次向社会公众公开招股的发行方式)。

业务问题构造

数据分析的目标通常都是为了回答某些业务问题，因此，构造符合业务现状以及利益相关群体需求的问题是数据分析的基础。

1. 利益相关群体

上文提到：间接融资、定向增发、股票交易是新三板市场中小微企业资金流动性和融资成果的重要手段，在本次数据分析项目中针对股票是否交易的背景因素进行探讨，主要涉及两个群体**公司**和**投资人**，对于公司和投资人而言，新三板中市场因素繁多，公司质量参差不齐，投资盈利模式尚不成熟，亟需数据分析的思路和结果提供相关投资参考。同时通过对公司这一群体的行为信息特征及企业背景进行分类分析，帮助投资人更加了解新三板市场，综合多方面特征发现优质公司，预测公司未来走向，从而帮助其更好地制定投资策略等。

2. 业务问题

从投资人的角度来看，**股票的流动性**是反映公司被市场或者其他投资人接受认可的程度，同时也能够暗示投资变现的能力。

流动水平越高，股东投资变现的能力越高，对投资人而言，公司的价值就越高。

而反映股票流动性的一个直接指标：公司是否会发生股票交易。

因此，本次案例的业务问题定为：

“基于逻辑回归模型，使用公司特征预测公司是否发生交易。”

字段检视

了解字段的含义是数据理解的第一步，其目的在于熟悉数据字段中所描述的对象及关系，同时可以将数据与业务联系起来，方便后续从业务角度来取用数据。不同的业务问题或者子问题的解决，需要使用不同的数据，将问题与所需数据字段对应可以更好地帮助我们进行后续的数据获取。

表 1：数据表定义及说明

英文表名	中文表名	备注
q_company_code	部分公司代码表	共有 5127 条公司记录，这些公司具有 2016 年年报，而且在各年解析、抽取（限以下数据表字段）中没有错误记录
q_company_code_keep_2018_08	留存至 2018 年 8 月的公司代码表	NEEQ 系统上下载 2018 年 8 月公司行业分类表获得当期存续公司列表，共 4252 家公司保留，可以获得哪些公司后来离开了新三板（摘牌+转板）
q_company	公司信息表	公司代码表所对应公司信息，每年每个公司一条记录，有所冗余，代表了每年公司基本情况
q_annoucement	公司公告表	目前没有把所有公告记录保留，只保留了年报记录，如果要针对公告表现聚类则需要对应公司的所有公告记录
q_indirect_financing	间接融资表	共有 3755 家公司发生过间接融资，记录了每个公司的融资情况，每一条记录是一次公司融资
q_customers	公司客户表	公司客户情况，年报仅记录前五大客户，表中每条记录代表某公司的一个客户
q_suppliers	公司供应商表	公司供应商情况，类似于客户
q_holders	公司股东表	公司股东情况，记录每年公司的所有股东，数据可能有空缺
q_executives	公司高管表	公司高管情况，与股东情况类似
q_trading_quotations_valid	有效交易表	每日每个公司一条记录，表中只包含有交易的日期，即 trading_amount!=0，共有 3338 个公司发生过交易
q_company_finance	公司财务数据表	公司财报数据，采用长表的方式记录，项目名称被记录在 name 列中
Q_OUTWAY	公司去向表	公司最终去向情况

注：数据来源于现有公开网络的记录爬取所得，共 12 张数据表，160 个字段，同时这些字段将支撑新三板的 3 个案例。

本案例的业务问题为："**根据公司特征预测公司是否会发生股票交易**"

与业务问题相关联的数据：①公司的不同特征

②公司是否发生股票交易记录

金融数据分析中指标特征丰富，结构复杂，往往在分析过程中先采用尽量多的指标并在后续筛选强相关指标进行分析。

基于业务理解，我们计划提取的 47 个字段如下表所示：

表 2 案例提取字段表

范畴	英文名称	特征名称	范畴	英文名称	特征名称
基本情况	company_code	公司代码	财务业绩	business_income	营业收入
	company_name	公司名称		gross_profit_rt	毛利率
	industry	第三级投资型行业		net_profit_rt	净利润
	listing_days	挂牌时长		roa	净资产收益率
	transfer_mode	转让方式		asset	资产总计
	layer	公司分层		liability	负债总计
	province	省份		net_asset_per_share	每股净资产
	maker_num	做市商数量		earning_per_share	基本每股收益
	registered_capital	注册资本		asset_liablity_rt	资产负债率
	register_days	注册时长		current_rt	流动比率
	employee_num	雇员数量		gross_cash_flow	现金流量总额
	board_num	董事会人数		asset_gr	总资产增长率
	supervisor_num	监事会人数		income_gr	营业收入增长率
	executive_num	高管人数		net_profit_gr	净利润增长率
	accounting_firm	会计事务所		ave_executive_age	高管平均年龄
	broker	主办券商		ave_executive_edu	高管平均学历
供应商客户	customer_rt_1st	第一大客户占比	人员	highest_executive_edu	高管最高学历
	customer_rt_5all	五大客户占比		holder_rt_1st	第一大股东占比
	supplier_rt_1st	第一大供应商占比		holder_rt_10all	前十大股东占比
	supplier_rt_5all	五大供应商占比		contain_org_holder	存在机构股东
交易	trading_days	交易日期		financing_amount_wan	融资金额（万）
	trading_price	交易平均价格	其他情况	financing_cnt	融资次数
	total_transaction_amount	交易总金额		outway	公司去向
	total_stock_equity	总股本			

本次案例的主要任务就是利用以上字段，建立模型，对公司是否会发生交易进行分类。常用的分类算法很多，本次案例我们选择**逻辑回归模型**。

补充资料：

数据库中设置的字段类型主要包括为数字、文本、布尔值（true/false）、时间（如YYYY/MM/DD）、位置（经纬度）等。

这些字段类型在一般的数据库中都有对应的表示方法：
text-文本(在形式上和 varchar 基本相同),float-浮点数,datetime-时间,int-整数,varchar-可变长字符串，bigint-长整数（取值范围比 int 大），smallint-短整数（取值范围比 int 小），括号内的数字如果形如 XX（a），表示设定字段的最大长度为 a；如果形如 XX（a，b），表示对于小数设定小数点前的数字最大长度为 a,小数点后数字的最大长度为 b。decimal-小数(与 float、double 不同的是，以字符串的形式保存数值)。

数据获取

在确定数据与业务关联后，就可以采用一定的数据获取技术从一个或多个数据库中获取所需的数据字段，最常用的数据库查询方法是 **SQL 语言**。

SQL（Structured Query Language）是用于访问和处理数据库的标准的计算机语言。SQL 的最基础的语法包含 SELECT 语句、DISTINCT 关键词、WHERE 子句等。根据数据获取内容的不同 SQL 语句的复杂程度会随之改变。

本案例从现有的 12 个表中获取上本 47 个字段形成表 feature.xlsx，但表中字段 *accounting_firm*（该公司的会计师事务所）和字段 *broker*（该公司的证券公司）存在名称统一问题，经人工比照后对表中会计师事务所和主办债商的名称进行规划化后，在 excel 里使用 vlookup 函数在表 feature 中生成新字段 local_accounting_firm 和 local_broker。

因此通过 sql 导出的是 features.xlsx 中的 47 个字段，同时为规范两个字段的名称，在原表中又多加了两个字段，共 49 个字段。

注：更多 SQL 语法可以参考网上教程（如 http://www.w3school.com.cn/sql/index.asp）