机器学习项目目录

1、航空公司客户价值分析

1、分析航空公司现状

- 目前航空公司已积累了大量的会员档案信息和其乘坐航班记录。
- 以 2014-03-31 为结束时间，选取宽度为两年的时间段作为分析观测窗口，抽取观测窗口内有乘机记录的所有客户的详细数据形成历史数据，44 个特征，总共 62988 条记录。数据特征及其说明如右表所示。

	特征名称	特征说明
客户基本信息	MEMBER_NO	会员卡号
	FFP_DATE	入会时间
	FIRST_FLIGHT_DATE	第一次飞行日期
	GENDER	性别
	FFP_TIER	会员卡级别
	WORK_CITY	工作地城市
	WORK_PROVINCE	工作地所在省份
	WORK_COUNTRY	工作地所在国家
	AGE	年龄

表 名	特征名称	特征说明
乘机信息	FLIGHT_COUNT	观测窗口内的飞行次数
	LOAD_TIME	观测窗口的结束时间
	LAST_TO_END	最后一次乘机时间至观测窗口结束时长
	AVG_DISCOUNT	平均折扣率
	SUM_YR	观测窗口的票价收入
	SEG_KM_SUM	观测窗口的总飞行公里数
	LAST_FLIGHT_DATE	末次飞行日期
	AVG_INTERVAL	平均乘机时间间隔
	MAX_INTERVAL	最大乘机间隔
积分信息	EXCHANGE_COUNT	积分兑换次数
	EP_SUM	总精英积分
	PROMOPTIVE_SUM	促销积分
	PARTNER_SUM	合作伙伴积分
	POINTS_SUM	总累计积分
	POINT_NOTFLIGHT	非乘机的积分变动次数
	BP_SUM	总基本积分

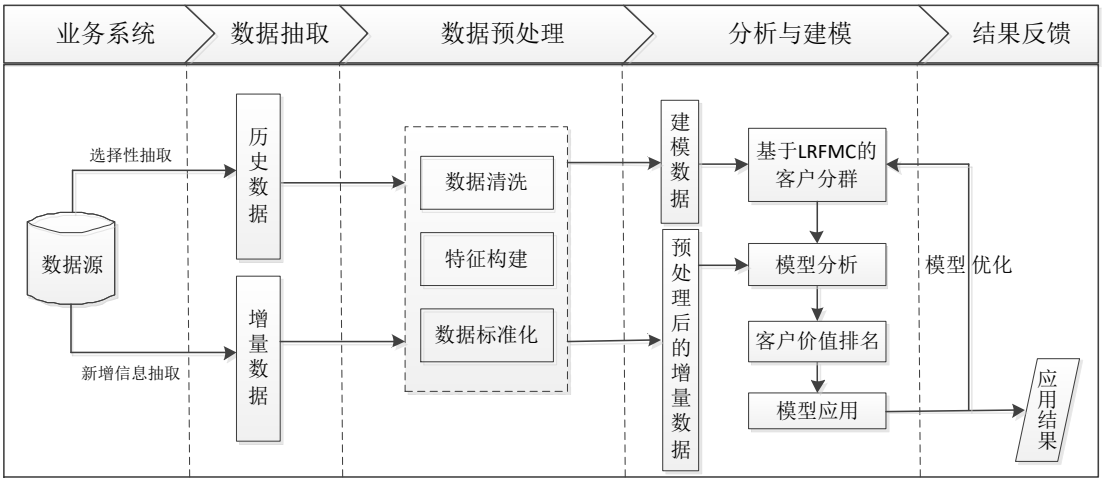
2、项目目标

结合目前航空公司的数据情况，可以实现以下目标。

- 借助航空公司客户数据，对客户进行分类。
- 对不同的客户类别进行特征分析，比较不同类别客户的客户价值。
- 对不同价值的客户类别提供个性化服务，制定相应的营销策略。

3、熟悉航空客户价值分析的步骤与流程

航空客户价值分析项目的总体流程如图所示。



2、财政收入预测分析

1、 财政收入简介和需求

财政收入，是指政府为履行其职能、实施公共政策和提供公共物品与服务需要而筹集的一切资金的总和。财政收入表现为政府部门在一定时期内（一般为一个财政年度）所取得的货币收入。财政收入是衡量一国政府财力的重要特征，政府在社会经济活动中提供公共物品和服务的范围和数量，在很大程度上取决于财政收入的充裕状况。

在我国现行的分税制财政管理体制下，地方财政收入不但是国家财政收入的重要组成部分，而且具有其相对独立的构成内容。如何制定地方财政支出计划，合理分配地方财政收入，促进地方的发展，提高市民的收入和生活质量是每个地方政府需要考虑的首要问题。因此，地方财政收入预测是非常必要的。

2、财政收入预测数据基础情况

考虑到数据的可得性，本项目所用的财政收入分为地方一般预算收入和政府性基金收入。地方一般预算收入包括以下 2 个部分。

- 税收收入。主要包括企业所得税与地方所得税中中央和地方共享的 40%，地方享有的 25%的增值税，营业税和印花税等。
- 非税收收入。包括专项收入、行政事业性收费、罚没收入、国有资本经营收入和其他收入等。

政府性基金收入是国家通过向社会征收以及出让土地、发行彩票等方式取得收入，并专项用于支持特定基础设施建设和社会事业发展的收入。

由于 1994 年我国对财政体制进行了重大改革，开始实行分税制财政体制，影响了财政收入相关数据的连续性，在 1994 年前后不具有可比性。由于没有合适的方法来调整这种数据的跃变，因此本项目仅对 1994 年至 2013 年的数据进行分析（本项目所用数据均来自《统计年鉴》）。

各项特征名称及特征说明如下（共 13 项）：

- **社会从业人数(x1)**：就业人数的上升伴随着居民消费水平的提高，从而间接影响财政收入的增加。

- **在岗职工工资总额(x2)**：反映的是社会分配情况，主要影响财政收入中的个人所得税、房产税以及潜在消费能力。
- **社会消费品零售总额(x3)**：代表社会整体消费情况，是可支配收入在经济生活中的实现。当社会消费品零售总额增长时，表明社会消费意愿强烈，部分程度上会导致财政收入中增值税的增长；同时当消费增长时，也会引起经济系统中其他方面发生变动，最终导致财政收入的增长。
- **城镇居民人均可支配收入(x4)**：居民收入越高消费能力越强，同时意味着其工作积极性越高，创造出的财富越多，从而能带来财政收入的更快和持续增长。
- **城镇居民人均消费性支出(x5)**：居民在消费商品的过程中会产生各种税费，税费又是调节生产规模的手段之一。在商品经济发达的如今，居民消费的越多，对财政收入的贡献就越大。
- **年末总人口(x6)**：在地方经济发展水平既定的条件下，人均地方财政收入与地方人口数呈反比例变化。
- **全社会固定资产投资额(x7)**：是建造和购置固定资产的经济活动，即固定资产再生产活动。主要通过投资来促进经济增长，扩大税源，进而拉动财政税收收入整体增长。
- **地区生产总值(x8)**：表示地方经济发展水平。一般来讲，政府财政收入来源于即期的地区生产总值。在国家经济政

策不变、社会秩序稳定的情况下，地方经济发展水平与地方财政收入之间存在着密切的相关性，越是经济发达的地区，其财政收入的规模就越大。

- **第一产业产值(x9)**：取消农业税、实施三农政策，第一产业对财政收入的影响更小。
- **税收(x10)**：由于其具有征收的强制性、无偿性和固定性特点，可以为政府履行其职能提供充足的资金来源。因此，各国都将其作为政府财政收入的最重要的收入形式和来源。
- **居民消费价格指数(x11)**：反映居民家庭购买的消费品及服务价格水平的变动情况，影响城乡居民的生活支出和国家的财政收入。
- **第三产业与第二产业产值比(x12)**：表示产业结构。三次产业生产总值代表国民经济水平，是财政收入的主要影响因素，当产业结构逐步优化时，财政收入也会随之增加。
- **居民消费水平(x13)**：在很大程度上受整体经济状况 GDP 的影响，从而间接影响地方财政收入。

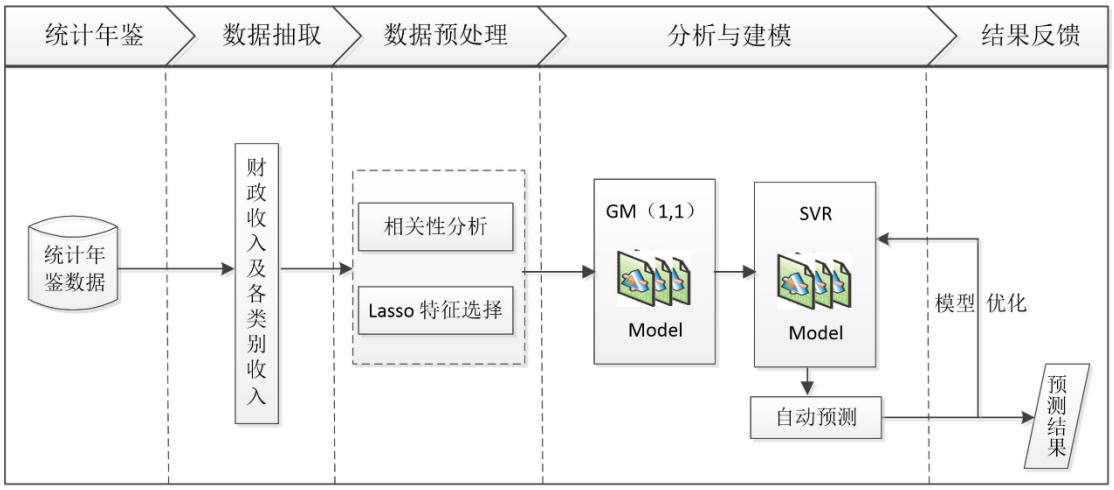
3、财政收入预测分析目标

结合财政收入预测的需求分析，本次数据分析建模目标主要有以下 2 个。

- 分析、识别影响地方财政收入的关键特征。

➢ 预测 2014 年和 2015 年的财政收入。

4、熟悉财政收入预测的步骤与流程



3、家用热水器用户行为分析与事件识别

1、了解热水器采集数据基本情况

国内某热水器生产厂商新研发的一种高端智能热水器，在状态发生改变或者有水流状态时，会采集各项数据。抽取 200 个热水器用户的用水记录作为原始建模数据，热水器采集到用户用水数据如下表所示。

热水器编号	发生时间	开关机状态	加热中	保温中	有无水流	实际温度	热水量	水流量	节能模式	加热剩余时间	当前设置温度
R_00001	20141019160855	开	开	关	无	47°C	25%	0	关	4分钟	50°C
R_00001	20141019160954	开	开	关	无	47°C	25%	0	关	2分钟	50°C
R_00001	20141019161040	开	开	关	无	48°C	25%	0	关	2分钟	50°C
R_00001	20141019161042	开	开	关	无	48°C	25%	0	关	1分钟	50°C
R_00001	20141019161106	开	开	关	无	49°C	25%	0	关	1分钟	50°C
R_00001	20141019161147	开	开	关	无	49°C	25%	0	关	0分钟	50°C
R_00001	20141019161149	开	关	开	无	50°C	100%	0	关	0分钟	50°C
R_00001	20141019172319	开	关	开	无	50°C	50%	0	关	0分钟	50°C

2、热水器数据特征说明

热水器采集的用水数据包含 12 个特征：热水器编码，发生时间，开关机状态，加热中，保温中，有无水流，实际温度，热水量，水流量，节能模式，加热剩余时间和当前设置温度。其解释说明如下表所示。

特征名称	说明
热水器编码	热水器出厂编号
发生时间	记录热水器处于某状态的时刻
开关机状态	热水器是否开机
加热中	热水器处于对水进行加热的状态
保温中	热水器处于对水进行保温的状态
有无水流	热水水流量大于等于10L/min为有水，否则为无
实际温度	热水器中热水的实际温度
热水量	热水器热水的含量
水流量	热水器热水的水流速度，单位：L/min
节能模式	热水器的一种节能工作模式
加热剩余时间	加热到设定温度还需多长时间
当前设置温度	热水器加热时热水能够到达的最大温度

3、熟悉家用热水器用户行为分析的步骤与流程

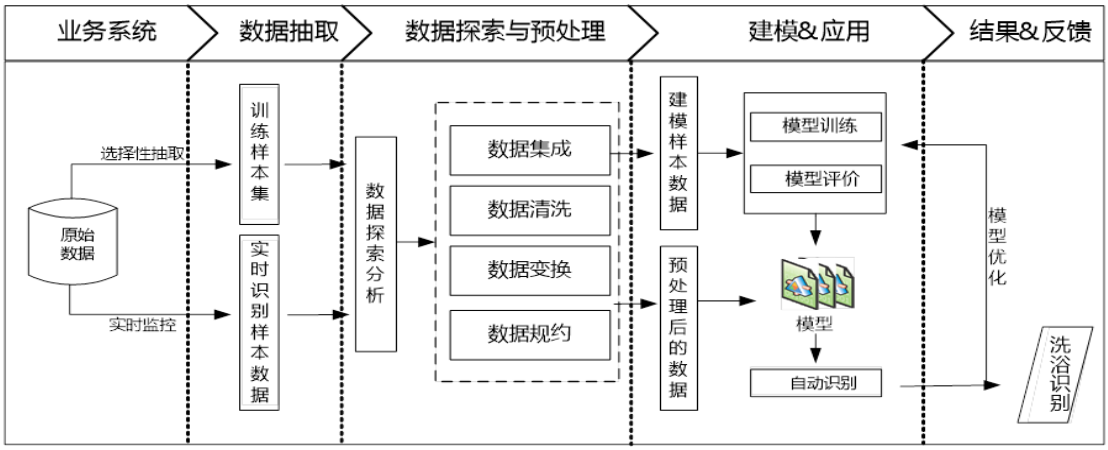
3.1 用水事件识别

在热水器用户行为分析过程中，用水事件识别是最为关键的环节。根据该热水器生产厂商提供的数据热水器用户用水事件划分与识别项目的整体目标如下。

（1）根据热水器采集到的数据，划分一次完整用水事件。

（2）在划分好的一次完整用水事件中，识别出洗浴事件。

3.2 总体流程



4、通信运营商用户流失分析与预测

1、通信运营商现状与需求

随着业务的快速发展、移动业务市场的竞争愈演愈烈。如何最大程度地挽留在网用户、吸取新客户，是电信企业最关注的问题之一。竞争对手的促销、公司资费软着陆措施的出台和政策法规的不断变化，影响了客户消费心理和消费行为，导致客户的流失特征不断变化。对于电信运营商而言，流失会给电信企业带来市场占有率下降、营销成本增加、利润下降等一系列问题。在发展用户每月增加的同时，如何挽留和争取更多的用户，是一项非常重要的工作。

随着机器学习技术的不断发展和应用，移动运营商希望能借助机器学习算法识别哪些用户可能流失，什么时候会发生流失。而通过建立流失预测模型，分析用户的历史数据和当前数据，提取辅助决策的关键性数据，并从中发现隐藏关系和模式，进而预测未来可能发生的行为，就可以帮助移动运营商实现这些要求。

2、通信运营商数据基本状况

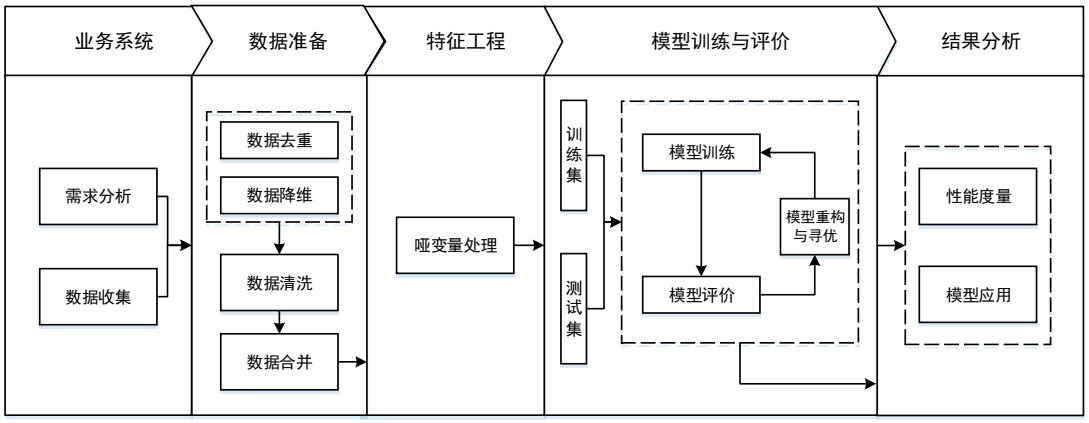
某运营商提供了不同用户的三个月使用记录共 900000 条数据，共 34 个特征，其中存在着重复值，缺失值与异常值，其字段说明如下表所示。

名称	字段描述
MONTH_ID	月份
USER_ID	用户ID
INNET_MONTH	在网时长
IS_AGREE	是否合约有效用户
AGREE_EXP_DATE	合约计划到期时间
CREDIT_LEVEL	信用等级
VIP_LVL	VIP等级
ACCT_FEE	本月费用（元）

名称	字段描述
CALL_DURA	通话时长（秒）
NO_ROAM_LOCAL_CALL_DURA	本地通话时长（秒）
NO_ROAM_GN_LONG_CALL_DURA	国内长途通话时长（秒）
GN_ROAM_CALL_DURA	国内漫游通话时长（秒）
CDR_NUM	通话次数（次）
NO_ROAM_CDR_NUM	非漫游通话次数（次）
NO_ROAM_LOCAL_CDR_NUM	本地通话次数（次）
NO_ROAM_GN_LONG_CDR_NUM	国内长途通话次数（次）
GN_ROAM_CDR_NUM	国内漫游通话次数（次）
P2P_SMS_CNT_UP	短信发送数（条）
TOTAL_FLUX	上网流量（MB）
LOCAL_FLUX	本地非漫游上网流量（MB）
GN_ROAM_FLUX	国内漫游上网流量（MB）
CALL_DAYS	有通话天数

名称	字段描述
CALLING_DAYS	有主叫天数
CALLED_DAYS	有被叫天数
CALL_RING	语音呼叫圈
CALLING_RING	主叫呼叫圈
CALLED_RING	被叫呼叫圈
CUST_SEX	性别
CERT_AGE	年龄
CONSTELLATION_DESC	星座
MANU_NAME	手机品牌名称
MODEL_NAME	手机型号名称
OS_DESC	操作系统描述
TERM_TYPE	终端硬件类型（0=无法区分，4=4g、3=3g、2=2g）
IS_LOST	用户在3月是否流失标记（1=是，0=否），1月和2月值为空

3、通信运营商客户流失分析与预测的步骤与流程



5、某移动公司客户价值分析

1、情景问题提出及分析

传统移动通信业对客户的管理以往是基于经验统计划分，无法细分有意义的高价值贡献客户，差异化服务得不到预想的效果。因此，建立灵活、精确的客户价值评估体系来辨别高价值客户，提供差异化营销服务成为了当前通信业发展的关键。本项目将以移动公司客户价值为研究对象，根据通信企业客户的特征及业务特点，利用 Python 来对客户信息进行分析，对客户群体进行分类，分析预测客户的潜在消费行为，对客户进行价值评估，在自己的客户群体中挖掘出特有的潜在客户。

2、客户价值分析过程

移动客户价值分析可以结合 RFM 分析模型和 K-Means 聚类算法共同对数据进行分析费型建分为以下 4 步。

第一步，读入数据并进行预处理。这一步的主要目的是处理读入的数据集中的缺失值、重复值以及无效值等，并对数据做量化处理。

第二步，根据客户价值分析中常用的 RFM 分析模型来对相应数据进行特征提取和标准化处理。

第三步，读入预处理好和标准化处理后的数据，结合 RFM 分析模型所计算的特征列使用 K-Means 聚类算法对客户进行聚类分析。

第四步，对聚类结果和相关数据进行数据可视化和数据分析。

本项目的数据集由阿里天池公开数据集获得。

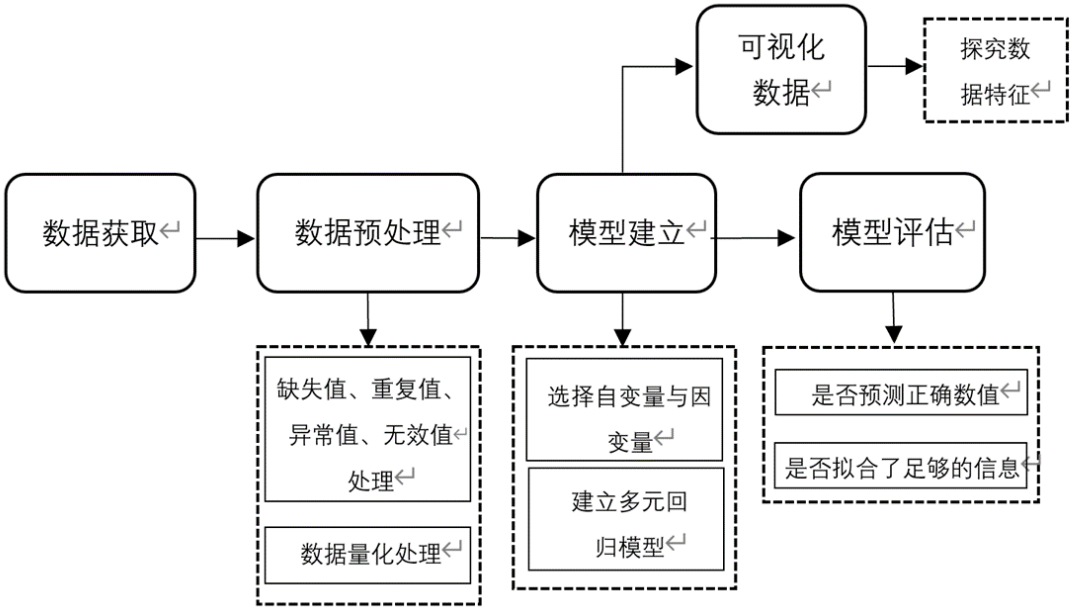
6、基于多元回归模型的房产估价

1、情景问题提出及分析

随着网络时代的来临，越来越多的用户选择在互联网上了解房源信息并选购房屋，如何利用这些房源信息尽可能帮助我们选房和对房产估价成了一个值得研究的问题。在二手房购买的选择过程中，房源的价格往往与位置、朝向、楼层和房屋面积等因素有关。本项目将利用这些信息首先对数据进行清洗，再通过建立多元回归模型的方式对房产进行估价。

本项目所提供的数据是截止到 2020 年 7 月 6 日的成都市二手房信息。

2、模型建立流程



7、基于决策树的电网负荷预测

1、情景问题提出及分析

电力系统的作用是对系统内各用户尽可能经济的提供可靠而合乎标准要求的电能。现代电网以系统运行的经济性为首要目标，再加之电能不能大量存储的特点，因此对电力系统的负载预测变得十分重要。

随着技术不断发展，当今越来越大的数据量配合着层出不穷的机器学习算法，已经大量运用在了电能预估当中，并且已经取得了一定的效果。

决定电力负荷的因素很多，比如有前期电力负荷、经济、社

会、气象等因素，本项目中将利用气象因素作为特征属性，通过建立决策树模型，完成对电网负载的预估。

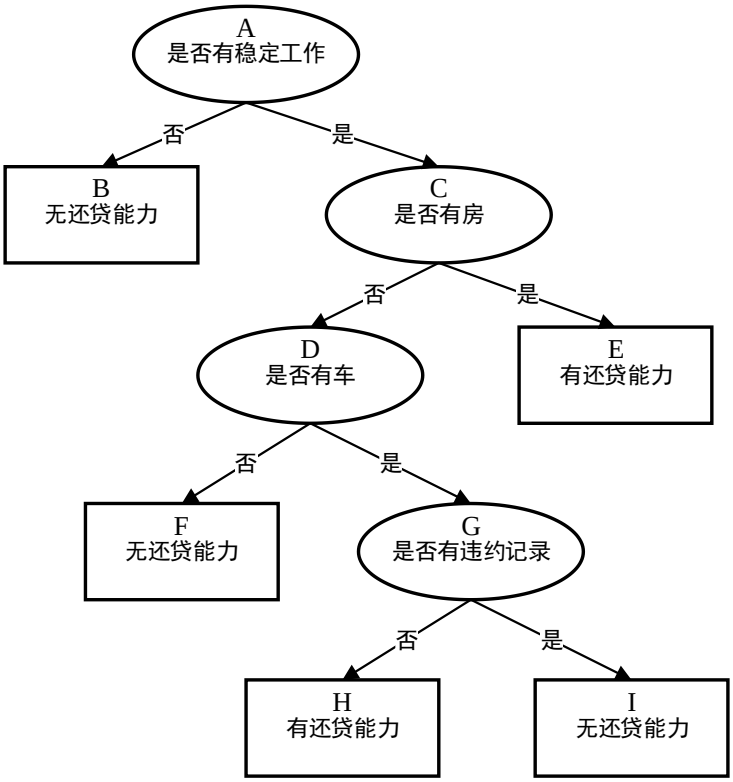
2、决策树算法简介

决策树(decision tree)是一种常见的机器学习方法。该方法属于监督学习的一种，它易于理解和实现，既可以用于分类，也可以用于回归。下面以决策树分类为例，对决策树算法做简要介绍。

例如，我们要判断一个人是否具有还贷能力，有以下几个特征：是否有稳定工作、是否有房、是否有车、是否有违约记录。决策过程包含着若干个子决策，下图展示了一个以作者主观思考所建立的决策树。

假设有以下测试集：

编号	是否有稳定工作	是否有房	是否有车	是否有违约记录
1	是	是	否	否
2	是	否	否	否



以编号 1，2 为例，具体说明决策树的决策过程。编号 1，首先进入根节点 A，判断是否有稳定工作？是，接着进入内部节点 C，判断是否有房？是，则进入叶子节点 E，决策结束，决策结果为有还款能力。编号 2，首先进入根节点 A，判断是否有稳定工作？是，进入内部节点 C，接着判断是否有房？否，进入内部节点 D，再判断是否有车？否，则进入叶子节点 F，决策结束，决策结果为无还款能力。

通俗来说决策树可简单理解为一个树和多个决策。它既可以用于分类也可以用于回归，通常决策树学习包括一下三个步骤：特征选择、决策树建树和决策树的修剪。

8、信用卡客户风险分析与评价

1、处理信用卡数据异常值

1.训练要点

(1)熟悉信用卡的基本业务知识。

(2)掌握异常值的识别与处理方法。

2.需求说明

为了推进信用卡业务良性发展，减少坏账风险，各大银行都进行了信用卡客户风险识别的相关工作，建立了相应的客户风险识别模型。某银行因旧的风险识别模型随时间推移不再适应业务发展需求，需要重新进行风险识别模型构建。目前，银行给出的信用卡信数据说明如表 7-11 所示。

3、实现思路

(1)读取信用卡数据。

(2)丢弃逾期、呆账、强制停卡、退票、拒往记录为 1、瑕疵户为 2 的记录。

(3)丢弃呆账、强制停卡、退票为 1、拒往记录为 2 的记录。

(4)丢弃频率为 5、刷卡金额不等于 1 的数据。

2、构造信用卡客户风险评价关键特征

1. 训练要点

(1)掌握信用卡模型的原理。

(2)构建信用卡用户风险分析关键特征。

2.需求说明

在信用卡相关的征信工作中，主要从 3 个方向判定客户的信用等级。信用等级分别为客户的历史信用风险，主要为客户的历史信用情况，包括了客户是否有逾期、呆账和强制停卡记录等；客户的现阶段经济状况，综合考虑了借款余额、个人月收入、个人月开销、家庭月收入，以及月刷卡额这类和个人经济水平息息相关的特征;客户的未来经济收入以及目前收入的稳定情况，客户的职业不同、年龄不同、房产信息不同，那么客户的经济稳定情况是不同的。

3.实现思路及步骤

(1)根据特征瑕疵户、逾期、 呆账，强制停卡记录 退票、拒往记录，构建历史行为特征。

(2)根据特征借款余额、个人月收入、个人月开销、家庭月收入和月刷卡额，构建出经济风险情况特征。

(3)根据特征职业、年龄、住家，构建出收入风险情况特征。

(4)标准化历史行为、经济风险情况、收人风险情况特征。

3、构建 K-Means 聚类模型

1.训练要点

(1)掌握 K-Means 聚类算法的应用。

(2)掌握聚类算法结果分析的方法。

2. 需求说明

构建信用卡高风险客户识别模型可以分为两部分:第一部分，根据构建的 3 个特征对客户进行分群，对客户做聚类分群；第二部分，结合业务对每个客户群进行特征分析，分析其风险，并对每个客户群进行排名。

3.实现思路及步骤

(1)构建 K-Means 聚类模型，聚类数为 5。

(2)训练 K-Means 聚类模型，并求出聚类中心、每类的用户数目。

9、企业所得税分析与预测

1、求取企业所得税各特征间的相关系数

1.训练要点

（1）掌握 Python 中的相关性分析方法，对每种分析方法进行比较。

（2）理解并会用 Python 实现企业所得税预测相关特征的相关性分析。

（3）对相关性分析结果进行解读。

2.需求说明

对影响企业所得税的原始特征进行相关性分析，对原始特征

间的相关性和原始特征与目标特征之间的相关性进行解读。

3.实现思路及步骤

（1）求取原始数据特征之间的 Pearson 相关系数。

（2）判断各特征之间的相关性。

2、选取企业所得税预测关键特征

1.训练要点

（1）理解 Lasso 回归模型，掌握其适用场景及优缺点。

（2）掌握使用 Lasso 回归进行特征选取的方法。

（3）理解并掌握上述过程的 Python 代码实现。

2.需求说明

对影响企业所得税的因素进行特征筛选，选取出对企业所得税有关键影响的特征，为下一步的模型构建奠定基础。

3.实现思路及步骤

（1）建立 Lasso 回归模型。

（2）对 Lasso 回归结果进行解读。

3、构建企业所得税预测模型

1.训练要点

(1)了解灰色预测模型，掌握其适用场景及优缺点。

(2)了解支持向量回归预测的由来，掌握其使用方法。

(3)理解并掌握回归预测模型评价指标。

(4)掌握此过程中所有的 Pyhon 代码实现。

2.需求说明

在对影响企业所得税的因素进行特征选取的基础上，建立单个特征的灰色预测模型和支持向向量回归预测模型，对 2014 年及 2015 年的企业所得税进行预测，并对模型进行评价。

3.实现思路及步骤

(1)使用灰色预测模型对各特征在 2014 年及 2015 年的值进行预测。

(2)建立支持向量回归预测模型。

(3)对上述建立的企业所得税预测模型进行评价。

10、电影数据分析

1、一元线性回归分析任务描述

在电影数据中，统计量日均票房=累计票房/放映天数。当日均票房不足百万元时一般将会在接下来的一周左右下档。我们可能会联想推测，日均票房与放映天数是否存在一定的相关性？在本项目中，我们将通过一元线性回归对两项数据进行简要的相关性分析，探讨是否可以通过计划放映天数预测电影的票房。

2、多项式回归分析任务描述

在电影数据中，统计量日均票房=累计票房/放映天数。当日均票房不足百万元时一般将会在接下来的一周左右下档。我们可能会联想推测，日均票房与放映天数是否存在一定的相关性？在本项目中，我们将通过多项式回归对数据进行简要的相关性分析，探讨是否可以通过计划放映天数预测电影的票房。此外，票房与多因素相关，使用多元线性回归分析多个影响因素。

3、任务实施

1.电影数据读取

2.数据清洗

3.模型建立

4.模型训练

5.数据预测与模型的可视化

11、基于聚类模型的观影评分数据分析

1、任务描述

本项目以观影评分数据为例说明不同聚类方法的使用。数据文件中存储了两列数据，分别表示用户对两部电影的评分。根据评分值的相似性，我们对观影用户进行分类，分成不同的客户群。实施不同的推荐内容。

2、任务实施

1. 电影评分数据读取
2. 数据清洗
3. 模型建立
4. 模型训练
5. 数据预测与模型的可视化

12、基于分类模型的身高与体重数据分析

1、任务描述

男性、女性的平均身高与体重不同，可否从身高、体重数据上找出与性别的关联。如果能够找出关联，那么我们就可以根据身高、体重数据来鉴别性别。在项目中，我们将使用逻辑回归、朴素贝叶斯、决策树、支持向量机等方法对数据进行简要的分析，探讨分类结果的准确性。

2、任务实施

1. 身高数据读取
2. 数据清洗
3. 模型建立
4. 模型训练
5. 数据预测与模型的可视化