# 第三阶段 从数据分析到数据挖掘

# SPSS Modeler-[数据分析与挖掘2]

## 网校相关视频

第二章 SPSS Modeler模型篇

## 明确本次课知识点，明确重点难点

### 【知识点目标】

* SPSS Modeler 数据挖掘--回归：多重线性回归/Logistic回归
* SPSS Modeler 数据挖掘--分类：决策树

## 本次课程任务讲解

### 【知识点1】SPSS Modeler 数据挖掘--回归

#### SPSS Modeler数据挖掘--多重线性回归

多重线性回归原理回顾

特点与前置条件

假设1：因变量是连续变量

假设2：自变量不少于2个（连续变量或分类变量都可以）

假设3：具有相互独立的观测值

假设4：自变量和因变量之间存在线性关系

假设5：等方差性

假设6：不存在多重共线性

假设7：不存在显著的异常值

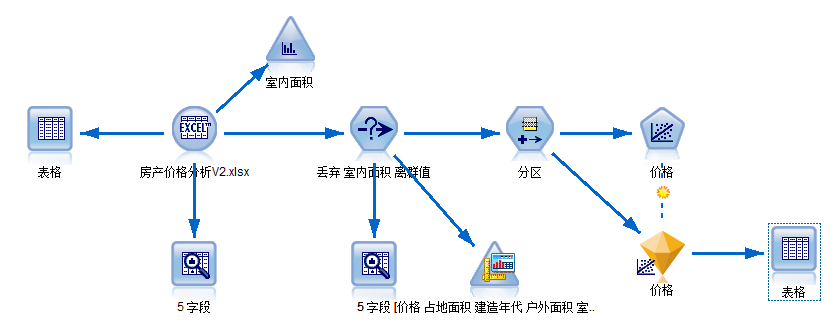
假设8：残差近似正态分布

SPSS Modeler中对建模的前置条件比较宽容，基本条件满足即可（假设1、6）

操作步骤

与一元线性回归类似，注意事先观察一下各变量之间的相关性

通过专家模式可以设置输出窗口的内容方便观察

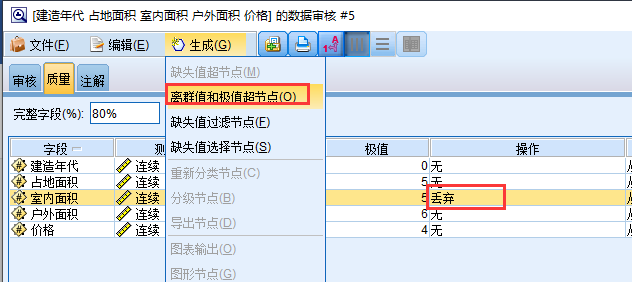
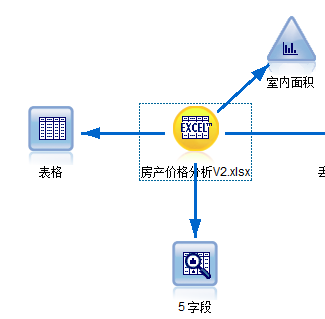


数据源：房产价格分析V2

1）、设置源节点

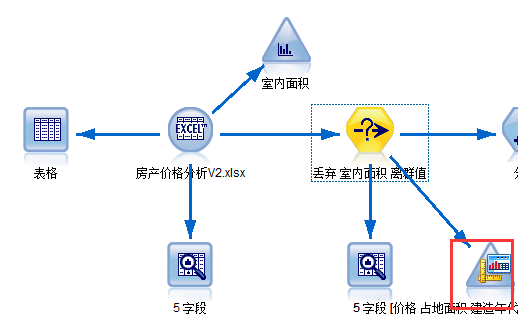


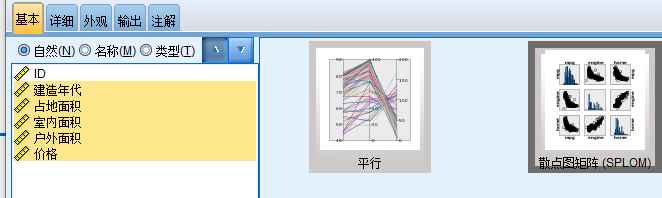
2）、观察室内面积的离群值



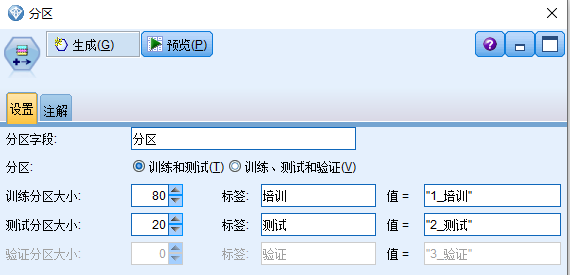


3）、通过矩阵散点图查看各变量的线性关系





4）添加分区

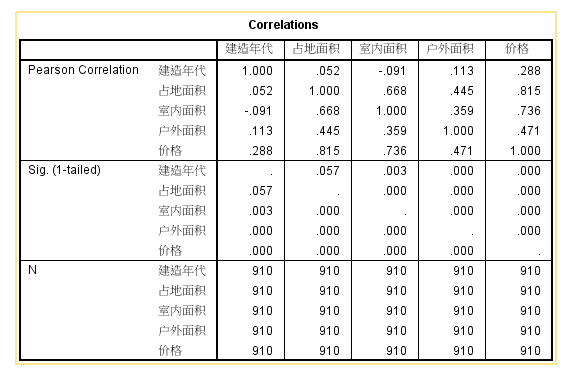


5）、建模选项卡——回归



结果解读

各变量间的相关性



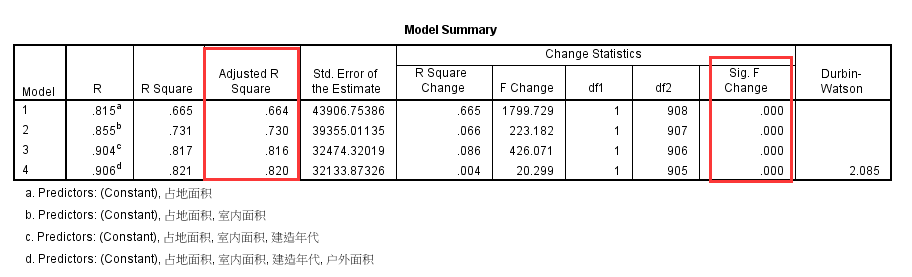
Pearson correlation：皮尔逊相关系数r，

若|r|≥0.8 时，视为高度相关；若0.5≤|r|＜0.8 时，视为中度相关；

当0.3≤|r|＜0.5时，视为低度相关；当|r|＜0.3时，可视为不相关

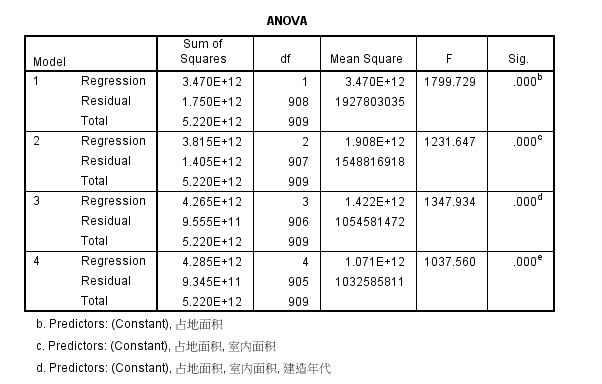
采用步进法如下：

模型汇总及拟合情况

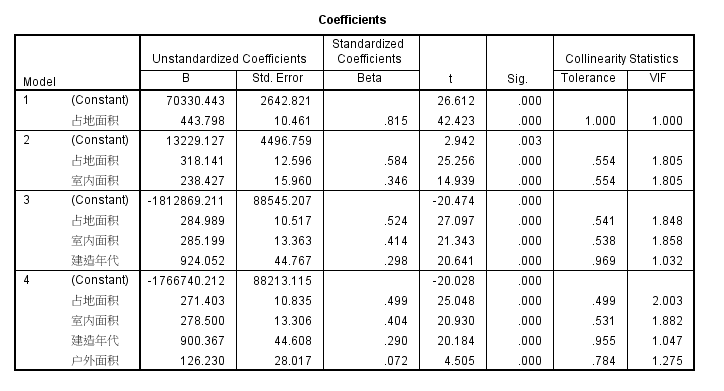


多元看调整r平方：adjusted r square

模型的显著性检验



系数及共线性检查



#### SPSS Modeler数据挖掘--二分类Logistic回归

Logistic回归原理

特点与前置条件

前置条件

假设1：因变量（结局）是二分类变量。

假设2：有至少1个自变量，自变量可以是连续变量，也可以是分类变量。

假设3：每条观测间相互独立。分类变量（包括因变量和自变量）的分类必须全面且每一个分类间互斥。

假设4：最小样本量要求为自变量数目的15倍，但一些研究者认为样本量应达到自变量数目的50倍。

假设5：连续的自变量与因变量的logit转换值之间存在线性关系。

假设6：自变量之间无多重共线性。

假设7：没有明显的离群点、杠杆点和强影响点

操作步骤

哑变量设置

把一个N分类的分类变量转换成N-1个二分类变量。SPSS Modeler会自动设置。

字段设置

**数据源：telco.sav**

**功能：建模选项卡——logistic节点**

字段的测量级别：

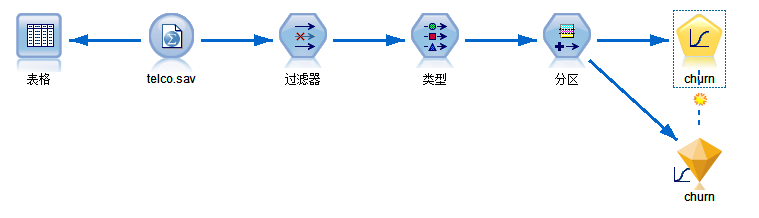
**连续型** – 用于描述数值，如0-100 或者0.75-1.25 内的连续值一个连续值可以是整数、实数或日期/时间

**分类（离散型）**– 用于当一个具体值的精确数量未知时描述字符串，一旦数据被读取，其类型就会转换成标记、名义或者无类型

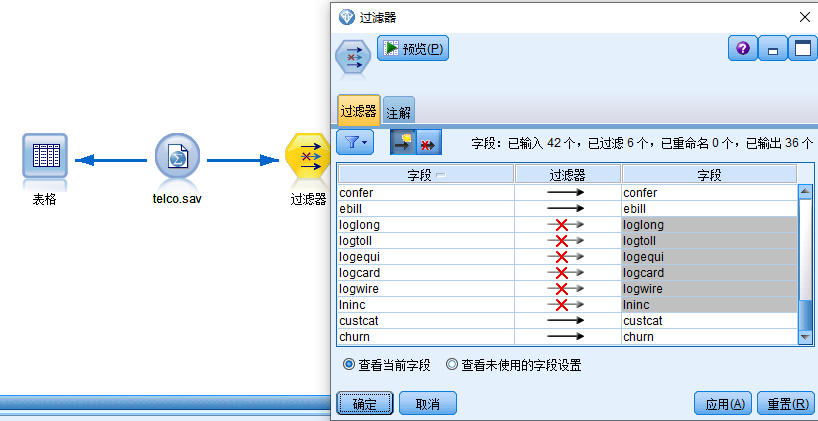
**名义** – 用于描述带有多个具体值的数据（黄、绿、蓝）

**标记** – 用于只取两个具体值的数据（真、假，1、0）

**无类型** – 用于不符合上述任一种类型的数据或者含有太多元素的集合类型数据

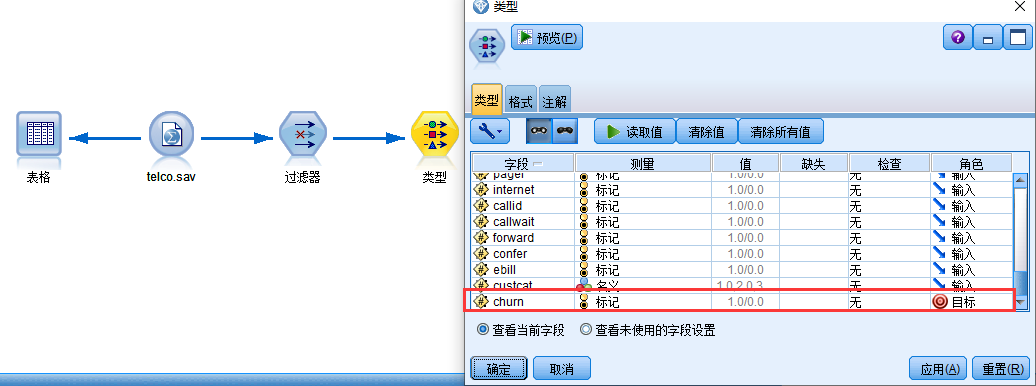


* + - * 1. 过滤掉不需要的字段



2）更改测量尺度：churn(上个月的客户流失)：名义——标记

gender：名义——标记



3）、添加分区字段



4）添加建模节点

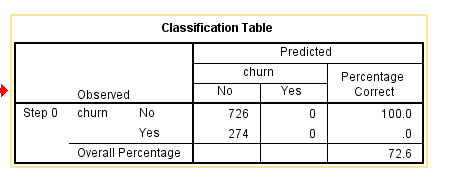




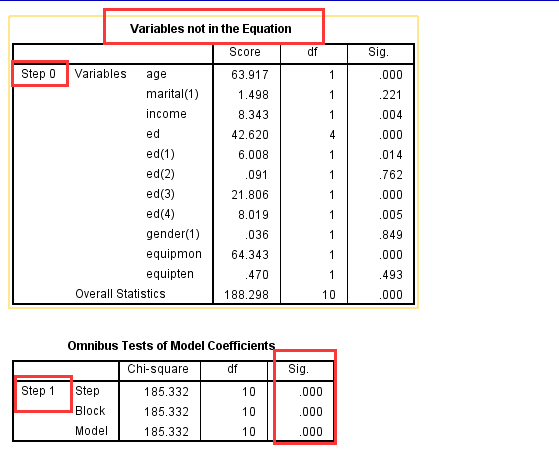
哑变量（分类变量）设置，注意基准为第一个，意味着计算OR比时以第一个类别为基准进行比较。

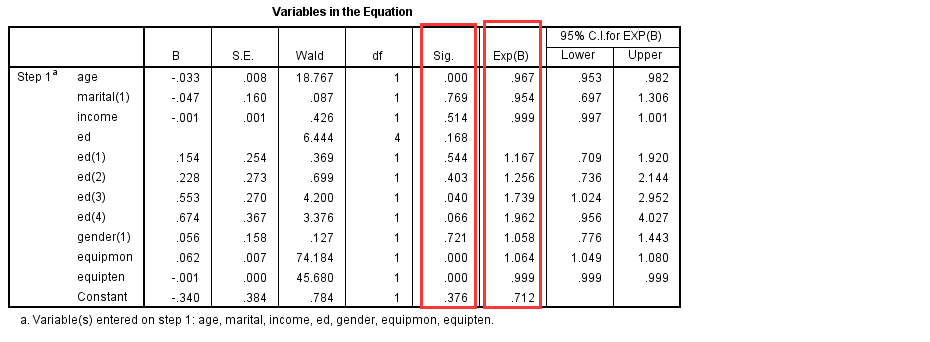
结果解读

原始数据因变量的混淆矩阵



加入自变量后的回归方程是否显著





年龄增大一岁，客户流失的可能性会降低3.3%

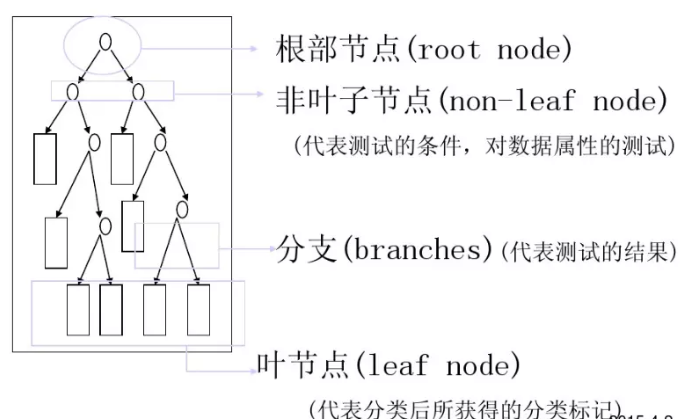
### 【知识点2】SPSS Modeler 数据挖掘--分类

#### 决策树算法简介

基本思路：通过设定条件（自变量），对数据（因变量）进行分组

分类决策树模型是一种描述对实例进行分类的树形结构。决策树由结点和有向边组成。结点有两种类型：内部结点和叶节点。内部节点表示一个特征或属性，叶节点表示一个类。  
决策树（决策树），又称为判定树，是一种以树结构（包括二叉树和多叉树）形式表达的预测分析模型。

* 通过把实例从根节点排列到某个叶子节点来分类实例
* 叶子节点为实例所属的分类
* 树上每个节点说明了对实例的某个属性的测试，节点的每个后继分支对应于该属性的一个可能值



1. 各种不同决策树的差异

|  |  |
| --- | --- |
| IMG_256 | 分类和回归 (C&R) 树节点生成可用于预测或分类未来观测值的决策树。该方法通过在每个步骤最大限度降低不纯洁度，使用递归分区来将训练记录分割为组。如果树中某个节点中 100% 的观测值都属于目标字段的一个特定类别，那么该节点将被认定为“纯洁”。**目标和输入字段可以是数字范围或分类（名义、有序或标志）；所有分割均为二元分割（即仅分割为两个子组）。** |

|  |  |
| --- | --- |
| IMG_257 | **CHAID 使用卡方统计量来生成决策树**，以确定最佳的分割。CHAID 与 C&R 树和 QUEST 节点不同，它**可以生成非二元树**，这意味着有些分割将有多于两个的分支。**目标和输入字段可以是数字范围（连续）或分类。**Exhaustive CHAID 是 CHAID 的修正版，它对所有分割进行更彻底的检查，但计算时间比较长。 |

|  |  |
| --- | --- |
| IMG_258 | QUEST 节点可提供用于构建决策树的二元分类法，此方法的设计目的是减少大型 C&R 树分析所需的处理时间，同时也减少在分类树方法中发现的趋势以便支持允许有多个分割的输入。**输入字段可以是数字范围（连续），但目标字段必须是分类。所有分割都是二元的。** |

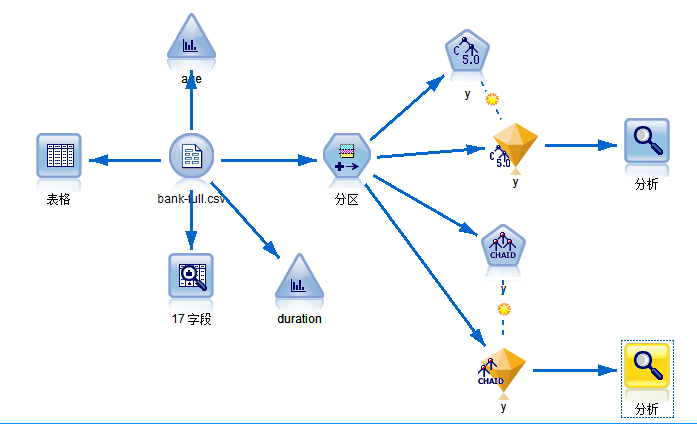
|  |  |
| --- | --- |
| IMG_259 | C5.0 节点构建决策树或规则 集。该模型的工作原理是根据在每个级别提供最大信息收获的字段分割样本。**目标字段必须为分类字段。允许进行多次多于两个子组的分割。** |

|  |  |
| --- | --- |
| IMG_260 | 树 AS 节点类似于现有的 CHAID 节点；但是，“树 AS”节点旨在处理大量数据以创建单个树，并在 SPSS® Modeler V17 中添加的输出查看器中显示生成的模型。此节点通过使用卡方统计 (CHAID) 来识别最优拆分，从而生成决策树。 对 CHAID 的这一使用可生成非二元树，意味着某些拆分将具有两个以上的分支。 **目标和输入字段可以是数字范围（连续）或分类。**Exhaustive CHAID 是 CHAID 的修正版，它对所有分割进行更彻底的检查，但计算时间比较长。 |

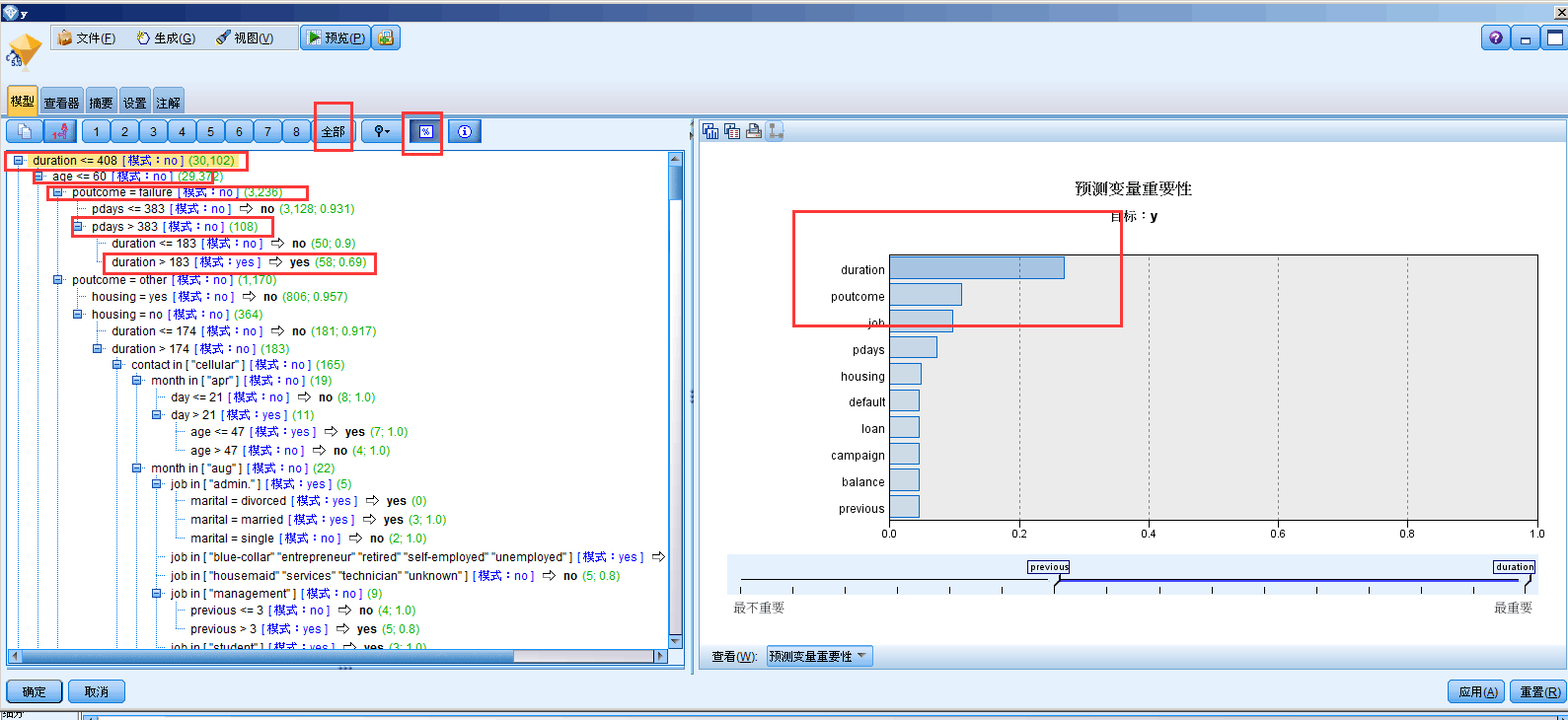
|  |  |
| --- | --- |
| IMG_261 | “随机树”节点类似于现有 C&RT 节点；但是，“随机树”节点旨在处理大数据以创建单个树，并在 SPSS Modeler V17 中添加的输出查看器中显示生成的模型。“随机树”节点将生成您可以对未来观测值进行预测或分类的决策树。通过在每个步骤最大限度降低不纯洁度，此方法使用递归分区将训练记录分割为多个段。如果树中某个节点的全部个案都属于目标字段的一个特定类别，那么系统会将该节点视为纯洁。**目标和输入字段可以是数字范围或分类（名义、有序或标志）；所有分割均为二元分割（即仅分割为两个子组）。** |

1. 实例

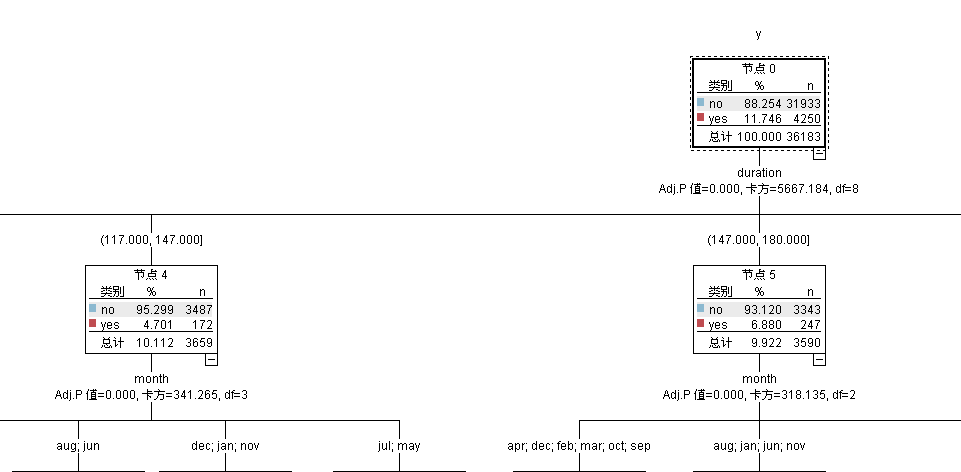
使用C5.0处理银行客户信息，数据流如下，y判断客户是否购买该产品



模型的解读（注意规则要右前至后，由上至下把一条规则全部说明完整）



部分决策树



通过分析节点可以判定针对Yes和No的客户的分别不同的准确率

## 自主学习作业讲解

无

## 课程总结

* SPSS Modeler 数据挖掘--回归
* 多重线性回归
* 原理
* 操作步骤
* 结果解读
* Logistic回归
* 原理
* 操作步骤
  + 注意哑变量设置
* 结果解读
* SPSS Modeler 数据挖掘--分类
* 决策树
* 基本思路
* 生长与剪枝（前后）
* 各种不同决策树算法的差异
* 实际操作
* 结果解读

## 下次自主学习任务布置

#### 观看预习视频

第2章 模型篇（关联、聚类）

#### 课后作业

见作业文档