# 第一阶段 业务分析师

# 建模数据分析师

## 网校相关视频

SPSS MODELER数据挖掘

## 明确本次课知识点，明确重点难点

### 【知识点目标】

* 主要算法介绍
* 自动建模器
* 分类模型
* 聚类模型
* 关联模型
* SPSS MODELER的输出和导出操作
* SPSS MODELER建模操作
* 建模操作
* 数据理解的重要性
* 从商业理解到数据理解
* 缺失值定义
* 缺失值分析
* 缺失值处理——填充
* 异常值分析——单字段分析
* 异常值处理
* 观察数据项之间联系

### 【复习巩固作业讲解

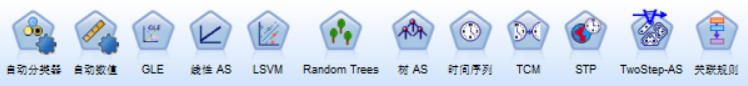
无

## 本次课程任务讲解

### 【知识点1】主要算法介绍

IBM SPSS Modeler以图形化的界面、简单的拖拽方式来快速构建数据挖掘分析模型著称，它提供了完整的统计挖掘功能，包括来自于统计学、机器学习、人工智能等方面的算法和数据模型，包括如关联、分类、预测等完整的数据挖掘分析功能。

1. 大数据类模型



1. 分类模型



1. 关联型模型



1. 细分类模型



### 【知识点2】自动建模器

首先，针对刚入门数据挖掘领域的初学者来说，即使你不懂数据挖掘算法，一样可以使用SPSS Modeler进行建模，这里提供了自动建模器，可以帮助你自动选择最优算法，包括有：

**自动分类器节点**：用于创建和对比二元结果（是或否，流失或不流失等）的若干不同模型，使用户可以选择给定分析的最佳处理方法。由于支持多种建模算法，因此可以对用户希望使用的方法、每种方法的特定选项以及对比结果的标准进行选择。节点根据指定的选项生成一组模型并根据用户指定的标准排列最佳候选项的顺序。应用于分类目标。

**自动数值节点：**使用多种不同方法估计和对比模型的连续数字范围结果。此节点和自动分类器节点的工作方式相同，因此可以选择要使用和要在单个建模传递中使用多个选项组合进行测试的算法。受支持的算法包括神经网络、C&R树、CHAID、线性回归、广义线性回归以及支持向量机(SVM)。可基于相关度、相对错误或已用变量数对模型进行对比。应用于数值目标。

**自动聚类节点：**估算和比较识别具有类似特征记录组的聚类模型。节点工作方式与其他自动建模节点相同，在一次建模运行中即可试验多个选项组合。模型可使用基本测量进行比较，以尝试过滤聚类模型的有效性以及对其进行排序，并提供一个基于特定字段的重要性的测量。

**时间序列节点：**可为时间序列估计指数平滑模型、单变量综合自回归移动平均(ARIMA) 模型和多变量ARIMA（或变换函数）模型并基于时间序列数据生成预测。

### 【知识点3】分类模型

分类是数据挖掘里面最常用的分析方法，它的计算逻辑是自我学习的过程，即有监督的学习，需要给他一个目标和一些影响因素，它会自动找到影响因素与目标之间隐藏的规则。

**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。

**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，使用递归分区将训练记录分割为组。如果节点中100% 的观测值都属于目标字段的一个特定类别，则树中的该节点将被认定为“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）。

**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。

**CHAID节点**：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。

**决策列表节点**：可标识子组或段，显示与总体相关的给定二元结果的似然度的高低。例如，你或许在寻找那些最不可能流失的客户或最有可能对某个商业活动作出积极响应的客户。通过定制段和并排预览备选模型来比较结果，可以将自己的业务知识体现在模型中。决策列表模型由一组规则构成，其中每个规则具备一个条件和一个结果。规则依顺序应用，相匹配的第一个规则将决定结果。

**线性模型节点**：根据目标与一个或多个预测变量间的线性关系来预测连续目标。

**线性回归节点**：是一种通过拟合直线或平面以实现汇总数据和预测的普通统计方法，它可使预测值和实际输出值之间的差异最小化。

**因子/主成分分析节点**：提供了用于降低数据复杂程度的强大数据缩减技术。主成份分析（PCA）可找出输入字段的线性组合，该组合最好地捕获了整个字段集合中的方差，且组合中的各个成分相互正交（相互垂直）。因子分析则尝试识别底层因素，这些因素说明了观测的字段集合内的相关模式。这两种方式的目标都是找到有效概括原始字段集中的信息的一小部分导出字段。

**神经网络节点**：使用的模型是对人类大脑处理信息的方式简化了的模型。此模型通过模拟大量类似于神经元的抽象形式的互连简单处理单元而运行。神经网络是功能强大的一般函数估计器，只需要最少的统计或数学知识就可以对其进行训练或应用。

**特征选择节点**：会根据某组条件（例如缺失值百分比）筛选可删除的输入字段；对于保留的输入，将相对于指定目标对其重要性进行排序。

**Logistic 回归节点**：是一种统计方法，它可根据输入字段的值对记录进行分类。它类似于线性回归，但采用的是类别目标字段而非数字范围。

**判别式分析节点**：所做的假设比logistic回归的假设更严格，但在符合这些假设时，判别式分析可以作为logistic回归分析的有用替代项或补充。

**“广义线性”模型节点**：对一般线性模型进行了扩展，这样因变量通过指定的关联函数与因子和协变量线性相关。另外，该模型允许因变量呈非正态分布。它包括统计模型大部分的功能，其中包括线性回归、logistic回归、用于计数数据的对数线性模型。

**Cox回归节点**：可为时间事件数据构建预测模型。该模型会生成一个生存函数，该函数可预测在给定时间t 内对于所给定的预测变量值相关事件的发生概率。

**支持向量机(SVM)节点**：使用该节点，可以将数据分为两组，而无需过度拟合。SVM 可以与大量数据集配合使用，如那些含有大量输入字段的数据集。

**贝叶斯网络节点**：可以利用该节点对真实世界认知的判断力并结合所观察和记录的证据来构建概率模型。该节点重点应用了树扩展简单贝叶斯(TAN) 和马尔可夫毯网络，这些算法主要用于分类问题。

**自学响应模型(SLRM) 节点**：利用该节点可以构建这样的模型：随着数据集的增长，可以不断对其进行更新或重新估计，而不必每次使用整个数据集重新构建该模型。例如，如果有若干产品，而希望确定某位客户获得报价后最有可能购买的产品，那么这种模型将十分有用。此模型可用于预测最适合客户的报价，以及该报价被接受的概率。

**KNN (k-最近相邻元素)节点**：将新的个案关联到预测变量空间中与其最邻近的k个对象的类别或值（其中k为整数）。类似个案相互靠近，而不同个案相互远离。

### 【知识点4】聚类模型

聚类算法我们也称为无监督学习，是与分类算法的有监督学习相对而言。一般在数据初探的时候，经常会用到，比如客户细分、市场细分等。

**K-Means节点**：将数据集聚类到不同分组（或聚类）。此方法将定义固定的聚类数量，将记录迭代分配给聚类，以及调整聚类中心，直到进一步优化无法再改进模型。k-means 节点作为一种非监督学习机制，它并不试图预测结果，而是揭示隐含在输入字段集中的模式。

**Kohonen节点**：会生成一种神经网络，此神经网络可用于将数据集聚类到各个差异组。此网络训练完成后，相似的记录应在输出映射中紧密地聚集，有差异的记录则应彼此远离。可以通过查看模型块中每个单元所捕获观测值的数量来找出规模较大的单元。这将让对聚类的相应数量有所估计。

**TwoStep节点**：使用两步聚类方法。第一步完成简单数据处理，以便将原始输入数据压缩为可管理的子聚类集合。第二步使用层级聚类方法将子聚类一步一步合并为更大的聚类。TwoStep 具有一个优点，就是能够为训练数据自动估计最佳聚类数。它可以高效处理混合的字段类型和大型的数据集。

### 【知识点5】关联模型

关联分析在零售行业应用最为广泛，最传统的就是购物篮分析，例如淘宝上的“猜你喜欢”。

**先验节点**：从数据抽取一组规则，即抽取信息内容最多的规则。“先验”节点提供五种选择规则的方法并使用复杂的索引模式来高效地处理大数据集。对于大问题而言，“先验”通常用于训练时比GRI 处理速度快；它对可保留的规则数量没有任何限制，而且可处理最多带有32 个前提条件的规则。“先验”要求输入和输出字段均为分类型字段，但因为它专为处理此类型数据而进行优化，因而处理速度快得多。

**CARMA节点**：使用关联规则发现算法来发现数据中的关联规则。例如，可以使用此节点生成的规则来查找一系列产品或服务（条件），其结果是要在此假期内进行促销的项目。

**序列节点**：可发现连续数据或与时间有关的数据中的关联规则。序列是一系列可能会以可预测顺序发生的项目集合。例如，一个购买了剃刀和须后水的顾客可能在下次购物时购买剃须膏。序列节点基于CARMA 关联规则算法，该算法使用有效的两步法来发现序列。

### 【知识点6】SPSS MODELER的输出和导出操作

**“输出”选项卡**

SPSS Modeler的输出不仅仅是ETL过程，还包括了对数据的统计分析报告输出，如表、矩阵、分析、数据审核、统计量等。输出节点提供了用于获取数据和模型的相关信息的方法。还提供了以各种格式导出数据以与其他软件工具相互作用的机制。



“表”节点以表格式显示数据，这些数据还可以写入到文件中。

“矩阵”节点将创建一个字段关系表。此节点最常用于显示两个符号字段间的关系。

“分析”节点评估预测模型生成准确预测的能力。

“数据审核”节点将首先全面检查数据，这些数据包括每个字段的汇总统计量、直方图和分布以及有关离群值、缺失值和极值的信息。数据：历史数据.xlsx

“统计量”节点可提供有关数值字段的基本汇总信息。它可计算单个字段以及字段间的相关性的汇总统计量。

**“导出”选项卡**

SPSS Modeler导出的格式与“源”选项卡类似，包含数据库、Excel、SAS导出、Statistics导出等，用来对处理后的结果输出成相应格式。



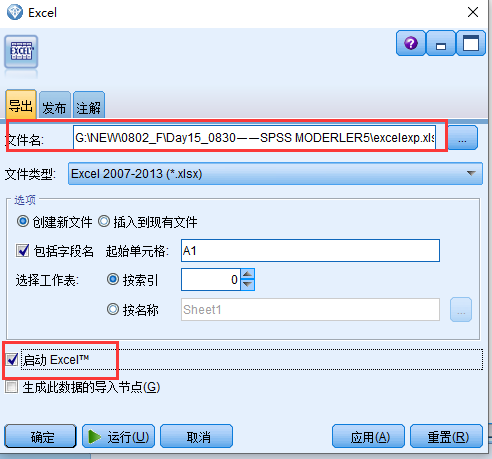
“数据库导出”节点：将数据写入与 ODBC 兼容的关系数据源。要将数据写入 ODBC 数据源，数据源必须存在且必须对其具有写许可权。

平面文件导出节点：将数据输出到已分隔的文本文件。这对导出可由其他分析或电子表格软件读取的数据非常有用。

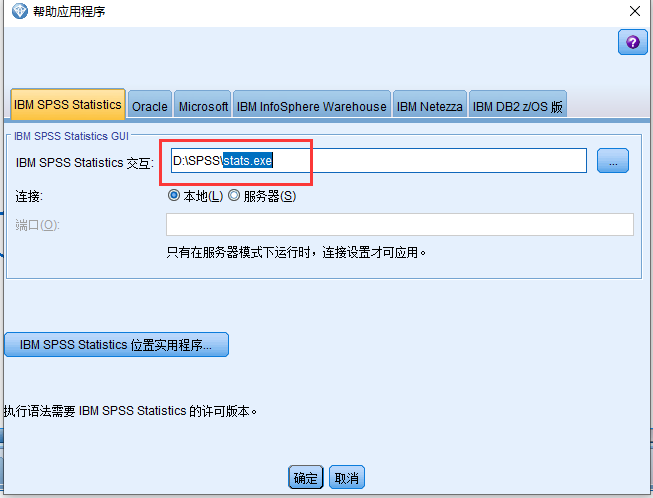
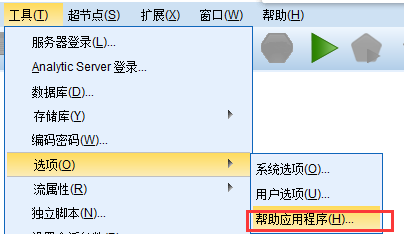
Statistics 导出节点：以 SPSS .sav 或 .zsav 格式输出数据。

“SAS 导出”节点：可以 SAS 格式输出数据，以便读入 SAS 或与 SAS 兼容的软件包中。有三种可用的 SAS 文件格式：SAS for Windows/OS2、SAS for UNIX 或 SAS Version 7/8。

Excel 导出节点：以 Microsoft Excel .xlsx 文件格式输出数据。 还可选择在执行完此节点后自动启动 Excel 并打开导出的文件。



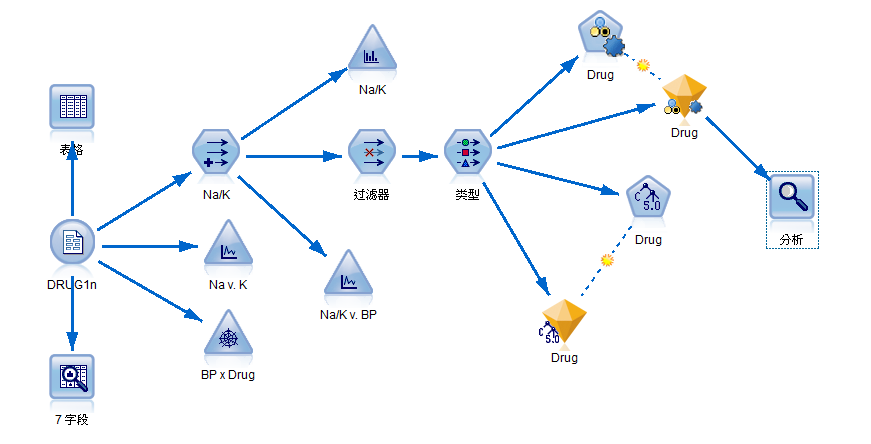
打开SPSS文件的方法：

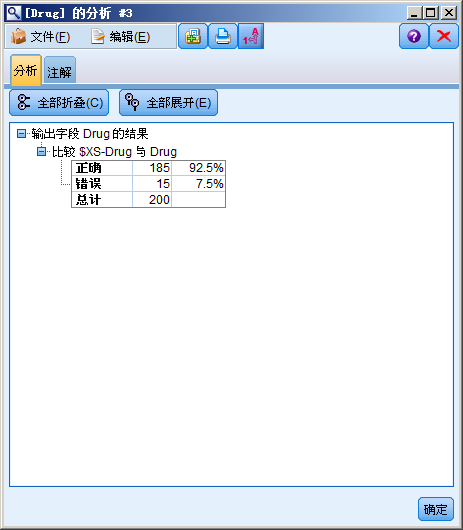


### 【知识点7】SPSS modeler建模操作

一般步骤：

数据挖掘的建模操作并非直接将建模节点放入数据流中执行即可，而是需要根据项目的实际情况，了解具体背景和需求，做好业务理解和数据理解，在此基础上完成对数据的准备之后，再进入到建模的操作环节。以下通过一个药物试验的范例，来对建模操作的步骤加以说明。





### 【知识点8】建模操作

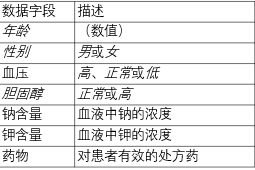
1. 背景介绍：

XX病是一种常见疾病，现有5种药物可以对其进行治疗，分别是A、B、C、X、Y。不同的药物对病人有不同的疗效。历史上，医院往往根据医生的经验去判断对特定的病人应该如何选择药物，但由于老医生的离开和新医生的加入，这种仅靠经验判断的做法会因为不同的医生而有不同的判断，造成了很多误诊。

医院有比较完善的病历留存，为了改变上述局面，决定使用数据挖掘技术对历史数据进行分析研究，希望能够建立一套有效的药物选择决策支持系统。

2、数据说明

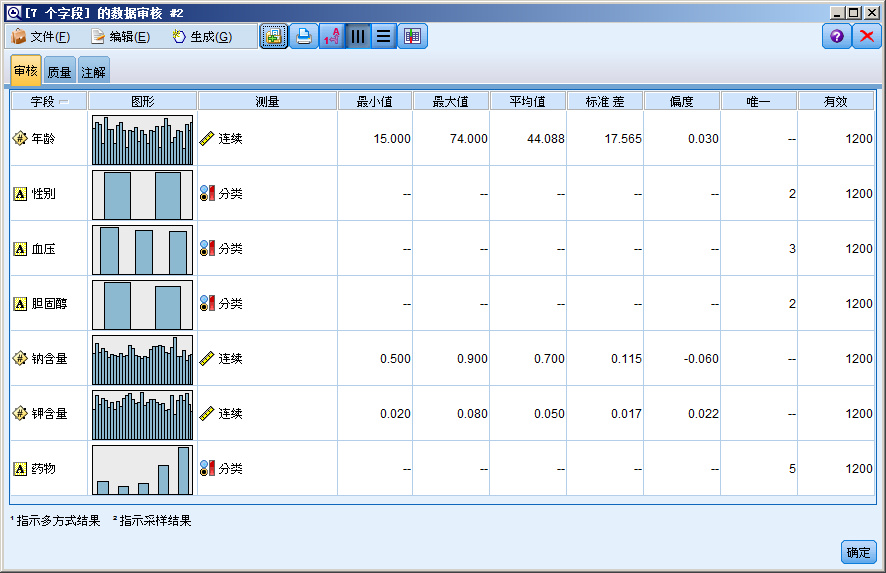
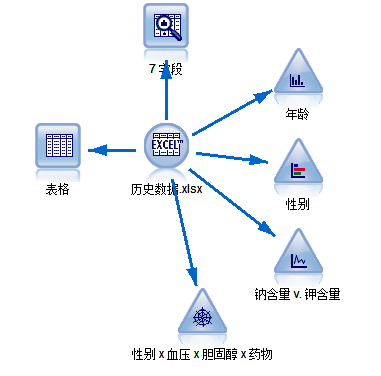
从医院的病历中整理出1200条相关数据，根据和医生的交流，最终从该数据中选择了以下7个字段记入数据库。



3、业务理解和数据理解

业务理解：此业务是要根据病人的个人情况和身体生化指标来确定何种药物对他最合适。

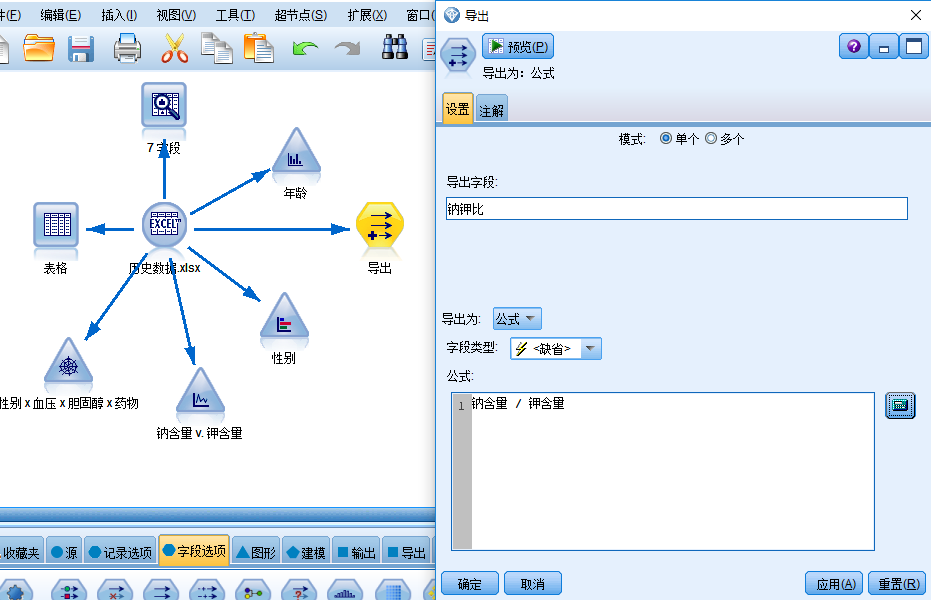
数据理解：通过对数据质量、数据之间的基本关系进行探索性分析，了解其基本情况，并初步观察病人的情况和身体特征是否与所选药物关系明显。可以通过建立各种图形来帮助完成这一过程。此外，通过数据审核节点，也可以快速就各字段生成相关评估图形供初步分析研判。



4、数据准备

数据准备主要完成对不同数据源的整合，并对数据进行适当变换，使之适应数据挖掘的需要。对于预测性的分类模型，通常还需要把原始数据集拆分成训练数据集和检验数据集。

1）导出新的字段测试



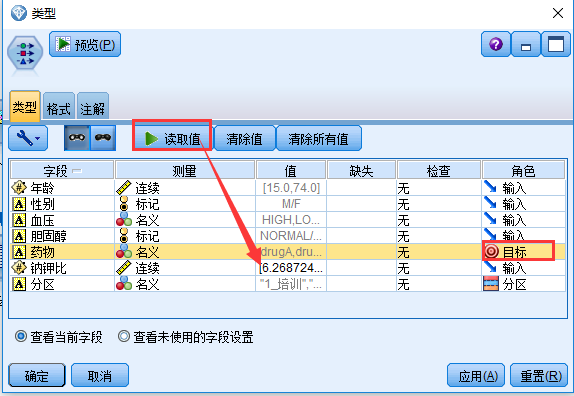
2）通过分区节点设定训练和测试分区



3）过滤掉干扰字段



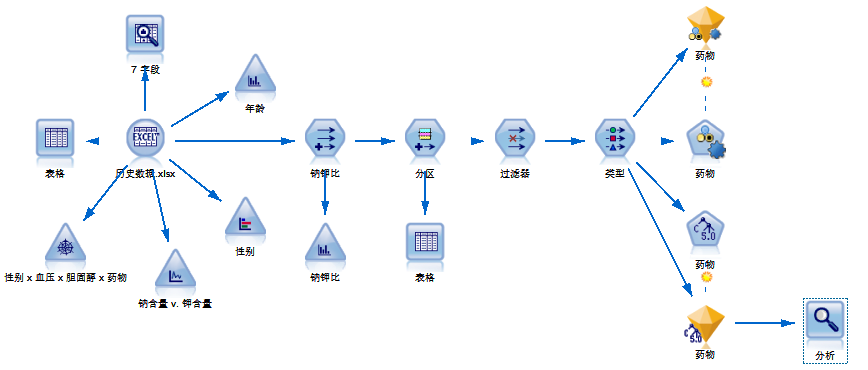
4）设置字段，将药物设为目标



5、建模

通过过滤节点和类型节点对数据源进行设置，而后可以选择相关的模型进行建模的工作了。

对药物这个分类变量进行预测，在建模菜单下可使用自动分类器得到通过算法计算的模型。

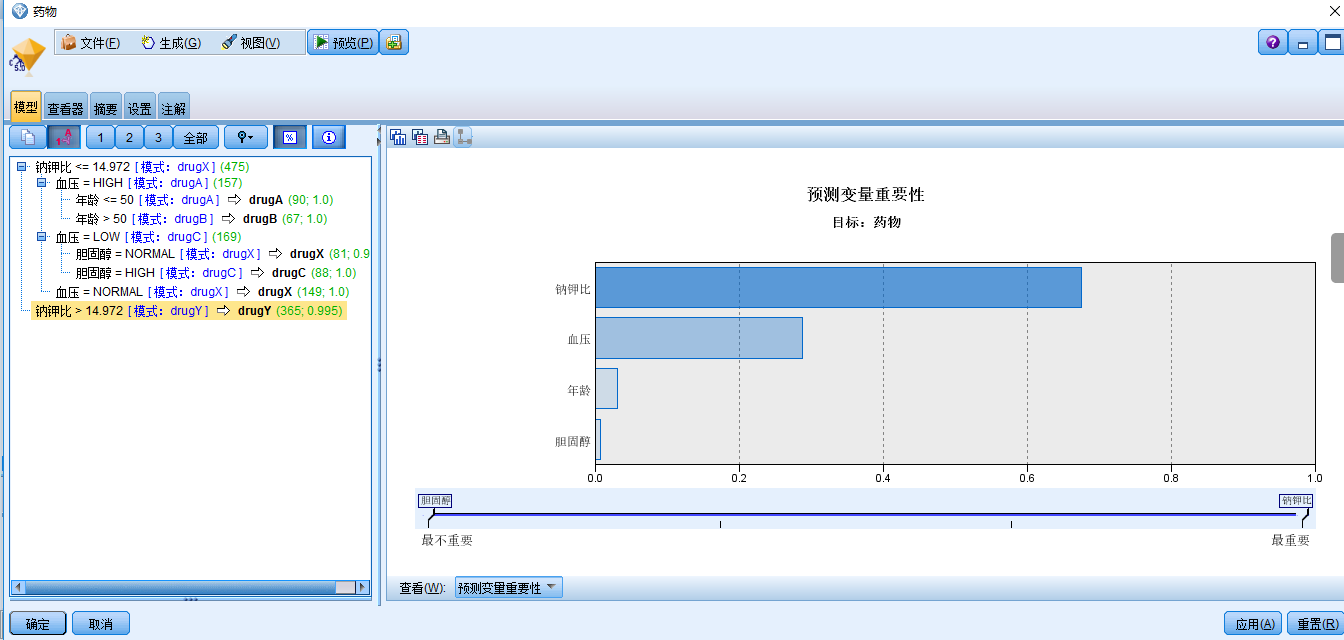


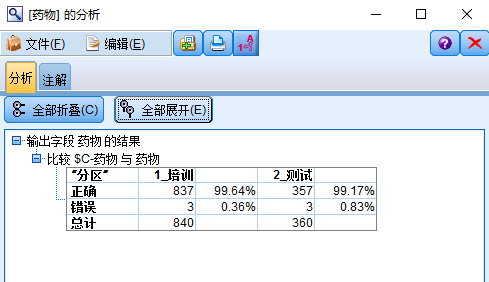
6、评估

分析模型运行结果，并可通过检验数据集再次验证模型的效果

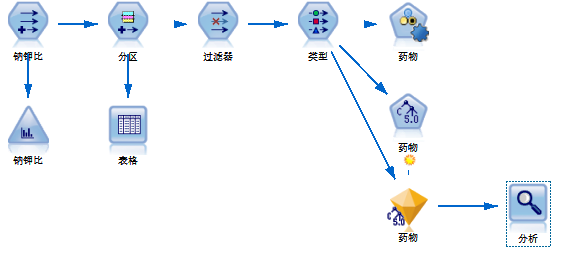
分析模型运行结果

当钠钾比大于14.972的时候，推荐使用Y药物，并且有356条数据支持，准确率达到99.5%





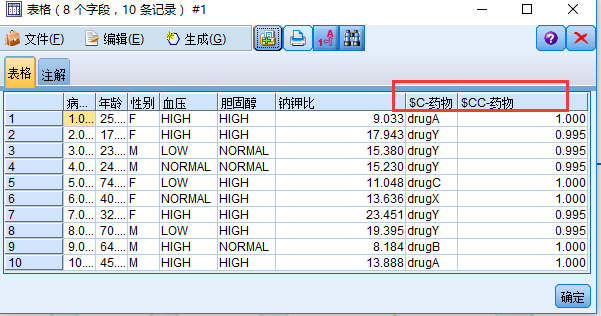
连接到验证数据集验证



7、部署

针对新病历数据（新数据.xlsx），利用预测模型获得推荐药物，并通过表或数据库方式发布。后续可以视情况开发相关推荐系统，以做到实时药物推荐。





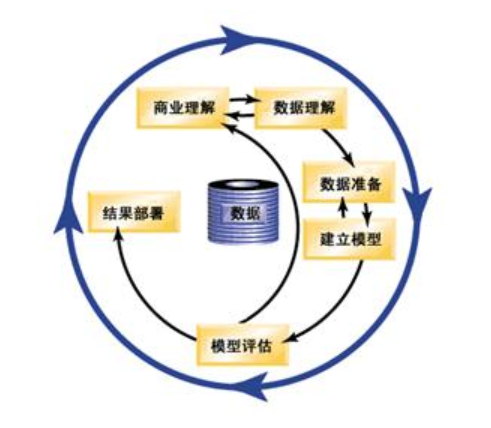
新增的两个字段

$C-药物：推荐使用的药物

$CC-药物：置信度/准确率

### 【知识点9】数据理解的重要性

在数据挖掘项目中，数据理解常常不被重视。但其实数据理解在整个数据挖掘项目中扮演着非常重要的角色，可以说是整个项目的基石。在计算机领域有一句话，“Garbage in，garbage out.” 意思就是说，如果项目的输入数据没有经过科学的预处理，那所得到的结果必将是错误的。通过数据理解，可以理解数据的特性和不足，进而对数据进行预处理，使得将来得到的模型更加稳定和精确。其次通过理解数据项之间的关系，可以为建模时输入数据项和模型的选择提供重要的信息。



### 【知识点10】从商业理解到数据理解

以某超市的市场推广活动为例，从商业理解开始，了解如何用Modeler 进行数据理解。

**商业理解**

现状：

某超市新增加了体育服饰用品营业部。开业一段时间，由于体育服饰用品地处二楼，很多顾客还不知道，营业额没有达到预期。

目标：

经理决定进行一次促销活动，具体活动是向会员中的部分用户邮寄打折优惠卡。考虑到优惠卡制作费用，邮寄费用，经理希望能够向那些最有购买潜力的客户邮寄优惠卡。使这些潜在用户了解本超市的体育品牌和刺激他们进行消费。

活动计划：

首先调取自体育用品部营业来的所有销售记录，得到购买体育用品的会员记录，建立模型，对本超市所有会员进行预测，对那些最有可能购买体育用品且尚未购买的客户邮寄优惠卡。以刺激这些潜在客户的消费。

**数据理解**

首先对超市内现有的数据进行分析：

1. 会员基本信息：会员申请会员卡时登记的信息，其中包含了会员年龄，职业，学历，电话，工作，收入，住址等信息。

2. 消费信息：会员消费的明细记录。

了解了现有数据后，发现有以下问题：

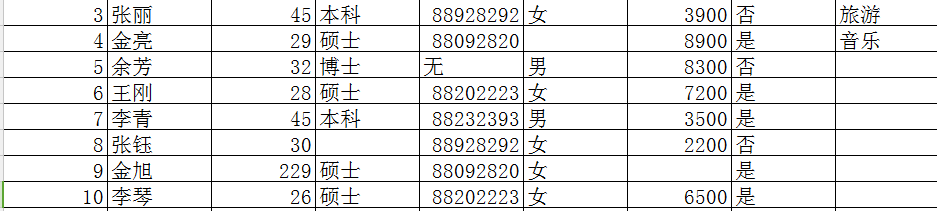
1. 会员基本信息是会员提供的，里边有很多值是缺失的，有的看起来是错误的。

2. 而消费信息由于是每次消费后电脑生成，信息是完整的，但是信息却很庞杂，不利于分析。

3. 可用的数据项非常多，到底用哪些数据项来进行数据预测呢？

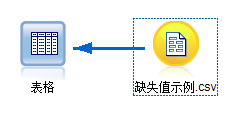
### 【知识点11】缺失值定义

缺失值就是指数据文件中的某些数据项是未知的值。几乎所有的商业数据挖掘中，都要遇到缺失值的问题，有可能是数据采集中的失误，有可能客户不愿意提供某些信息，面对这样的数据，应该如何处理? 使用这样的数据可能会对后期的建模产生不可预料的影响。如果丢弃，这些数据中可能包含着宝贵的信息。

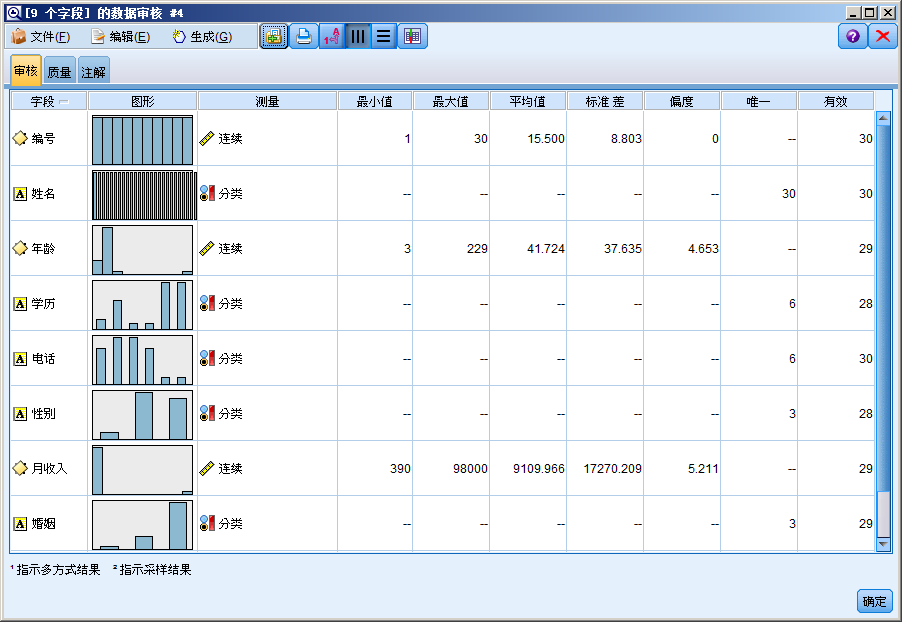


### 【知识点12】缺失值分析

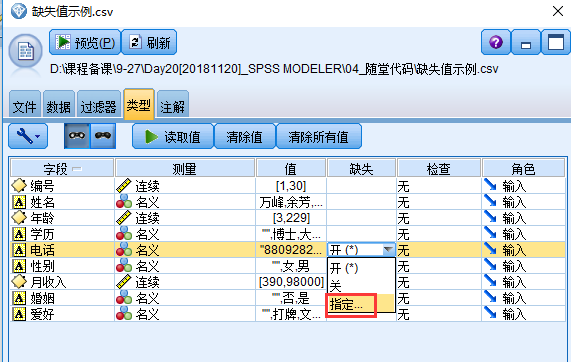
1、确定数据文件中缺失值的类型和数量

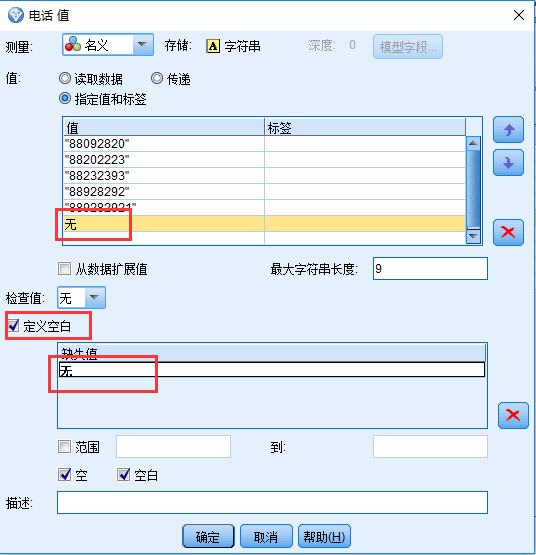


2、利用“数据审核”节点审核数据，观察有效数据的数量，不足30的都是有缺失值



3、电话里有“无”是有效字符串，需要定义缺失值，利用类型节点设置缺失值定义





4、通过数据审核节点进行缺失值处理设定

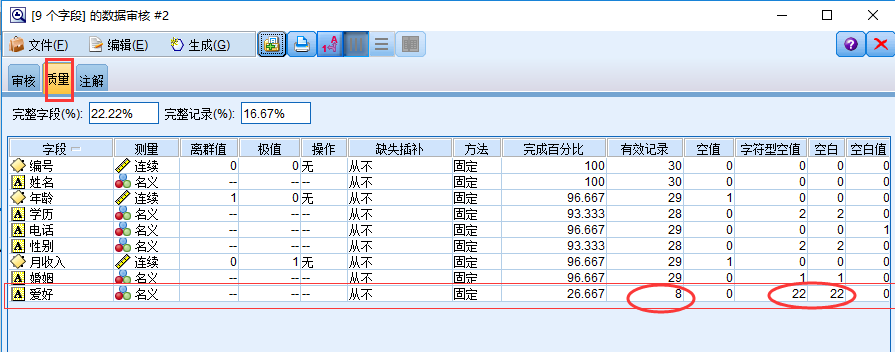
原则：

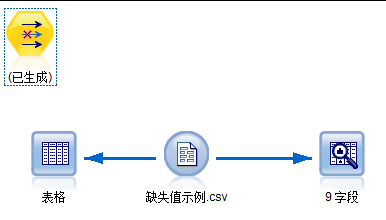
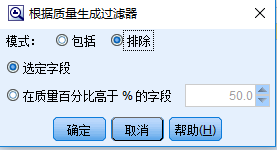
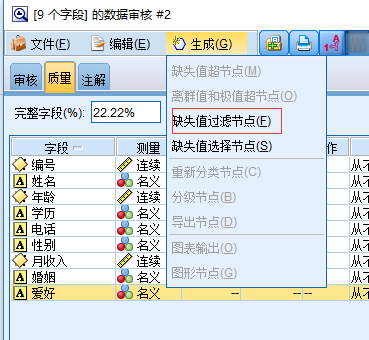
1）、完整的字段占的比例高，

一般应该过滤掉包含缺失值的字段

2）、 完整记录所占比例较高

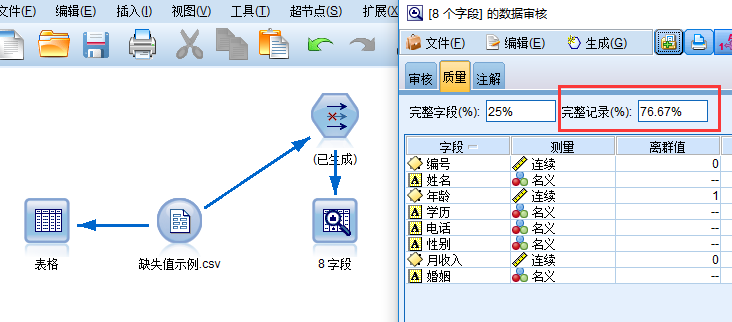
一般应该删除含有缺失值的记录





5、通过数据审核节点生成的处理节点删除字段或记录

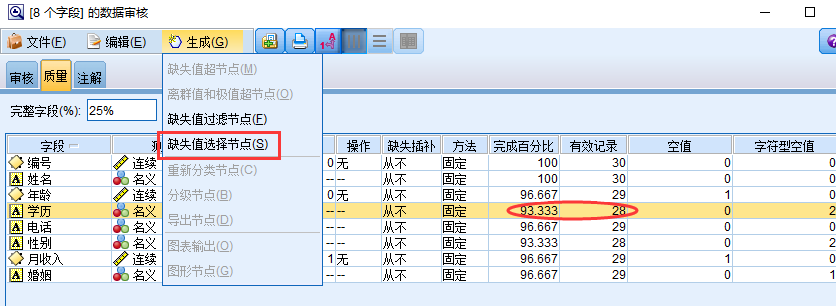


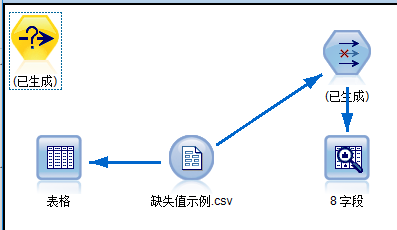
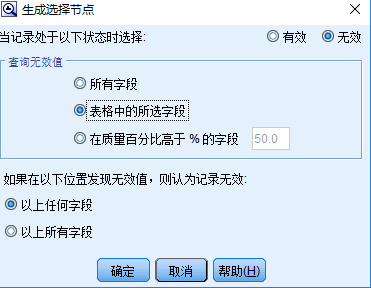


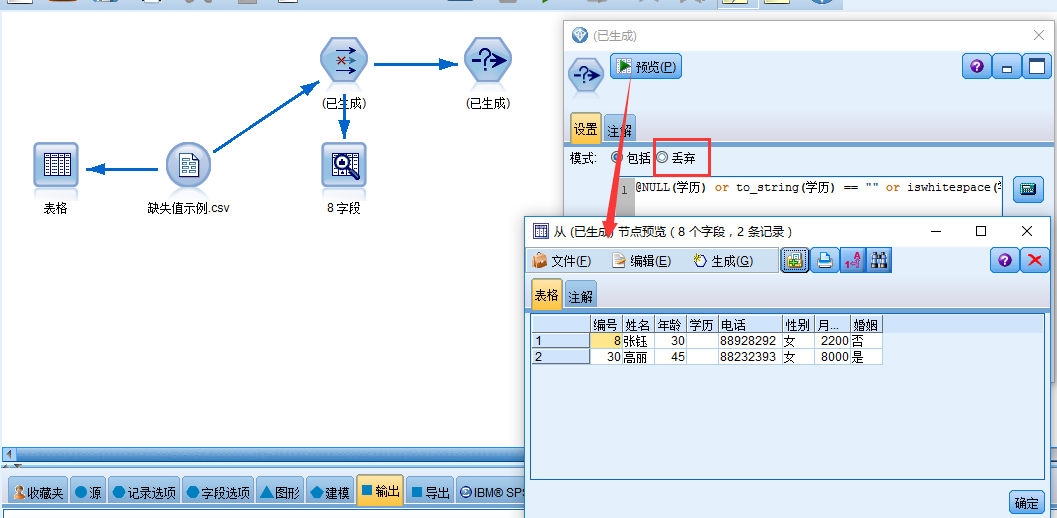
6、通过数据审核节点生成的处理节点删除字段或记录

现在完整记录高于完整字段，删除有缺失值的记录

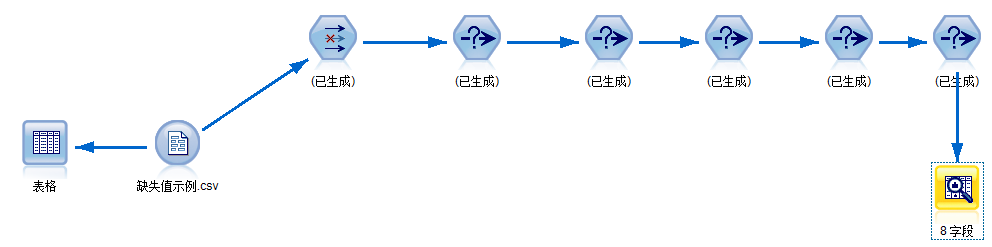
比如学历有缺失值





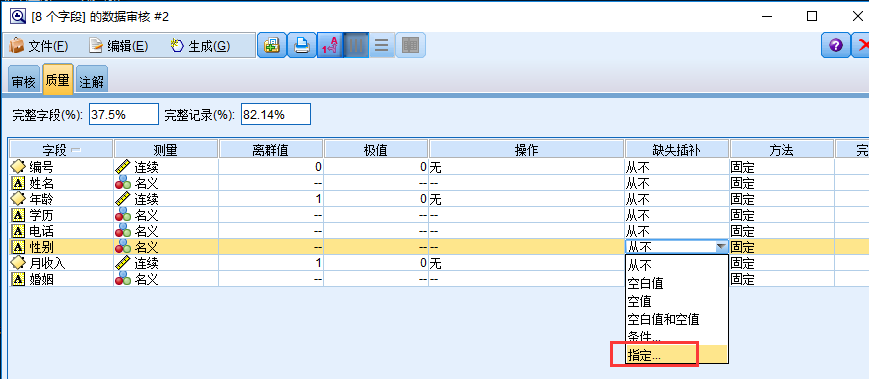


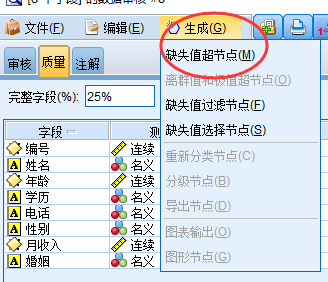
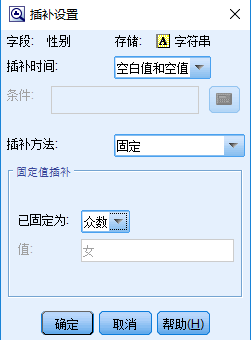
再对其他有缺失值字段进行相同的处理

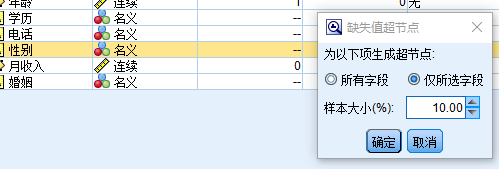


### 【知识点13】缺失值处理——填充

1、通过数据审核节点设定缺失值插补方式

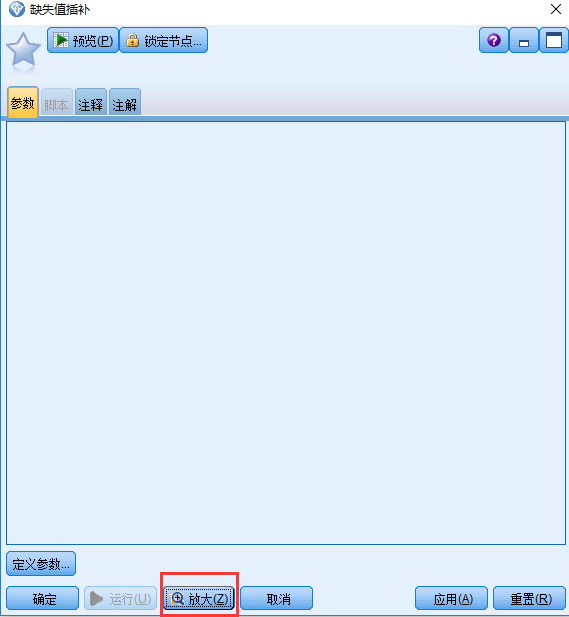


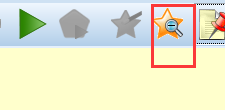
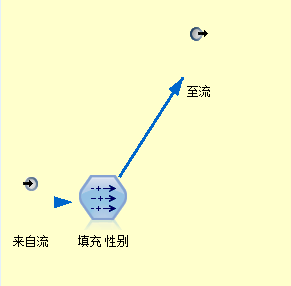




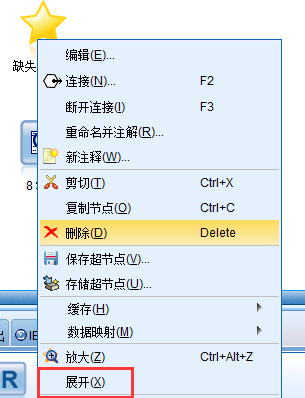
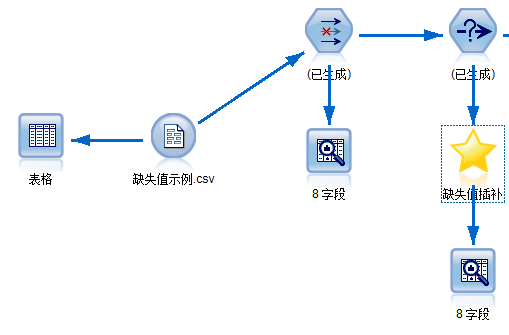


2、通过数据审核节点设定缺失值插补方式





3、通过数据审核节点设定缺失值插补方式



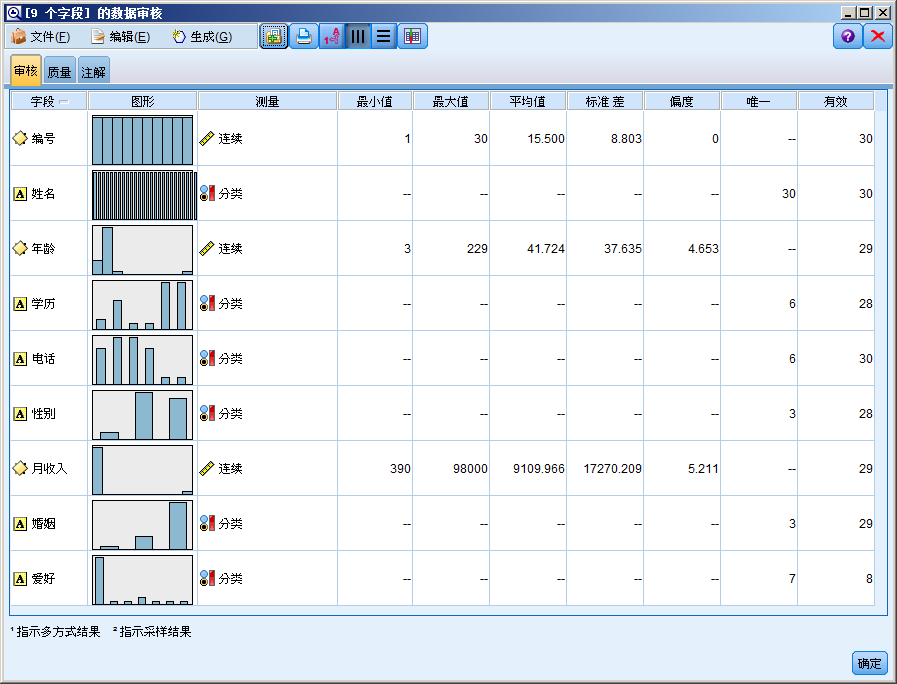
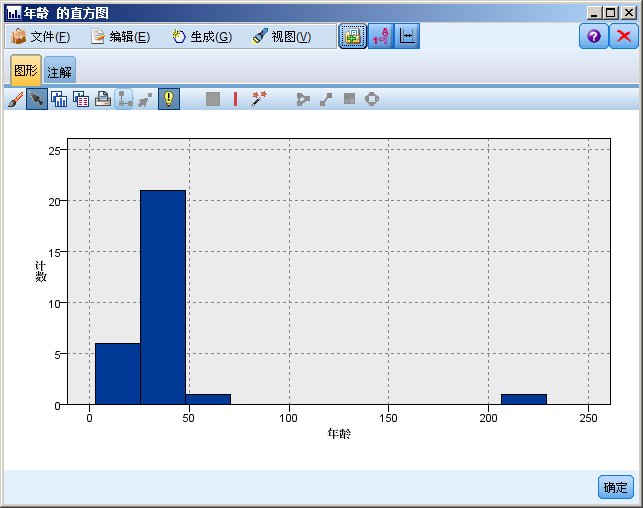


### 【知识点14】异常值分析——单字段分析

异常值就是数据文件中那些和其它值相比有明显不同的值，它们可以通过观察数据分布来确定。

单字段异常值：出现在一个字段中的异常值（极值或离群值）

连续型数据异常值：通过数据审核节点查看

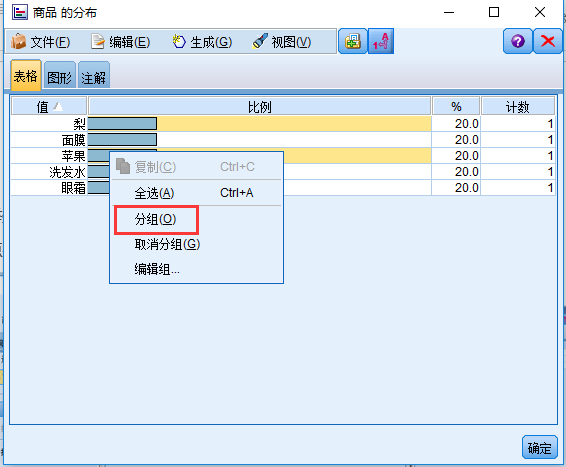
****

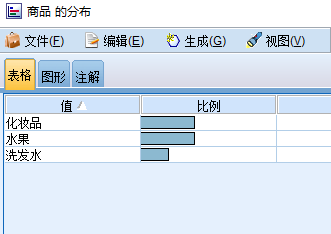
小数据量级枚举类数据的抽象分组

枚举类数据：不连续，可穷举（比如购买的商品，商品种类不连续）

数据源：A001用户消费记录.csv

利用分布图节点实现分组

****

****

### 【知识点15】异常值分析——多字段分析

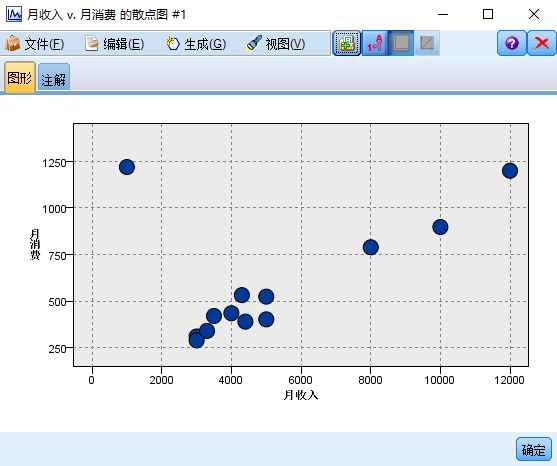
多字段异常值：需要多列组合才能发现的异常值，也称为联合分布异常值

数据源：收入消费.csv

****

联合分布异常值分析：通过散点图进行

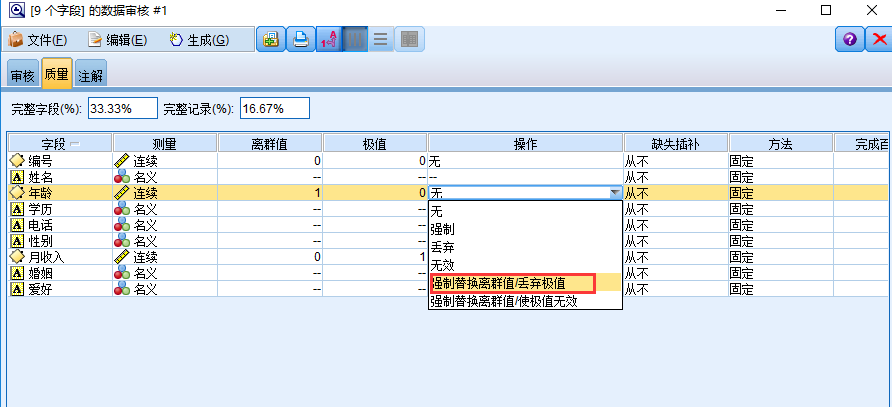


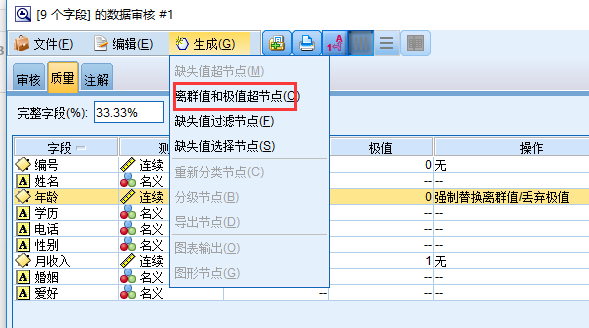


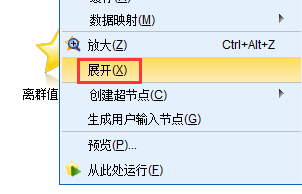
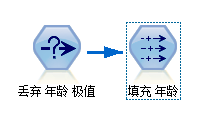
### 【知识点16】异常值处理

单字段异常值处理：通过数据审核节点设置

数据源：缺失值示例.csv

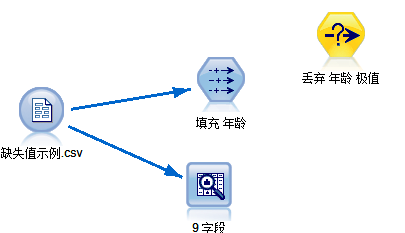






单字段异常值处理：

通过数据审核节点生成的超节点处理

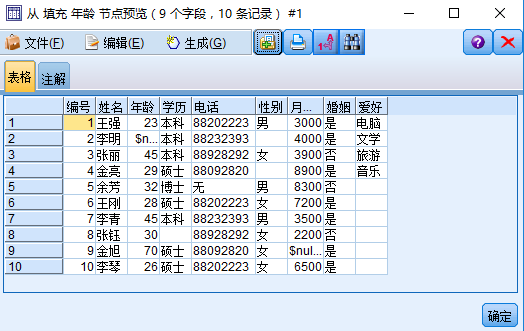






单字段异常值处理：自定义填充节点，以取得更好效果





### 【知识点17】观察数据项之间联系

对于数据挖掘来说在进行真正的建模之前，通过观察数据项之间的关系，特别是输入数据项和目标数据项之间的关系，是非常有意义的，它能快速的让我们对数据之间的关系有个大概了解，精简一些不必要的数据项，提高建模速度和稳定性。

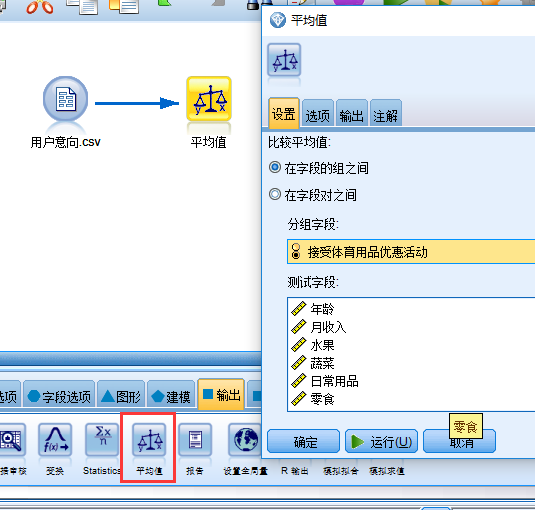
通过均值节点可以观察数值类字段对目标字段的影响力

通过网络图节点可以观察非数值类字段对目标字段的影响力



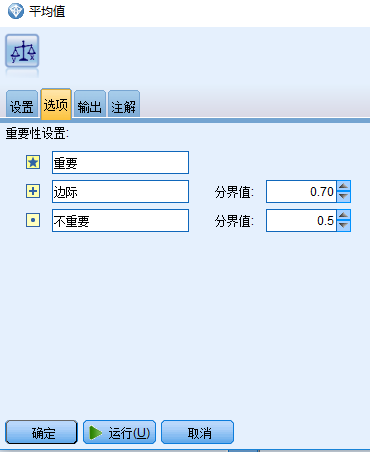
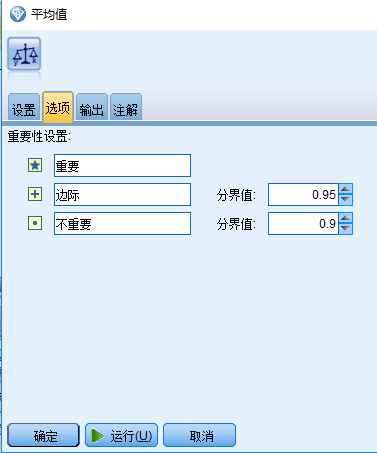
通过均值节点观察数值类字段对目标字段的影响力

数据源：用户意向.csv



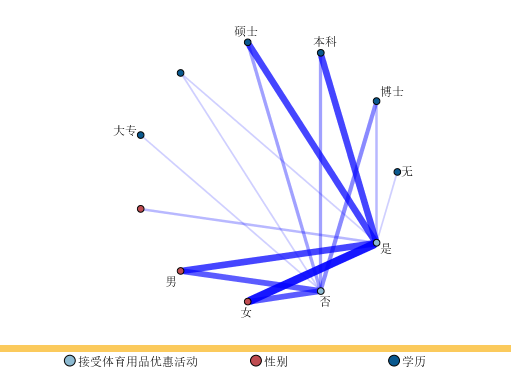
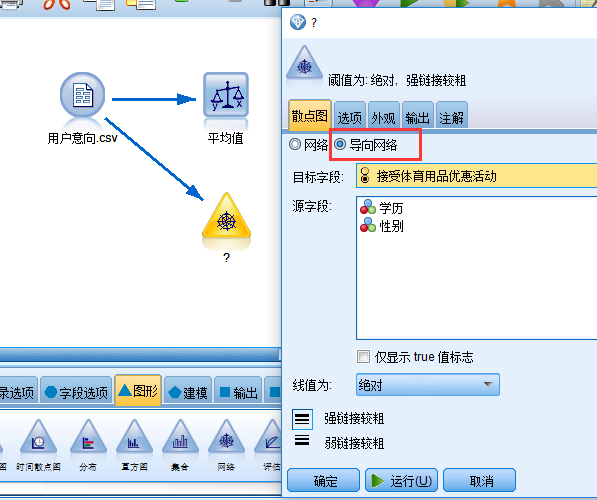


适当的放低点要求，来查看各因素对优惠活动的影响





通过网络图节点观察非数值类字段对目标字段的影响力



## 自主学习作业讲解

无

## 课程总结

* 主要算法介绍
* 自动建模器
* 分类模型
* 聚类模型
* 关联模型
* SPSS MODELER的输出和导出操作
* SPSS MODELER建模操作
* 建模操作
* 数据理解的重要性
* 从商业理解到数据理解
* 缺失值定义
* 缺失值分析
* 缺失值处理——填充
* 异常值分析——单字段分析
* 异常值处理
* 观察数据项之间联系

## 下次自主学习任务布置

#### 观看预习视频

数据建模分析师——Python语音基础

#### 课后作业

1. 参见直播平台
2. Qq群课后习题