**西南财经大学《 数据挖掘 》实验报告**

**年 月 日**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **实验题目** | | 住房贷款信用违约识别与预测案例 | | |
| **选课课号** | |  | | |
| **学 院** | |  | **班 级** |  |
| **姓 名** | |  | **学 号** |  |
| **理论课教师** | | 张英 | **上机指导教师** | 张英 |
| **实验目的及要求：**   * 1. 针对问题环境确定挖掘任务   2. 熟悉在SAS EM 中实施分类挖掘任务的总体步骤和流程   3. 熟悉常用数据探查和预处理节点的使用   4. 熟悉常用分类建模节点中涉及的关键参数设置及结果解读   5. 熟悉利用模型对新数据进行预测的操作   6. 通过实验加深对相应理论知识的理解，增强数据挖掘的实际实施能力   按照要求填入问题的解答或操作截图，及时提交报告到指定位置 | | | | |
| **实验的详细操作步骤：**  **注意：以下截图时，请确保项目和流程图的名称显示在截图中！！！**  **准备工作：**   * **在D盘（如果D盘不可用，则在C盘）建立以你的姓名（拼音）为名称的文件夹，并将案例数据文件拷贝到该文件夹；** * **将实验报告另存为“贷款学号姓名.DOCX”，并存放在上述新建目录下。**   **一、阅读案例背景和数据说明，回答以下问题：**  **1.本案例的挖掘目标是什么？属于哪类挖掘任务？**  **建立 预测模型，预测新的住房贷款申请人未来是否会出现信用违约**  **2. 目标变量是：BAD**  **你关注的目标事件是：申请人出现信用违约**  **是否有变量不能作为输入变量？ 否**  **二、操作**   1. **建立逻辑库，新建挖掘项目**  * **启动SAS9.4（中文），建立逻辑库lab指向准备工作中新建的你的姓名（拼音）为名称的文件夹。** * **导入案例数据集，形成逻辑库lab下的同名数据集，打开该数据集，查看数据，确保导入成功。关闭SAS9.4（中文）** * **打开SAS EM，新建挖掘项目，以你的名字（拼音）dk为项目名，项目对应目录为“d:\dm” （如果D盘不可用，则在C盘）** * **建立逻辑库lab指向准备工作中新建的你的姓名（拼音）为名称的文件夹。** * **新建数据源hmeq, 对应lab逻辑库下的hmeq数据集，了解数据集的规模，设定变量的角色、测量水平。**   **观察Name 和 Type 两列的背景均为灰色，表明这两列为SAS数据集中不可更改的信息。Type 显示了一个变量或者是字符型变量（char）或者是数值型变量（num）。**  **SAS EM使用元数据对如何使用每一个变量做初步估计。在默认的情况下，EM选择2000个随机样本并用样本信息为每一个变量分配一个model role 和一个measurement level. EM同时也根据这一样本计算描述性统计数据并显示出来。**  **如果变量只含有两个非缺失值， EM将其measurement level视为binary；**  **如果数值变量所包含的数值为整数，不同的取值个数为2到10个之间，则EM将其measurement level视为ordinal;**  **如果数值型变量在元数据样本中拥有超过10个不同的数值，则EM将其measurement level视为interval ；**  **如果字符变量取值超过了两个值，则被视为nominal；**  **如果变量名以“target\_”开头，则其model role视为target，否则，视为inpput（日期型变量视为reject）。**  **将变量BAD设置为目标变量，我们关注“违约”情况的发生，所以目标事件是BAD=1，一般我们将目标事件排在前面，因此排序应该选择降序排列。**  **其它nominal变量的排列顺序隐含为升序，排列顺序对以后建立回归模型的系数的估计和解释是有影响的。**  **将变量DELINQ的测量水平设为ORDINAL(有序)**   * **新建流程图，以hmeq你的学号为流程图名**  1. 基本分析和探查  * 将数据源拖入SAS EM数据流设计区内，形成一个输入数据（input data source）节点。   右键点击该节点，选择“编辑变量”查看并确认各变量的测量水平和模型角色。  在该节点的左边属性窗口中，将属性“汇总”设置为“是”，运行该节点，查看变量的统计指标。   * 加入“统计探查”节点，并与输入数据节点相连   **有多少比例的客户是信用违约的？（回答）**  **80.5%**  查看其它各变量的分布和统计指标，哪些变量存在缺失值？（回答并截图）    CLAGE,CLNO,DEBTINK,DELINQ,DEROG,JOB,MORTDUE,NINQ,REASON,VALUE,YOJ都存在缺失值  初步判定哪些输入变量和目标变量相关？（卡方检验结果截图并回答）      初步判断，所有变量都与目标变量相关。   * 加入一个data partition节点，并将其与数据输入节点相连。   Data partition属性窗口中选择分层抽样法（Stratified Sampling）。  将数据分为训练数据集、检验数据集和测试数据集的比例。但是无论怎样分配比例，各个比例的总合必须为100%。此处我们采用默认设置。  分区节点的“聚类”抽样法是针对聚类变量的取值进行随机抽样。  在Data partition属性窗口中点击“变量”旁的省略号，选择所有变量（选择一个变量，然后按CTRL+A）,单击“探查”按钮，可以再次探查各变量的分布直方图。 ？？？   1. 建立一个默认设置的回归模型   在data partition节点之后添加并连接一个Regression节点。  右键单击Regression节点，选择run。运行完毕选择Yes查看结果  对于提升图（选择Y轴指标为累计响应率%Response），本例中响应者就是那些违约的客户（BAD=1）。对于每一个客户而言，建模（此处为回归建模）结果显示了其违约的预测概率，所有的客户都被按照这一预测概率由高到低进行了排序（Descending）。然后，所有的被排好次序的客户又被分为20（或10）个有序组，每一组包含大约5%（或10%）的数据，根据变量BAD的实际值(0和1)，计算每一个有序组内违约的客户的实际比率并绘制了提升图。  如果这个模型是可用的，那么相对靠前的有序组（违约的客户的预测概率相对高）内违约的客户的实际比率应该高于那些相对靠后的有序组（违约的客户的预测概率相对高）内违约的客户的实际比率。  提升图（选择Y轴指标为累计捕获响应率%Captured Response）代表了有序组中包含的违约的客户数量占全部违约客户的百分比。如果我们随机抽取10%的客户，那么将有10%的违约的客户在这一抽样中被抽到。同样地，如果随机抽取20%的客户，就将有20%的违约的客户被抽到。因此，基本线为45°角的上升直线，  提升图（选择Y轴指标为提升度lift），用%Response除以目标事件的比列，其理想形态和前面提到的“提升图（选择Y轴指标为累计响应率%Response）”类似。在选择模型时，我们应考虑前面多大比例的客户是比较理想的拒绝对象。  回答下列问题，有图形依据的请截图：  当前模型是否可用？为什么？      如果可用，模型中应该包含哪些变量？找到对应的系数  如果不可用，可能的原因是什么？  **缺失值，** 4. 建立一个填补了缺失值的回归模型 在Data Partition节点后连接一个impute节点，再添加一个Regression节点。  impute节点可以为每一个有缺失值的变量按一定的规则填补缺失值。  对数值型变量和分类变量分别选择均值和众数（或者均选择决策树）方法进行填补。  运行impute节点和回归节点  查看回归模型的结果。回答下列问题，有图形依据的请截图：  当前模型是否可用？    如果可用，模型中应该包含哪些变量？找到对应的系数  对模型加以解释：哪些变量对目标变量有正相关作用？哪些变量对目标变量有负相关作用？（注意nominal变量的参照水平）    能否确定哪个变量对目标变量的影响程度最大？ 5.数据转换与建立缺失值指示变量 再次查看数据探查的结果（例如：统计探查、数据分区节点探查）  可以发现除部分变量有缺失值外，变量YOJ等分布偏斜、变量DEROG、DELINQ含有大量为0 的观察值，变量NINQ大部分的观察值为0，1或2。  我们试探对缺失值进行填补的同时建立指示变量；对yoj等分布偏斜的数值变量进行log纠偏处理，变量DEROG、DELINQ进行二元化（是否大于0），变量NINQ进行分组。   * 拷贝IMPUTE节点，并将该节点与数据分区节点相连，在impute节点属性窗口中 “指示变量”属性设为“唯一”（为每个有缺失值的变量建立对应的指示变量）   更改变量JOB的替换方法，将缺失值替换为“其它”   * 添加一个“交互式分组”节点，并与上述impute节点相连。   运行该节点。  将变量NINQ值分为三组，即值为0，值为1，和值大于等于2：  在“交互式分组”节点属性窗口中点击“交互式分组”旁的省略号，确保所有变量的新角色都为“输入”。  在左上角“变量选择”列表中选择变量NINQ，单击下面的“分组”页，  通过选择对应分组，并在其右键菜单中选择“合并”、“裂分”操作，实现目标（三组，即值为0，值为1，和值大于等于2）（截图）   * 添加一个数据转换节点，并与上述新加的“交互式分组”节点连接，   运行该节点。  在数据转换节点属性窗口中点击“变量”旁的省略号，通过改变变量对应“方法”，对变量YOJ采取对数变换（log-transformation）进行纠偏，注意：由于原变量YOJ的最小值为0，对0取对数没有定义，故新变量的公式为对原变量加1后再进行对数变换，然后查看新变量的分布情况。  在数据转换节点属性窗口中点击“公式”旁的省略号，创建两个新的变量INDEROG和INDELINQ，对应的定义公式分别为：DEROG>0，DELINQ>0  注意观察转换操作后，原变量的可用状态是否发生变化？  不进行任何转换的变量的转换方法都设为“无”  6.对转换后的数据建立回归模型  拷贝上一个回归节点，并将新的回归节点与数据转换节点相连  在回归节点的属性窗口中，将“模型选择”方法设为“逐步”，并不使用缺省设置，将Stay Significant Level修改为0.025。  运行回归节点，查看回归模型的结果。回答下列问题，有图形依据的请截图：  当前模型是否可用？  如果可用，模型中应该包含哪些变量？找到对应的系数  模型的分类效果如何？  7. 建立决策树模型   * 添加一个Tree节点，并与Data partition节点相连。   决策树采用二分支，节点裂分标准采用信息熵的下降。其它设置不修改。  决策树技术能够直接对缺失值进行处理而不需要进行替换，也会将数值型变量分组处理， 故可直接对原数据建模。  运行决策树节点，查看结果，回答下列问题，有图形依据的请截图：  当前模型是否可用？  如果可用，哪些变量对目标变量比较重要？  当前最佳决策树包含多少个叶子节点？  由树可以推导出怎样的判定规则用于判定客户的“好”“坏”？  模型的分类效果如何？   * 拷贝上述决策树节点，并将新生成的树节点与数据转换节点相连。   运行决策树节点，查看结果，与上一个决策树结果相比，哪个模型的效果更好？为什么？  8.对流程图重新布局  如果你的流程图  模型选择有点凌乱，在空白处右键单击，选择布局中的“水平”或“垂直”，系统会自动整齐排列你的节点。  9.模型选择  添加一个“模型比较”节点，并将上述较好的决策树模型、只填补了缺失值的回归模型、数据转换后的回归模型与该节点相连。  运行该节点，查看结果。  在提升图窗口，右键单击，选择“数据选项”，将“model”对应行的角色设为“分组”，按模型名称显示图形。  哪个模型较优？（截图并回答）  10.使用模型对新数据进行分类   * 创建一个打分节点   打分（Score）节点用于应用评价、保存、合并不同模型的打分代码。  在最后一个回归模型后添加并连接一个打分（Score）节点。   * 在SAS EM 中对新数据直接进行分类   新建数据源，对应sampsio逻辑库下的dmahmeq数据集，将其角色设置为“打分”。  拖动刚建的数据源节点到流程图工作区，将该节点连接到打分节点。  单击打分（Score）节点，在打分节点的属性窗口将“打分数据类型”设置为“视图”，将“验证”和“测试”属性设为“是”。  运行打分节点。 在打分节点属性窗口中查看“导出数据”项下的数据集名称（包括逻辑库名），找到以score开头的数据集，在SAS资源管理器中打开该数据集查看分类结果。  11. 使用Reporter节点生成报告  通过Reporter节点创建一份HTML报告  在模型选择节点之后添加并连接一个Reporter节点（“应用”节点组下），并运行此节点。生成的HTML报告显示了与Reporter节点相连的每一条处理流上的每一个节点的结果。有些结果直接显示在报告中，有些则通过超链接显示。所有的与报告相关的文件都存储在一个文件夹中，它位于项目文件夹的报告文件夹中  **12.** 思考  如果该项目的目标仅是对新的贷款申请者进行分类，而无须对决策原因进行解释，神经网络技术是否是我们的一个技术选择？  建立一个神经网络模型：  在数据转换节点后添加并连接一个“自动神经网络”节点，并与“模型比较”节点相连。  运行此节点并查看模型的预测性能指标、各链接的权重。  运行“模型比较”节点并查看结果，将该模型与先前的决策树、回归模型进行比较，神经网络模型性能如何？  13.完整的挖掘流程图（截图） | | | | |
| **实验体会、发现的问题及解决方法：** | | | | |
| **成 绩** |  | | | |