

数据仓库与数据挖掘 期末作业报告

组 员： 2017526019 刘禾子
 2017526002 邓晏宁
专 业： 计算机科学与技术
 网络工程
提交日期： 2018/1/8

1 分析数据集概述

(一) 数据来源

本实验数据 diamonds.csv 来自 kaggle 网页上供学习的数据集，链接如下：<https://www.kaggle.com/shivam2503/diamonds>

(二) 属性名称与属性类型

carat: 钻石的克拉重量 (0.2-5.01) ;
cut: 切割质量 (一般, 良好, 很好, 高级, 理想) ;
color: 彩色钻石的颜色, 从 J (最差) 到 D (最好) ;
clarity: 钻石清晰度 (I1 (最差), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (最好)) ;
table: 相对于最宽点金刚石的顶部的宽度 (43--95) ;
price: 价格以美元计 (326 美元 - 18,823 美元) ;
x: 长度 (mm) (0--10.74) ;
y: 宽度 (mm) (0--58.9) ;
z: 深度 (mm) (0--31.8) ;

(三) 数据规模

具有 53940 行和 10 个变量的数据。

(四) 数据样例

1	field1	carat	cut	color	clarity	table	price	x	y	z
2	1	0.23	Ideal	E	SI2	55	326	3.95	3.98	2.43
3	2	0.21	Premium	E	SI1	61	326	3.89	3.84	2.31
4	3	0.23	Good	E	VS1	65	327	4.05	4.07	2.31
5	4	0.29	Premium	I	VS2	58	334	4.2	4.23	2.63
6	5	0.31	Good	J	SI2	58	335	4.34	4.35	2.75
7	6	0.24	Very Good	J	VVS2	57	336	3.94	3.96	2.48
8	7	0.24	Very Good	I	VVS1	57	336	3.95	3.98	2.47
9	8	0.26	Very Good	H	SI1	55	337	4.07	4.11	2.53
10	9	0.22	Fair	E	VS2	61	337	3.87	3.78	2.49
11	10	0.23	Very Good	H	VS1	61	338	4	4.05	2.39
12	11	0.3	Good	J	SI1	55	339	4.25	4.28	2.73
13	12	0.23	Ideal	J	VS1	56	340	3.93	3.9	2.46
14	13	0.22	Premium	F	SI1	61	342	3.88	3.84	2.33
15	14	0.31	Ideal	J	SI2	54	344	4.35	4.37	2.71
16	15	0.2	Premium	E	SI2	62	345	3.79	3.75	2.27
17	16	0.32	Premium	E	I1	58	345	4.38	4.42	2.68
18	17	0.3	Ideal	I	SI2	54	348	4.31	4.34	2.68
19	18	0.3	Good	J	SI1	54	351	4.23	4.29	2.7
20	19	0.3	Good	J	SI1	56	351	4.23	4.26	2.71
21	20	0.3	Very Good	J	SI1	59	351	4.21	4.27	2.66
22	21	0.3	Good	I	SI2	56	351	4.26	4.3	2.71
23	22	0.23	Very Good	E	VS2	55	352	3.85	3.92	2.48
24	23	0.23	Very Good	H	VS1	57	353	3.94	3.96	2.41
25	24	0.31	Very Good	J	SI1	62	353	4.39	4.43	2.62
26	25	0.31	Very Good	J	SI1	62	353	4.44	4.47	2.59
27	26	0.23	Very Good	G	VVS2	58	354	3.97	4.01	2.41
28	27	0.24	Premium	I	VS1	57	355	3.97	3.94	2.47
29	28	0.3	Very Good	J	VS2	57	357	4.28	4.3	2.67
30	29	0.23	Very Good	D	VS2	61	357	3.96	3.97	2.4
31	30	0.23	Very Good	F	VS1	57	357	3.96	3.99	2.42
32	31	0.23	Very Good	F	VS1	57	402	4	4.03	2.41
33	32	0.23	Very Good	F	VS1	57	402	4.04	4.06	2.42
34	33	0.23	Very Good	E	VS1	59	402	3.97	4.01	2.42
35	34	0.23	Very Good	E	VS1	58	402	4.01	4.06	2.4

2 分析目标

挖掘目标：根据大量数据挖掘出钻石的切割质量好坏与各个参数的关系；



挖掘模型：训练出可根据各个参数判定钻石切割质量好坏的模型，并测试该模型。

3 分析流设计

modeler 预处理：

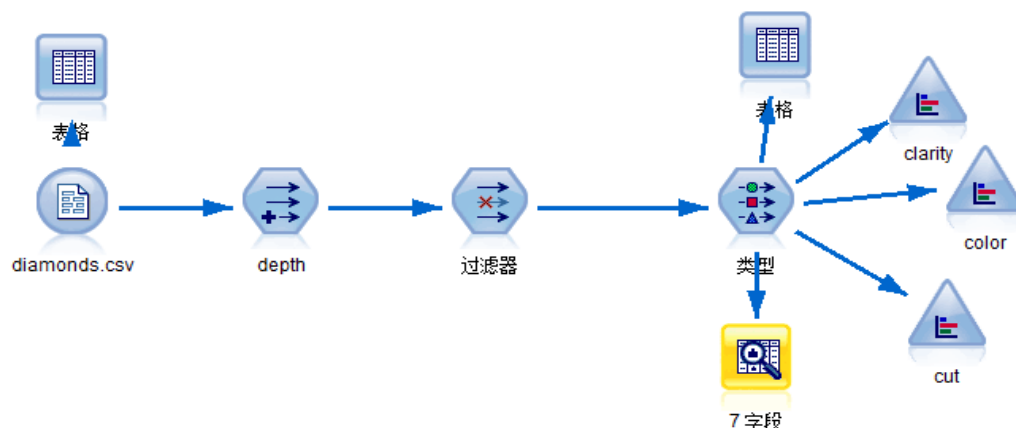
- ① 加入“类型”节点，确定各个字段测量值的属性；
- ② 加入“导出”节点，查资料根据钻石的长度、宽度、深度可计算出用于评估切割质量好坏的另外一个参数，导出新字段 depth，其含义是钻石总深度百分比，以“公式”方式导出，计算公式： $\text{depth} = z / \text{平均值}(x, y) = 2 * z / (x + y)$ ；
- ③ 加入“数据审核”节点，大体分析各个字段的分布情况以及初步分析各个字段之间的关联关系；
- ④ 加入“过滤节点”除去冗余数据，过滤掉计算出 depth 的三个字段 x, y, z。

建模步骤：

- ① 将预处理过的数据流连接“分区”节点，取 60% 用作训练，40% 用作测试；（后经调参发现，6：4 比可调出最佳正确率）
- ② 将数据流中“类型”节点中“角色”一栏下的 cut (切割质量) 字段设为目标  目标，其余变量均设置为输入  输入；
- ③ 将分好区的数据节点连接建模节点“C5.0”，生成模型并浏览；
- ④ 将生成的模型连接“分析”节点对所训练的模型进行分析。

4. 分析流实现

数据预处理：主要工作是除去冗余数据并导出新的字段，对各个字段的分布进行数据审核，分析其大概情况，数据流（文件中 first.str）如下图所示：



处理前的数据样例如下图所示：

	f...	carat	cut	color	clarity	table	price	x	y	z
1	1	0.230	Ideal	E	SI2	55	326	3.950	3.980	2.430
2	2	0.210	Premium	E	SI1	61	326	3.890	3.840	2.310
3	3	0.230	Good	E	VS1	65	327	4.050	4.070	2.310
4	4	0.290	Premium	I	VS2	58	334	4.200	4.230	2.630
5	5	0.310	Good	J	SI2	58	335	4.340	4.350	2.750
6	6	0.240	Very Good	J	VVS2	57	336	3.940	3.960	2.480
7	7	0.240	Very Good	I	VVS1	57	336	3.950	3.980	2.470
8	8	0.260	Very Good	H	SI1	55	337	4.070	4.110	2.530
9	9	0.220	Fair	E	VS2	61	337	3.870	3.780	2.490
10	10	0.230	Very Good	H	VS1	61	338	4.000	4.050	2.390
11	11	0.300	Good	J	SI1	55	339	4.250	4.280	2.730
12	12	0.230	Ideal	J	VS1	56	340	3.930	3.900	2.460
13	13	0.220	Premium	F	SI1	61	342	3.880	3.840	2.330
14	14	0.310	Ideal	J	SI2	54	344	4.350	4.370	2.710
15	15	0.200	Premium	E	SI2	62	345	3.790	3.750	2.270
16	16	0.320	Premium	E	I1	58	345	4.380	4.420	2.680
17	17	0.300	Ideal	I	SI2	54	348	4.310	4.340	2.680
18	18	0.300	Good	J	SI1	54	351	4.230	4.290	2.700
19	19	0.300	Good	J	SI1	56	351	4.230	4.260	2.710
20	20	0.300	Very Good	J	SI1	59	351	4.210	4.270	2.660

处理后的数据样例：

	carat	cut	color	clarity	table	price	depth
1	0.230	Ideal	E	SI2	55	326	61.286
2	0.210	Premium	E	SI1	61	326	59.767
3	0.230	Good	E	VS1	65	327	56.897
4	0.290	Premium	I	VS2	58	334	62.396
5	0.310	Good	J	SI2	58	335	63.291
6	0.240	Very Good	J	VVS2	57	336	62.785
7	0.240	Very Good	I	VVS1	57	336	62.295
8	0.260	Very Good	H	SI1	55	337	61.858
9	0.220	Fair	E	VS2	61	337	65.098
10	0.230	Very Good	H	VS1	61	338	59.379
11	0.300	Good	J	SI1	55	339	64.009
12	0.230	Ideal	J	VS1	56	340	62.835
13	0.220	Premium	F	SI1	61	342	60.363
14	0.310	Ideal	J	SI2	54	344	62.156
15	0.200	Premium	E	SI2	62	345	60.212
16	0.320	Premium	E	I1	58	345	60.909
17	0.300	Ideal	I	SI2	54	348	61.965
18	0.300	Good	J	SI1	54	351	63.380
19	0.300	Good	J	SI1	56	351	63.840
20	0.300	Very Good	J	SI1	59	351	62.736

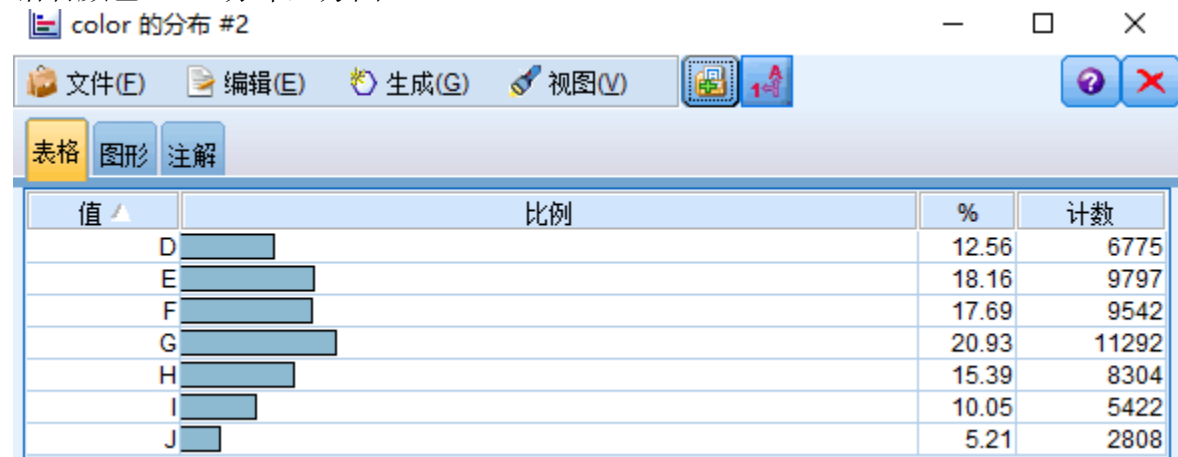
数据审核：

字段	样本图形	测量	最小值	最大值	平均值	标准差	偏度	唯一	有效
carat		连续	0.200	5.010	0.798	0.474	1.117	--	53940
cut		分类	--	--	--	--	--	5	53940
color		分类	--	--	--	--	--	7	53940
clarity		分类	--	--	--	--	--	8	53940
depth		连续	43.000	79.000	61.749	1.433	-0.082	--	53940
table		连续	43	95	57.449	2.240	0.784	--	53940
price		连续	326	18823	3932.800	3989.440	1.618	--	53940

钻石清晰度 clarity 分布直方图：



钻石颜色 color 分布直方图：



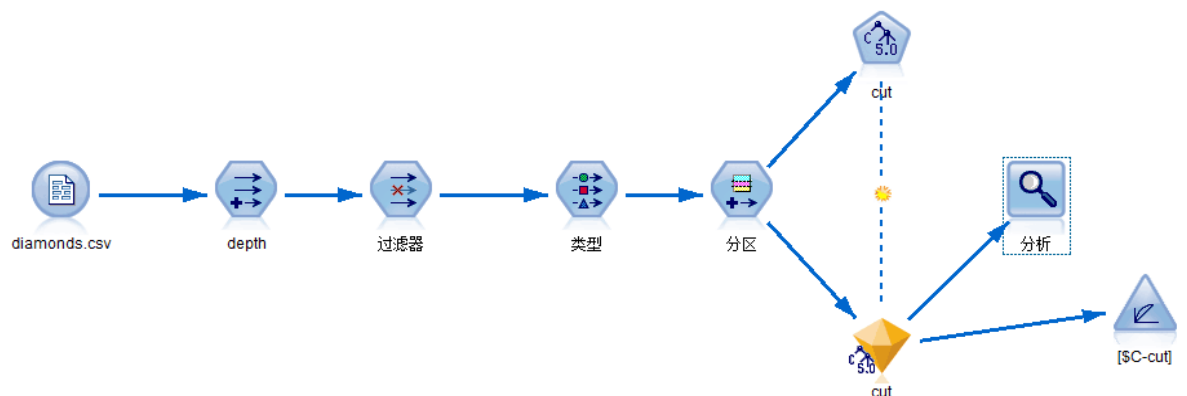
钻石切割质量分布直方图：



从数据预处理结果中，可以看出钻石质量，颜色，和清晰度的分布情况，方便后续分析建模中根据分布情况进行数据的筛选和处理。

分类分析及 C5.0 建模：

分类分析数据流(文件中 second.str)：



对数据类型角色进行定义：目标是训练出能够辨别钻石切割质量好坏(cut)的模型

类型

预览(P)

类型 格式 注解

读取值

清除值

清除所有值

字段	测量	值	缺失	检查	角色
carat	连续	[0.2,5.01]		无	输入
cut	标记	"Very Goo...		无	目标
color	名义	D,E,F,G,H,...		无	输入
clarity	名义	I1,IF,Si1,SI...		无	输入
table	连续	[43,95]		无	输入
price	连续	[326,18823]		无	输入
depth	连续	[0.0,619.2...		无	输入

☒ 查看当前字段
 ☐ 查看未使用的字段设置

确定

取消

应用(A)

重置(R)

数据的分区：

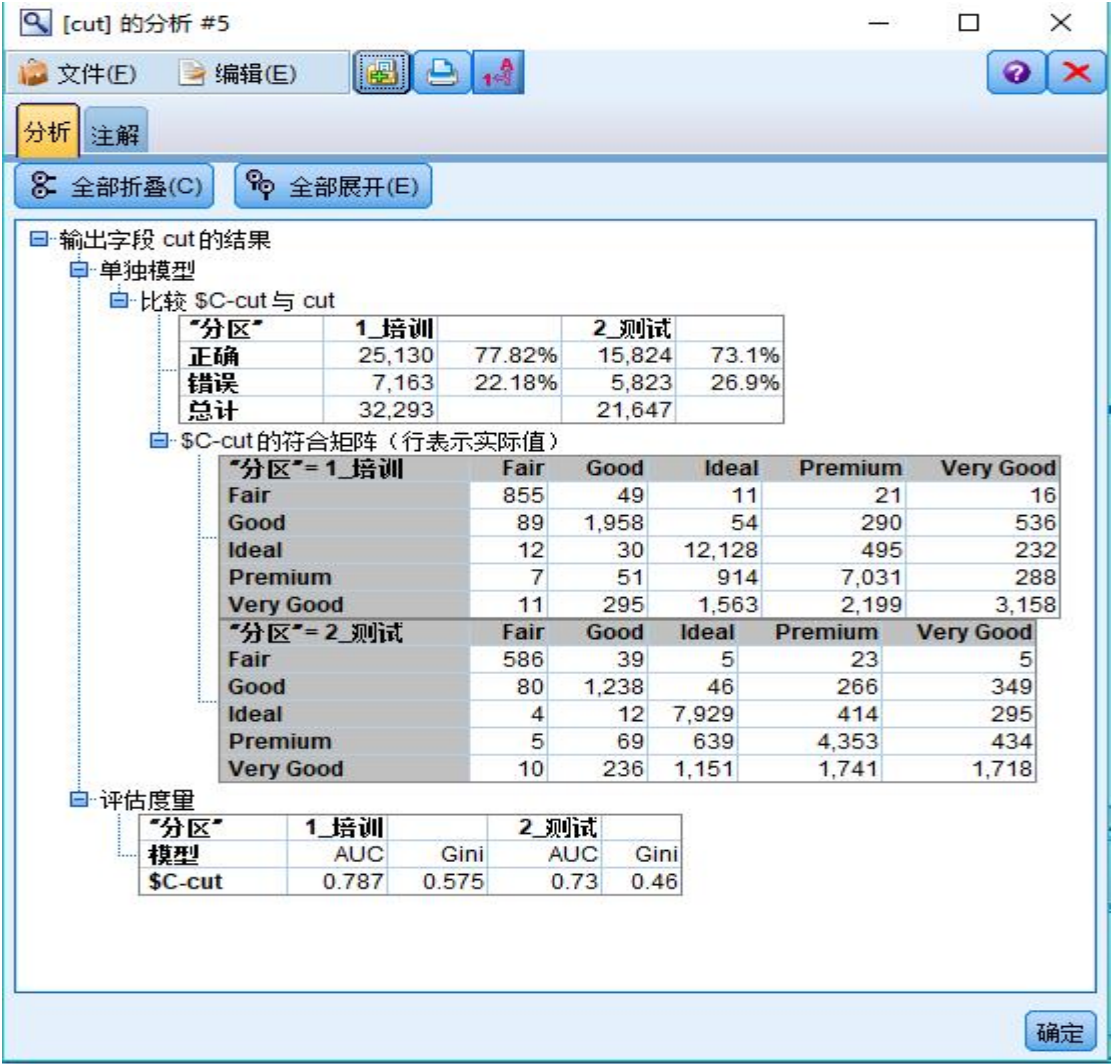


执行训练模型节点：



得到一个深度为 25 的决策树，可从中得出预测 cut 的好坏 depth 和 table 这两个字段为主导因素。

分析模型：

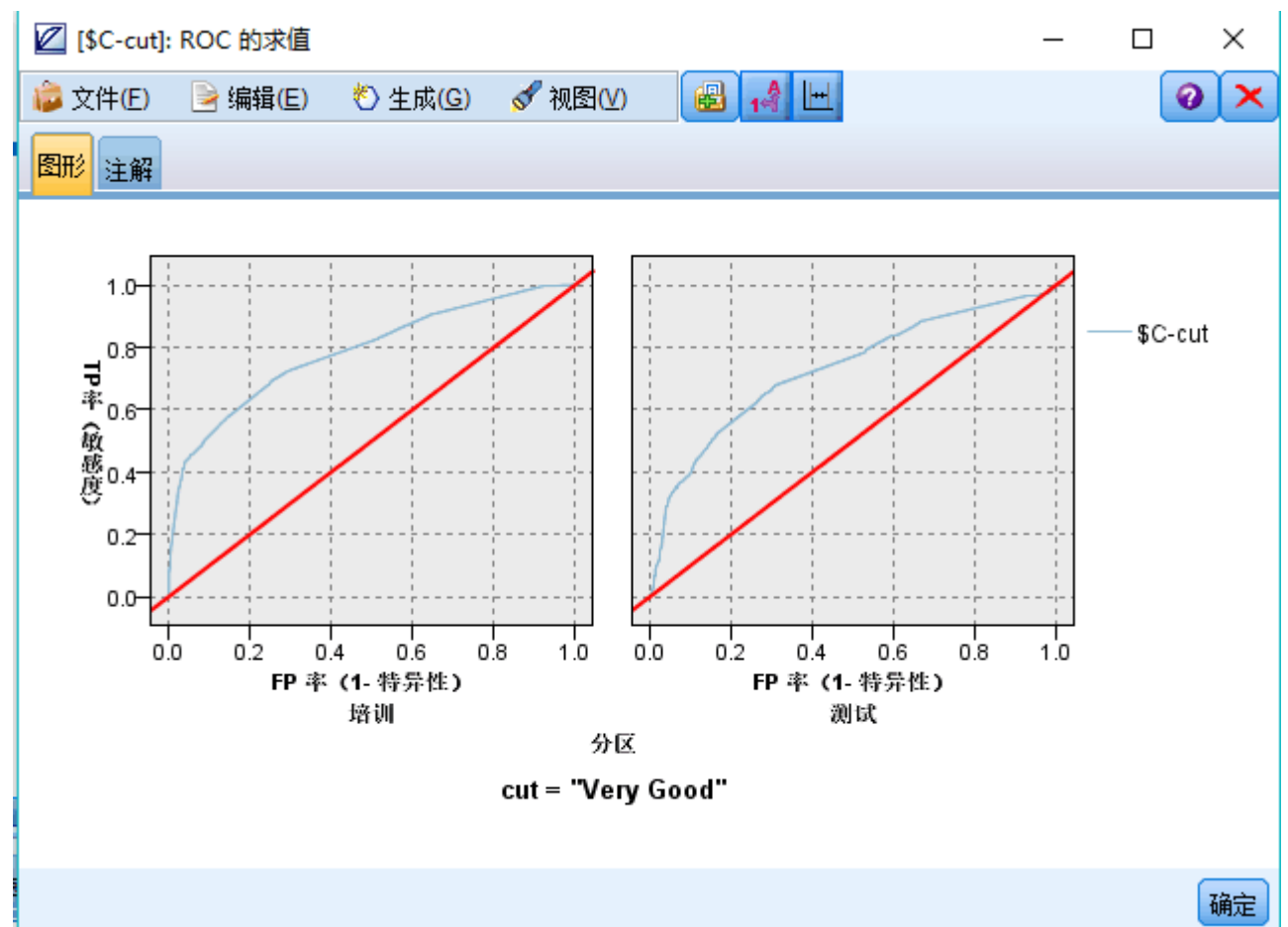


分析结果显示训练数据的预测正确性为 78.32%，测试数据的预测正确率 75.08%，下列符合矩阵的含义为：列元是钻石原本的切割质量，行元是预测的结果，例如测试分区的第一行，钻石原本的切割质量为 Fair 预测为 Fair 的数量为 855，预测为 Good 的数量为 48，预测为 Ideal 的数量为 12，预测为 Premium 为 21，预测为 Very Good 的数量为 16。我们根据符合矩阵计算出召回率和精确率：

"分区" = 1_培 训	Fair	Good	Ideal	Premium	Very Good	召回率
Fair	855	49	11	21	16	0.898109
Good	89	1,958	54	290	536	0.668944
Ideal	12	30	12,128	495	232	0.940374
Premium	7	51	914	7,031	288	0.848028
Very Good	11	295	1,563	2,199	3,158	0.437033
精确率	0.877823	0.821653	0.826721	0.700578	0.746572	

可看出 cut 在 Ideal 情况下召回率较高（达到 94%），Very Good 的召回率较低，

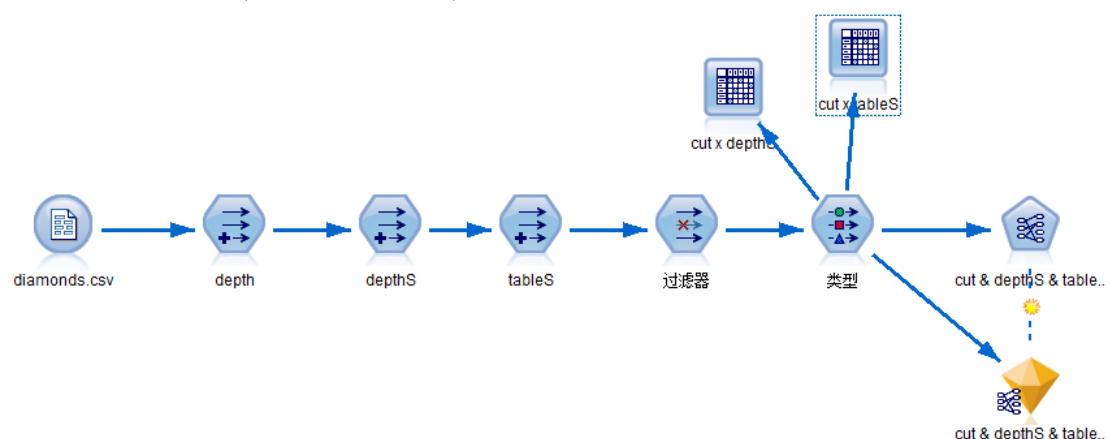
而预测钻石切割质量的精确率大体都在 70% 以上，训练效果良好。
ROC 曲线如下图：



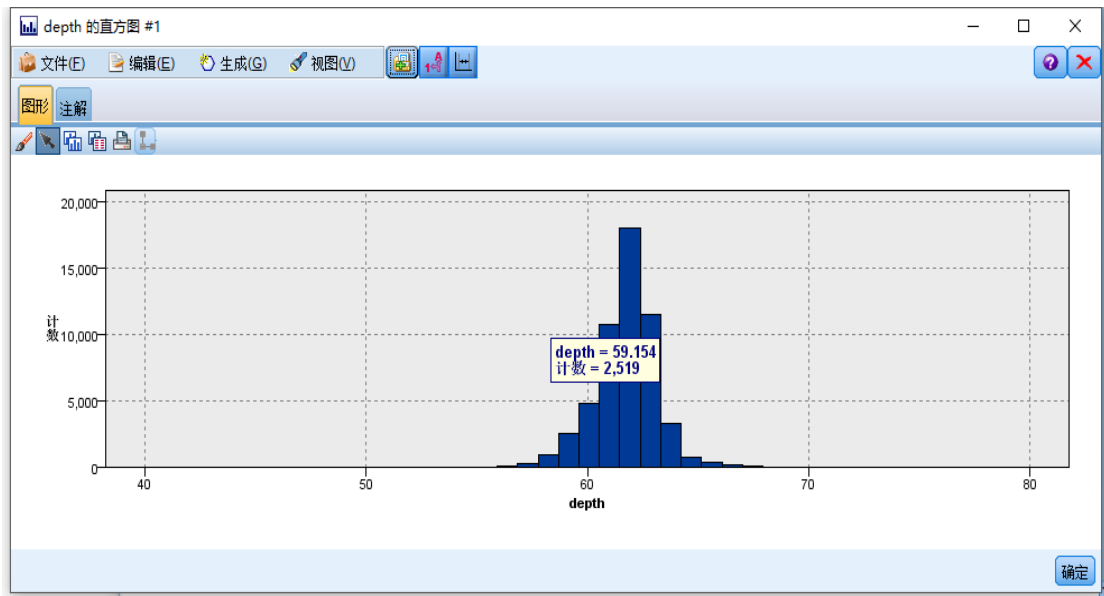
可见，预测 cut=very good（切割质量为非常好）的训练效果要好一些。

关联分析：

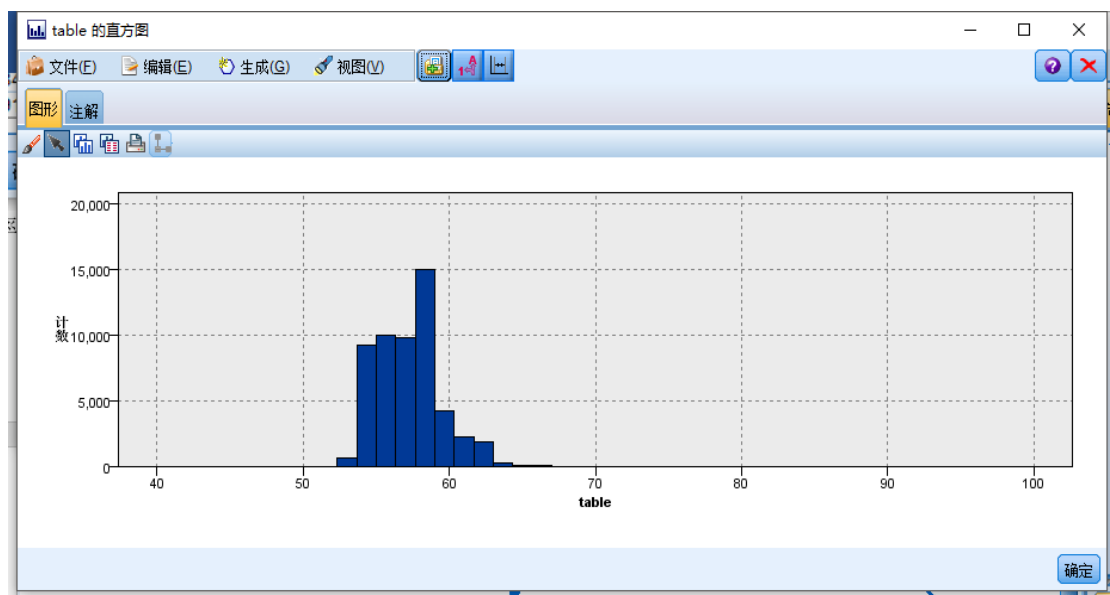
关联分析数据流(文件中 third.str)：



由预测变量重要性的结果对 depth，table，及 cut 三个字段进一步的关联分析：
根据 depth 直方图将 depth 按三个范围（low<61<=middle<63<=high）导出新的名义节点 depthS



再根据 table 直方图将 table 按四个范围
 $(low < 54.333 \leq middle < 58.333 \leq high < 59.667 \leq super)$ 导出新的名义节点 tableS



连接 Apriori 节点建立关联规则挖掘流：

分析结果：初始时建立置信度为 70%，得出的结果并不是很理想，我们期望由 tableS 和 depthS 的属性值作为前项推出 cut 属性值，但条件与结论相反，之后依次调整置信度为 50%、30%，并修改前后项之后分析结果如下图：

后项: cut

前项: depthS
tableS

cut

文件(E) 生成(G) 预览(P)

模型 设置 摘要 注解

排序依据: 置信度百分比 7 的 7

后项	前项	支持度百分比	置信度百分比
cut = Ideal	depthS = middle tableS = middle	29.494	68.332
cut = Premium	tableS = high	12.278	57.074
cut = Ideal	tableS = middle	64.247	52.962
cut = Ideal	depthS = middle	43.276	51.741
cut = Ideal	depthS = high tableS = middle	28.728	46.07
cut = Premium	tableS = super	16.846	42.478
cut = Ideal	depthS = high	41.707	39.961

分析表明，

- 在 depthS=middle（总深度百分比处于 middle）且 tableS=middle（相对于最宽点金刚石的顶部的宽度处于 middle）时得出 cut=ideal（切割质量为理想）的概率为 68.332%，说明在现实生活中判断一个钻石切割是否理想主要看它的总深度百分比和最宽点距离是否处于 middle 级别，这也可从执行中的矩阵节点得到：

cut X depthS 的矩阵

矩阵 外观 注解

	high	low	middle	super
Fair	51	254	79	1225
Good	2196	897	517	1294
Ideal	8990	449	12078	33
Premium	4819	2184	6777	9
Very Good	6441	1556	3892	192

单元格内容: 字段的交叉列表 (包括缺失值)

cut X tableS 的矩阵

矩阵 外观 注解

	high	low	middle	super
Fair	171	103	770	566
Good	490	239	2335	1842
Ideal	327	2653	18354	217
Premium	3780	120	6031	3860
Very Good	1855	460	7165	2602

单元格内容: 字段的交叉列表 (包括缺失值)

- 在 tableS=high 时得出 cut=Premium 的概率是 57.074%，在 tableS=middle 时推出 cut=ideal 的概率为 52.962%且该项的支持率也达到了 64.247%，说明大多数情况下，切割质量为优质的钻石其相对于最宽点金刚石的顶部的宽度是处于 high 级别的
- 在 tableS=middle 单独条件下和在 depthS=middle 单独条件下，cut=ideal 的概率也达到了 51%左右，比两者均为前项时得到的 cut=ideal 的概率要低一些
- 总结：当钻石的总深度百分比（depth）和其相对于最宽点金刚石顶部的宽度（table）都处于 middle 时，钻石的切割质量属理想型，当这两个参数有处于 high 或者 super 级别的钻石质量就次一些，切割质量为高级，进而

可推出普通的一般钻石其 table 值和 depth 值都是处于较低的水平（范围之在 low 和 middle 之间）。

5 数据分析总结

此次分析的初期发现钻石的切割质量与各个参数的关联不太明显，只发现部分参数与切割质量的关联关系，之后对数据进行 Apriori 关联分析并未得出理想的结果，然后筛选了部分参数进行针对性的关联分析得到了较高的置信度，与我们预计的结果有出入，我们的预计是 cut 与 clarity(清晰度)、depth(总深度百分比)关联较大，并且实验结果中 cut 是作为后项（即在 XX 条件下得到 cut 不同属性值的概率），最后调整了分析节点的前后项以及置信度，重新构造模型之后情况良好才得到了我们所需要的结论。整个实践过程中我们加强了自身对于建模软件的实践操作能力，对各个分析节点的特点及作用也有了较深的认识，但是还缺乏一些对所挖掘的数据背后隐藏的一些意义的觉察能力，在以后的学习生活中还有待提高。

6 分工说明

组员刘禾子：负责前期数据源的获取，分类分析建模，实验报告撰写；

组员邓晏宁：负责数据预处理，关联规则分析建模，实验报告撰写。