**评分卡——案例1**

**转自：http://blog.csdn.net/textboy/article/details/47008125**

**一、数据预处理**

**1、数据清洗（data cleaning）**

（1）缺失值处理（missingdata processing）

无缺失值。

（2）去噪声（noisy dataprocessing）

（未有时间研究）

（3）去异常值（outlierprocessing）

?

（4）共线性变量处理（pairwisecorrelations processing）

VIF （未有时间研究）

**2、数据集成（data integration）**

单一数据来源，**[数据结构](http://lib.csdn.net/base/datastructure" \o "算法与数据结构知识库" \t "_blank)**也一致。无需再集成。

**二、导入数据**

分析：

|  |  |
| --- | --- |
| 数据来源 | https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data) |
| 自变量-连续型 | V2,V5,V8,V11,V13,V16,V18 |
| 自变量-分类型 | V1,V3,V4,V6,V7,V9,V10,V12,V14,V15,V17,V19,V20 |
| 因变量y | V21 |
| 变量释义 | https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data) |

R程序：

|  |
| --- |
| rawdata = read.table("D:/personal/knowledge/dataMining/dataset/german/german.data",header=F)  colnames(rawdata)[21] <- "y"  # rename response variable  str(rawdata) |

**三、数据分区**

分析：

|  |  |
| --- | --- |
| 训练数据 | 从总样本中抽样600条 |
| 验证数据 | 剩余的400条 |

R程序：

|  |
| --- |
| trainIdx <- sample(nrow(rawdata), round(0.6\*nrow(rawdata)))  traindata <- rawdata[trainIdx,]  validdata <- rawdata[-trainIdx,]  nrow(traindata)  # result: 600 |

**四、交互式分组（discretization）**

**1、连续型数据离散化**

（1）利用最优准则（基于ConditionalInference Trees）进行分组

R程序：

|  |
| --- |
| # 需转换y从1-2变量变为0-1变量才到调用smbinning  replace2to0 <- function(x) {    n <- nrow(x);    for (i in 1:n) {      if (x[i,21] %in% c("2")) {        x[i,21] <- 0;      }    }    return(x);  }  updtraindata = replace2to0(traindata)    # binning cutoff calculation  library(smbinning)  V2bin=smbinning(df=updtraindata, y="y", x="V2", p=0.05)  V2bin$ivtable  V2bin$bands  # need install package "smbinning" |

结果：

|  |
| --- |
| <= 11, <= 26, <= 72 |

R程序：

|  |
| --- |
| # binning  bin <- function(x, cutoffmin, cutoffmax) {    n <- length(x);    for (i in 1:n) {      if (cutoffmin < x[i] && x[i] <= cutoffmax) {        x[i] <- 1;      } else {        x[i] <- 0;      }    }    return(x);  }  V2bin1 <- bin(updtraindata$V2,0,11)  V2bin2 <- bin(updtraindata$V2,11,26)  V2bin3 <- bin(updtraindata$V2,26,72) |

这只是V2，其它像V5,V13也一样处理~~，如下：

R程序：

|  |
| --- |
| V5bin=smbinning(df=updtraindata, y="y", x="V5", p=0.05)  V5bin$ivtable  V5bin$bands  V5bin1 <- bin(updtraindata$V5,250,6110)  V5bin2 <- bin(updtraindata$V5,6110, 15945)    V13bin=smbinning(df=updtraindata, y="y", x="V13", p=0.05)  V13bin  # 结果竟然是"No Bins" |

V13结果竟然是"No Bins"，不知是不是均匀分布不能分箱了，网上也查不到，那就不分吧。

其它，V8,V11,V16,V18实为分类型变量。如：

R程序：

|  |
| --- |
| summary(updtraindata$V8) |

结果：

|  |
| --- |
| Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    1.000   2.000   3.000   3.042   4.000   4.000 |

变量合并，R程序：

|  |
| --- |
| # 插入新V2, V5  updtraindata <- cbind(updtraindata,V2bin1)  updtraindata <- cbind(updtraindata,V2bin2)  updtraindata <- cbind(updtraindata,V2bin3)  updtraindata <- cbind(updtraindata,V5bin1)  updtraindata <- cbind(updtraindata,V5bin2)  # 转换格式  updtraindata$V2bin1 <- as.factor(updtraindata$V2bin1)  updtraindata$V2bin2 <- as.factor(updtraindata$V2bin2)  updtraindata$V2bin3 <- as.factor(updtraindata$V2bin3)  updtraindata$V5bin1 <- as.factor(updtraindata$V5bin1)  updtraindata$V5bin2 <- as.factor(updtraindata$V5bin2)  # 删除原V2, V5  updtraindata$V2 <- NULL  updtraindata$V5 <- NULL  str(updtraindata) |

结果：

|  |
| --- |
| # updtraindata结构  'data.frame':   600 obs. of  24 variables:   $ V1    : Factor w/ 4 levels "A11","A12","A13",..: 1 4 2 2 3 1 4 4 4 2 ...   $ V3    : Factor w/ 5 levels "A30","A31","A32",..: 2 5 4 3 5 3 5 5 3 5 ...   $ V4    : Factor w/ 10 levels "A40","A41","A410",..: 1 5 2 5 1 5 4 1 6 2 ...   $ V6    : Factor w/ 5 levels "A61","A62","A63",..: 5 1 5 1 5 1 1 1 1 2 ...   $ V7    : Factor w/ 5 levels "A71","A72","A73",..: 3 5 4 3 4 4 1 5 2 1 ...   $ V8    : int  4 4 3 2 4 1 2 4 1 3 ...   $ V9    : Factor w/ 4 levels "A91","A92","A93",..: 3 3 3 2 3 4 3 3 2 3 ...   $ V10   : Factor w/ 3 levels "A101","A102",..: 1 1 1 1 1 3 1 1 1 1 ...   $ V11   : int  2 4 4 2 2 1 3 1 4 4 ...   $ V12   : Factor w/ 4 levels "A121","A122",..: 2 3 2 1 1 1 3 1 1 3 ...   $ V13   : int  40 46 36 22 37 34 31 38 23 27 ...   $ V14   : Factor w/ 3 levels "A141","A142",..: 3 3 1 3 1 3 3 3 3 3 ...   $ V15   : Factor w/ 3 levels "A151","A152",..: 2 2 1 2 2 2 2 2 1 2 ...   $ V16   : int  2 2 2 1 2 2 1 2 1 1 ...   $ V17   : Factor w/ 4 levels "A171","A172",..: 2 3 4 3 2 3 4 2 3 3 ...   $ V18   : int  2 1 2 1 2 1 1 2 1 1 ...   $ V19   : Factor w/ 2 levels "A191","A192": 1 2 2 1 1 2 1 1 2 1 ...   $ V20   : Factor w/ 2 levels "A201","A202": 1 1 1 1 1 1 1 1 1 1 ...   $ y     : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 2 2 2 2 ...   $ V2bin1: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 2 1 ...   $ V2bin2: Factor w/ 2 levels "0","1": 2 2 2 1 2 1 1 2 1 1 ...   $ V2bin3: Factor w/ 2 levels "0","1": 1 1 1 2 1 2 1 1 1 2 ...   $ V5bin1: Factor w/ 2 levels "0","1": 2 2 1 2 2 1 2 2 2 2 ...   $ V5bin2: Factor w/ 2 levels "0","1": 1 1 2 1 1 2 1 1 1 1 ... |

（2）使用WoE进行离散化处理

（见WoE建模阶段处理）

**2、分类型数据离散化**

（暂不处理）

**五、模型选择**

**1、GLM-logistic回归（GLM logistic regression）**

**（1）WoE建模（Modeling）**

我们结合使用信用评分卡中的WoE（Weight of Evidence证据权重）对连续型变量进行离散化处理。

R程序：

|  |
| --- |
| library(klaR)  woemodel <- woe(y~., data = updtraindata, zeroadj=0.5, appont = TRUE)  # 需安装klaR包，install.packages("klaR") |

**（2）IV检验（Examine）**

分析：

使用IV（Information Value 信息价值）检验，检验标准如下：

|  |  |
| --- | --- |
| **Information Value** | **Predictive Power** |
| < 0.02 | useless for prediction |
| 0.02 to 0.1 | Weak predictor |
| 0.1 to 0.3 | Medium predictor |
| 0.3 to 0.5 | Strong predictor |
| >0.5 | too good to be true |

R程序：

|  |
| --- |
| woemodel |

结果：

|  |
| --- |
| IV  V1     0.6948970820  V3     0.3634078216  V4     0.3014986700  V2bin1 0.2214788425  V12    0.1827822608  V7     0.1598300489  V6     0.1584984650  V2bin3 0.1380258581  V15    0.0746645819  V5bin2 0.0738721662  V14    0.0699081960  V5bin1 0.0697554006  V20    0.0636595749  V9     0.0415308555  V10    0.0185753500  V19    0.0170747941  V17    0.0078521265  V2bin2 0.0002055111 |

通过结果观测，我们发现<0.02的变量有：V2bin2, V10, V17, V19，>0.5的变量有：V1。

V1: Status of existing checking account

V2bin2: 11 < Duration in month<= 26

V10: Other debtors / guarantors

V17: Job

V19: Telephone

由此得知，V1, V2bin2, V10, V17,V19都不应直接放入模型。（就这样就行?）

**（3）logistic建模（Modeling）**

Logistic Regression with Weight of Evidence。

R程序：

|  |
| --- |
| woedata <- predict(woemodel, updtraindata, replace = TRUE)  woedata$woe.V1 <- NULL  woedata$woe.V2bin2 <- NULL  woedata$woe.V10 <- NULL  woedata$woe.V17 <- NULL  woedata$woe.V19 <- NULL  str(woedata)    logit.glm <- glm(y~., family=binomial, data=woedata) |

结果：

|  |
| --- |
| > str(woedata)  'data.frame':   600 obs. of  19 variables:   $ V8        : int  4 4 3 2 4 1 2 4 1 3 ...   $ V11       : int  2 4 4 2 2 1 3 1 4 4 ...   $ V13       : int  40 46 36 22 37 34 31 38 23 27 ...   $ V16       : int  2 2 2 1 2 2 1 2 1 1 ...   $ V18       : int  2 1 2 1 2 1 1 2 1 1 ...   $ y         : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 2 2 2 2 ...   $ woe.V3    : num  1.2341 -0.851 0.1797 0.0805 -0.851 ...   $ woe.V4    : num  0.506 -0.537 -1.05 -0.537 0.506 ...   $ woe.V6    : num  -0.56 0.241 -0.56 0.241 -0.56 ...   $ woe.V7    : num  0.0448 -0.2993 -0.4645 0.0448 -0.4645 ...   $ woe.V9    : num  -0.176 -0.176 -0.176 0.194 -0.176 ...   $ woe.V12   : num  0.12648 -0.00817 0.12648 -0.6117 -0.6117 ...   $ woe.V14   : num  -0.136 -0.136 0.537 -0.136 0.537 ...   $ woe.V15   : num  -0.183 -0.183 0.349 -0.183 -0.183 ...   $ woe.V20   : num  0.0405 0.0405 0.0405 0.0405 0.0405 ...   $ woe.V2bin1: num  0.179 0.179 0.179 0.179 0.179 ...   $ woe.V2bin3: num  -0.219 -0.219 -0.219 0.638 -0.219 ...   $ woe.V5bin1: num  -0.118 -0.118 0.593 -0.118 -0.118 ...   $ woe.V5bin2: num  -0.121 -0.121 0.613 -0.121 -0.121 ... |

**（4）z统计量及AIC检验（Examine）**

R程序：

|  |
| --- |
| summary(logit.glm) |

结果：

|  |
| --- |
| Coefficients:               Estimate Std. Error z value     Pr(>|z|)  (Intercept)   1.43947    1.02810   1.400     0.161475  V8           -0.32997    0.10459  -3.155     0.001606 \*\*  V11           0.06341    0.10359   0.612     0.540483  V13           0.01640    0.01072   1.529     0.126213  V16           0.05905    0.19656   0.300     0.763847  V18          -0.42953    0.29474  -1.457     0.145023  woe.V3       -0.87996    0.18595  -4.732 0.0000022198 \*\*\*  woe.V4       -1.09751    0.19591  -5.602 0.0000000212 \*\*\*  woe.V6       -1.09784    0.28430  -3.862     0.000113 \*\*\*  woe.V7       -0.75943    0.27101  -2.802     0.005076 \*\*  woe.V9       -1.45651    0.55785  -2.611     0.009029 \*\*  woe.V12      -0.84312    0.29247  -2.883     0.003942 \*\*  woe.V14      -0.95227    0.38731  -2.459     0.013945 \*  woe.V15      -0.42942    0.43532  -0.986     0.323915  woe.V20      -0.67652    0.49786  -1.359     0.174189  woe.V2bin1   -0.77827    0.25723  -3.026     0.002481 \*\*  woe.V2bin3   -0.56849    0.31997  -1.777     0.075615 .  woe.V5bin1   13.95697  752.97692   0.019     0.985211  woe.V5bin2  -13.93934  728.95510  -0.019     0.984743  ---  Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1    (Dispersion parameter for binomial family taken to be 1)        Null deviance: 758.15  on 599  degrees of freedom  Residual deviance: 569.55  on 581  degrees of freedom  AIC: 607.55 |

通过结果观测，我们发现V2bin3大于0.1显著性水平，(Intercept)、V11、V13、V16、V18、V15、V20、V5bin1、V5bin2大于0.05显著性水平，这些变量接受原假设，对因变量信用风险无显著影响。

V5：Credit amount

V11：Present residence since

V13：Age in years

V15：Housing

V16：Number of existing credits at this bank

V18：Number of people being liable to provide maintenance for

V20：foreign worker

AIC值为607.55，后面逐步回归时及模型比较时会用上。

**（5）逐步回归建模（Modeling）**

我们使用逐步回归分析来解决参数检验不显著的情况，应用 stepwise logistic regression。

R程序：

|  |
| --- |
| logit.glm.step <- step(logit.glm, direction="both") |

最后一次叠代结果：

|  |
| --- |
| Df Deviance    AIC  <none>            575.72 599.72  - woe.V20     1   577.75 599.75  + V13         1   573.77 599.77  + V18         1   574.38 600.38  + woe.V5bin2  1   574.92 600.92  + woe.V5bin1  1   574.94 600.94  + woe.V15     1   575.12 601.12  + V11         1   575.25 601.25  + V16         1   575.60 601.60  - woe.V14     1   581.19 603.19  - woe.V2bin3  1   581.25 603.25  - woe.V9      1   581.97 603.97  - V8          1   584.55 606.55  - woe.V2bin1  1   586.57 608.57  - woe.V7      1   586.80 608.80  - woe.V12     1   589.09 611.09  - woe.V6      1   593.11 615.11  - woe.V3      1   606.66 628.66  - woe.V4      1   609.98 631.98 |

**（6）z统计量及AIC检验（Examine）**

R程序：

|  |
| --- |
| summary(logit.glm.step) |

结果：

|  |
| --- |
| Coefficients:              Estimate Std. Error z value     Pr(>|z|)  (Intercept)   1.6426     0.3338   4.921 0.0000008619 \*\*\*  V8           -0.2939     0.1005  -2.925     0.003445 \*\*  woe.V3       -0.9415     0.1781  -5.286 0.0000001253 \*\*\*  woe.V4       -1.0735     0.1948  -5.512 0.0000000355 \*\*\*  woe.V6       -1.0961     0.2777  -3.947 0.0000792685 \*\*\*  woe.V7       -0.8667     0.2622  -3.306     0.000947 \*\*\*  woe.V9       -1.3254     0.5323  -2.490     0.012768 \*  woe.V12      -0.9126     0.2530  -3.607     0.000310 \*\*\*  woe.V14      -0.8914     0.3794  -2.349     0.018816 \*  woe.V20      -0.6444     0.4970  -1.296     0.194827  woe.V2bin1   -0.7825     0.2545  -3.075     0.002106 \*\*  woe.V2bin3   -0.6766     0.2877  -2.352     0.018672 \*  ---  Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1    (Dispersion parameter for binomial family taken to be 1)        Null deviance: 758.15  on 599  degrees of freedom  Residual deviance: 575.72  on 588  degrees of freedom  AIC: 599.72 |

在逐步回归之后，V5、V11、V13、V15、V16、V18去掉， V20保留。各参数除了V20全部通过显著性检验，这里V20依然保留是因为去掉后AIC反而更高。同时，AIC为599.72，小于原来的 607.55，表明优先考虑逐步回归后的模型。同时该AIC也比没有进行交互式分组的AIC值要小，说明交互式分组后模型更优。

**（7）其它检验（Examine）**

（a）ROC/AUC、Gini检验

以V2bin1为例：

R程序：

|  |
| --- |
| rcorr.cens(woedata$woe.V2bin1,woedata$y) |

结果：

|  |
| --- |
| C Index             Dxy       0.42293898     -0.15412204 |

分析：

C Index代表AUC，Dxy代表Gini系数

由于变量较多，以两个变量为例，归纳结果如下：

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | variable | attr | woe | IV | AUC | Gini | | V2bin1 | 0 | 0.1793342 | 0.221478843 | 0.42293898 | -0.15412204 | | V2bin1 | 1 | -1.2577013 | 0.221478843 | 0.42293898 | -0.15412204 | | V3 | A30 | 1.47707202 | 0.363407822 | 0.35755961 | -0.28488078 | | V3 | A31 | 1.23412584 | 0.363407822 | 0.35755961 | -0.28488078 | | V3 | A32 | 0.08050692 | 0.363407822 | 0.35755961 | -0.28488078 | | V3 | A33 | 0.17968477 | 0.363407822 | 0.35755961 | -0.28488078 | | V3 | A34 | -0.85104637 | 0.363407822 | 0.35755961 | -0.28488078 | |

（尼马?!, 难道全部不通过，所有变量（不仅V2、V3）Gini<0.02, AUC<0.5，￥%#￥%）

**2、GAM-logistic回归（GAM logistic regression）**

（后补）

**3、模型比较（Model comparison）**

（后补）

**4、模型验证（Model validation）**

理论：

|  |
| --- |
| Logit变换 - |

R程序：

|  |
| --- |
| updvaliddata = replace2to0(validdata)    V2bin1 <- bin(updvaliddata$V2,0,11)  V2bin2 <- bin(updvaliddata$V2,11,26)  V2bin3 <- bin(updvaliddata$V2,26,72)  V5bin1 <- bin(updvaliddata$V5,250,6110)  V5bin2 <- bin(updvaliddata$V5,6110, 15945)  updvaliddata <- cbind(updvaliddata,V2bin1)  updvaliddata <- cbind(updvaliddata,V2bin2)  updvaliddata <- cbind(updvaliddata,V2bin3)  updvaliddata <- cbind(updvaliddata,V5bin1)  updvaliddata <- cbind(updvaliddata,V5bin2)  updvaliddata$V2bin1 <- as.factor(updvaliddata$V2bin1)  updvaliddata$V2bin2 <- as.factor(updvaliddata$V2bin2)  updvaliddata$V2bin3 <- as.factor(updvaliddata$V2bin3)  updvaliddata$V5bin1 <- as.factor(updvaliddata$V5bin1)  updvaliddata$V5bin2 <- as.factor(updvaliddata$V5bin2)  updvaliddata$V2 <- NULL  updvaliddata$V5 <- NULL  str(updvaliddata)    validWoeData <- predict(woemodel, updvaliddata, replace = TRUE)  pred.val <- predict(logit.glm.step, validWoeData, type = "response")  pred.val |

结果（前16条）：

|  |
| --- |
| 1         4         5         8        10        12  0.9913149  0.6774469  0.3637323  0.8274460 0.4732830  0.2124960 |

理论：

http://img.blog.csdn.net/20150722192029357?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQv/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center

R程序：

|  |
| --- |
| p.pred.val = exp(pred.val) / (1 + exp(pred.val))  p.pred.val |

结果（前16条）：

|  |
| --- |
| 1         4         5         8        10        12  0.7293476  0.6631686  0.5899436 0.6958146  0.6161605  0.5529250 |

**5、评分**

（1）获取WoE

R程序：

|  |
| --- |
| woemodel$woe |

结果：

|  |
| --- |
| $V1         A11        A12        A13        A14   0.8352181  0.4861704 -0.5888862 -1.1367785    $V3          A30         A31         A32         A33         A34   1.47707202  1.23412584  0.08050692  0.17968477 -0.85104637    $V4         A40        A41       A410        A42        A43        A44        A45   0.5062357 -1.0497671  0.4356181  0.0159684 -0.5373679  2.1095946  0.5897688         A46        A48        A49   0.5156609 -0.8861377  0.2034248    $V6         A61        A62        A63        A64        A65   0.2406038  0.1452224 -0.5134624 -1.1862423 -0.5600462    $V7          A71         A72         A73         A74         A75  -0.01066896  0.76497292  0.04475184 -0.46454320 -0.29932616    $V9          A91         A92         A93         A94   0.47198579  0.19445609 -0.17618339  0.08144633    $V10         A101        A102        A103  -0.03114749  0.54097866 -0.10337835    $V12          A121         A122         A123         A124  -0.611700848  0.126484147 -0.008165826  0.702680932    $V14        A141       A142       A143   0.5367143  0.4550362 -0.1358321    $V15        A151       A152       A153   0.3486068 -0.1831058  0.4869114    $V17         A171        A172        A173        A174   0.16368443  0.07667305 -0.06910192  0.14227034    $V19       A191      A192   0.104261 -0.164003    $V20         A201        A202   0.04051583 -1.57928487    $V2bin1           0          1   0.1793342 -1.2577013    $V2bin2            0           1   0.01757426 -0.01169407    $V2bin3           0          1  -0.2190058  0.6375334    $V5bin1           0          1   0.5926800 -0.1183796    $V5bin2           0          1  -0.1211926  0.6132993 |

（2）套用公式

http://img.blog.csdn.net/20150722192046308?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQv/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center

woe=ln(odds)，beta为回归系数，alpha为截距，n为变量个数，offset为偏移量（视风险偏好而定），比例因子factor。

http://img.blog.csdn.net/20150722192053883?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQv/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center

总评分。

比例因子和偏移量为相信是人为设定的，可根据实际情况而定。

因为变量较多，现以两个变量为例：

|  |
| --- |
| http://img.blog.csdn.net/20150722192224270?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQv/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center |

**六、模型预测**

从模型验证（Model validation）中抽取记录当作预测。