实行评分卡大纲

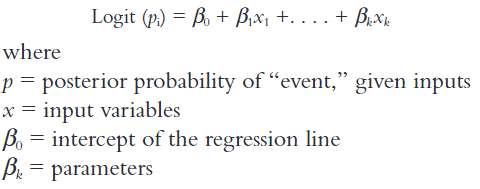
1. 数据分析
   1. 需要处理缺失value
      1. 如果有缺失value，你就有几个选择– 最好把缺失value集体在一起在一个attribute里
      2. 其他方法：
         1. 可以排除不完整的数据
            1. 可是很可能会把很多的数据都消除
         2. 可以impute缺失value
   2. Correlation
      1. **Example**: “PROC VARCLUS” on SAS “用一种主成分分析（PCA）去鉴定correlated characteristics. 你可以从每一组用一个或多个characteristic. 理论上,你选的characteristic会代表它的组里全部的information.” [Siddiqi 77]
         1. 比只用correlation figures好 – 因为correlation意外也会考虑到collinearity
   3. Multicollinearity 是不好的，可是不一定需要消除
      1. 大的样本量会减轻multicollinearity的影响
2. **Statistical Measures**
   1. 条款
      1. 特徵– 数据集里的一个value/变量/column
      2. Attribute – one WOE bin of a characteristic
   2. **WOE** - Generate WOE bins and values for **continuous** characteristics
      1. Distribution of Goods - % of Good Customers in a particular group
      2. Distribution of Bads - % of Bad Customers in a particular group
      3. WOE = In(% of non-events ➗ % of events)
      4. Separate bin for “Missing” groups
      5. General rule: “minimum 5% of data in each bucket”
      6. No groups without either good or bad values (divide by zero)
      7. Each bin has sufficiently different bad rate and WOE
      8. Bin WOE pattern should usually be monotonic (Or have **logical** relationships)
         1. Exceptions to be judged by person with business intelligence (exceptions: Siddiqi 84)
      9. **Nominal** (categorical/non-continuous) characteristics are grouped by combining attributes with similar WOE
   3. **IV** - Generate IV (Information Value) for characteristics
      1. IV = ∑ (% of non-events - % of events) \* WOE
      2. Indicates predictive power of characteristic for the data
      3. Usually use Medium/Strong variables

|  |  |
| --- | --- |
| **Information Value** | **变量预测能力** |
| 少于0.02 | 没有预测能力 |
| 0.02 到 0.1 | 弱的预测能力 |
| 0.1 到 0.3 | 中等预测能力 |
| 0.3 到0.5 | 强预测能力 |
| 大于0.5 | 可疑的预测能力 |

1. 初步概观
   1. 一般用8到15个characteristic
      1. Characteristics represent as many independent types of data as possible
   2. 评分卡的目的
      1. “模仿者一位经验丰富，聪明的评判员的思考过程”
      2. 给每一个subject发展一个“风险简介”
   3. 评分卡的随后的监测很重要
      1. 随着人口（人口统计，行为，文化）的变化，记分卡的准确性也会变
      2. 没月都应该测试评分卡的模型为了监控记分卡的准确性
2. 路基回归：用WOE values 做逻辑回归
   1. 线性回归使用在有连续变量的情况
      1. 线性回归不能预测概率
   2. 逻辑回归使用在目标变量是categorical（不是连续）的
      1. 可以预测输出属于哪个类别的概率
      2. 把输出限制到0和1之间
   3. 派生逻辑回归的函数(用一个输入变量, X)

|  |  |
| --- | --- |
| Logistic Function | Reframed Logistic Function |
|  |  |
| **Odds Ratio** defined as | Logit defined as |
|  |  |

* 1. Logit （用多个变量）



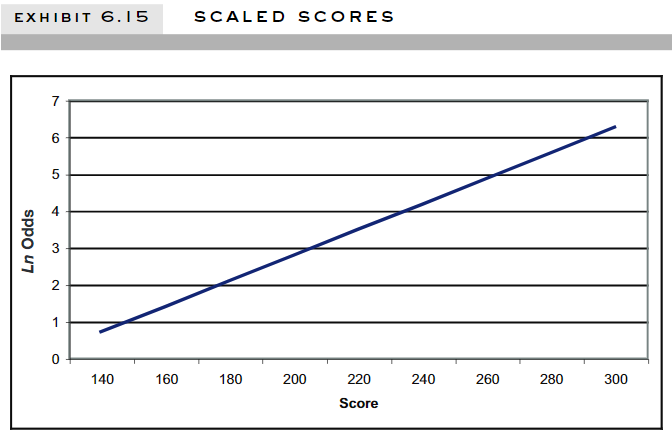
* 1. Use WOE of each grouping as the input
  2. 用回归找最好的模型(最佳的特徵组合)
     1. “所有可能”回归 – 详尽的搜索 – 用太多计算能力
     2. 有三种逻辑回归方法
        1. 预选
           1. 一个接一个地把最好的特徵加到模型里 (按预测能力，例如 p-value, Chi Square, etc.)直到达到某个阈值。

先测试每个变量，再把最好的特徵加到模型里。然后接着这样，每一次全测，再把最好的变量加进去。 选择最能改进模型的变量。

重复，直到达到阈值

* + - 1. 反向消除
         1. 这与预选相反
         2. 先把每一个特徵加在模型里。 按预测能力一个接一个把特徵去掉
      2. 步进式
         1. 预选和反向消除的组合
         2. 在这两方法之间交替
  1. Using regression to evaluate scorecards
     1. Model-ordering option in stepwise regression: two choices
        1. Single regression
        2. Multiple regression

1. **Reject Inference**
   1. 至今，我们的数据全部是从批准的贷款
   2. 也应该看被拒绝的情况
      1. 找这个信息有可能费劲
2. **设计评分卡[Siddiqi 92-119]**
   1. 标度改变评分
      1. Scores can be in many different forms
         1. Odds ratio
         2. “Defined min/max scale with a specified odds ratio at a certain point and specified rate of change of odds [Siddiqi 114]
      2. Scaling is purely cosmetic – does not affect performance
         1. Consider implementability with other tools, ease of understanding, and continuity with other scorecards
      3. A common scale used in industry
         1. “A scorecard with discrete scores scaled logarithmically, with the odds doubling at every 20 points”

****

* + 1. Scaling calculation
       1. Score = Offset + Factor\*ln(odds)

1. **模型评估**
   1. Cross-validation
   2. Confusion matrix
   3. ROC (Receiver Operating Characteristic) & AUC (Area Under Curve)
      1. ROC shows the classification power of the model compared to a purely random classifier
      2. The AUC is the integral of the ROC, numerically representing the models’ classification power
   4. Reject inference
      1. We do not have binary data on rejected applicants
      2. Seeks to augment data from rejected applicants