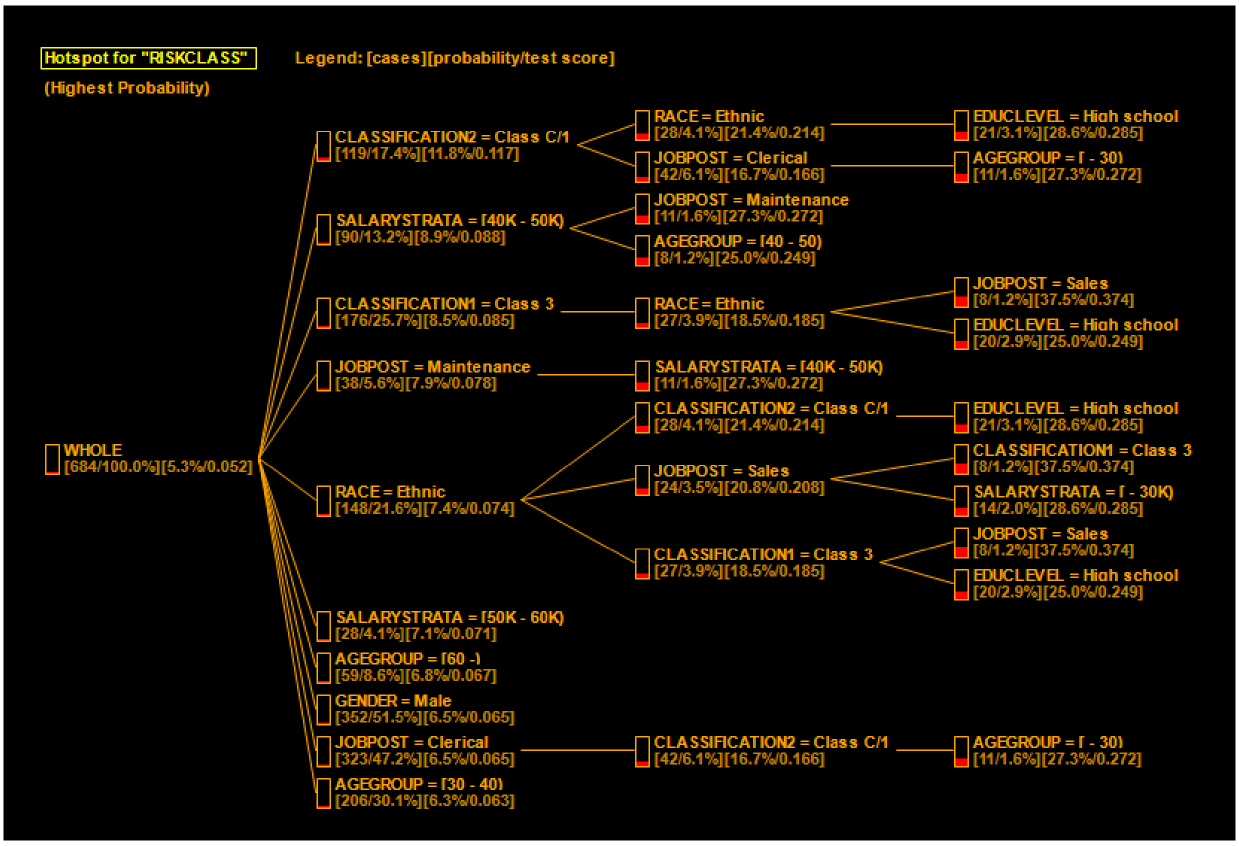
Variable Relevancy Analysis in Predictive Modeling

Variable Relevancy Analysis

The purpose of relevancy analysis is to find variables which contain *predictive information* to be used in predictive models. Low relevancy variables do not provide information that can be used in predictive models. Rather they can cause *overfitting*. Selection of modeling variables through relevancy analysis is very important to improve *predictive accuracy* and avoid overfitting. Most of predictive modeling work is spent on data preparation and variable relevancy analysis. The rest of modeling work is rather straightforward. Three relevancy analysis methods are described in the subsequent sections.

1. Hotspot Drill-down Profiling

CMSR Hotspot Drill-down tools can identify *profiles* of major risky customer segments. The following figure shows a hotspot drill-down example;

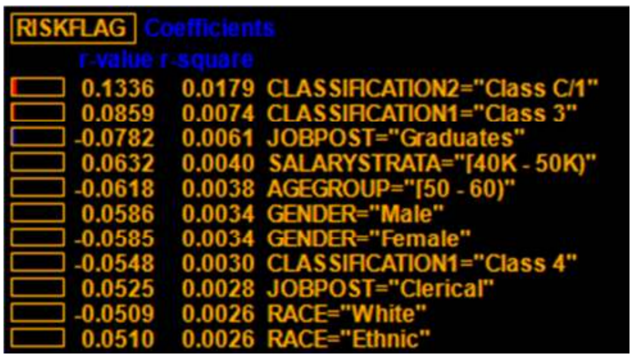


Hotspot for "RISKCLASS"
(Highest Probability)
Legend: [cases][probability/test score]
WHOLE [684/100.0%][5.3%/0.052]
CLASSIFICATION2 = Class C/1 [119/17.4%][11.8%/0.117]
SALARYSTRATA = [40K - 50K) [90/13.2%][8.9%/0.088]
CLASSIFICATION1 = Class 3 [176/25.7%][8.5%/0.085]
JOBPOST = Maintenance [38/5.6%][7.9%/0.078]
RACE = Ethnic [148/21.6%][7.4%/0.074]
SALARYSTRATA = [50K - 60K) [28/4.1%][7.1%/0.071]
AGEGROUP = [60 -) [59/8.6%][6.8%/0.067]
GENDER = Male [352/51.5%][6.5%/0.065]
JOBPOST = Clerical [323/47.2%][6.5%/0.065]
AGEGROUP = [30 - 40) [206/30.1%][6.3%/0.063]
RACE = Ethnic [28/4.1%][21.4%/0.214]
JOBPOST = Clerical [42/6.1%][16.7%/0.166]
JOBPOST = Maintenance [11/1.6%][27.3%/0.272]
AGEGROUP = [40 - 50) [8/1.2%][25.0%/0.249]
RACE = Ethnic [27/3.9%][18.5%/0.185]
SALARYSTRATA = [40K - 50K) [11/1.6%][27.3%/0.272]
CLASSIFICATION2 = Class C/1 [28/4.1%][21.4%/0.214]
JOBPOST = Sales [24/3.5%][20.8%/0.208]
CLASSIFICATION1 = Class 3 [27/3.9%][18.5%/0.185]
CLASSIFICATION2 = Class C/1 [42/6.1%][16.7%/0.166]
EDUCLEVEL = High school [21/3.1%][28.6%/0.285]
AGEGROUP = [- 30) [11/1.6%][27.3%/0.272]
JOBPOST = Sales [8/1.2%][37.5%/0.374]
EDUCLEVEL = High school [20/2.9%][25.0%/0.249]
EDUCLEVEL = High school [21/3.1%][28.6%/0.285]
CLASSIFICATION1 = Class 3 [8/1.2%][37.5%/0.374]
SALARYSTRATA = [- 30K) [14/2.0%][28.6%/0.285]
JOBPOST = Sales [8/1.2%][37.5%/0.374]
EDUCLEVEL = High school [20/2.9%][25.0%/0.249]
AGEGROUP = [- 30) [11/1.6%][27.3%/0.272]

Nodes of the figure show customer segments with highest ratios of "risk" occurrences. Red portions of boxes indicate ratios (or probabilities) of risk. *Variables and categorical items appearing in this output can be a good candidate for predictive modeling.*

2. Correlation Analysis

CMSR correlation analysis can be used to identify relevant variables that show significant (linear) correlation to risk. The following figure shows a CMSR correlation analysis output;

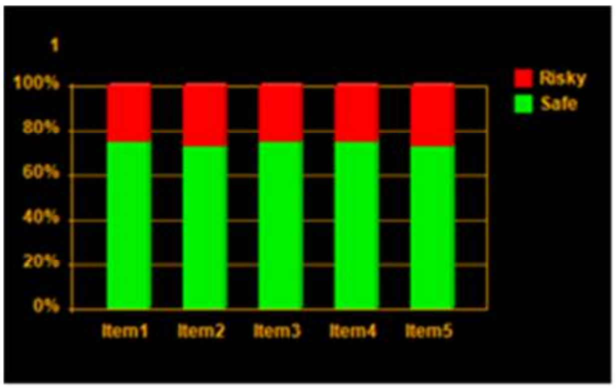


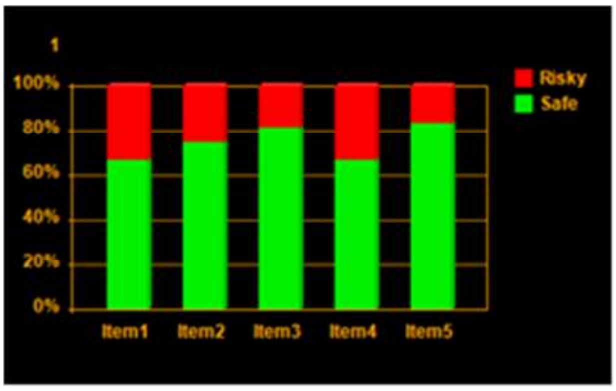
```
RISKFLAG Coefficients
      r-value r-square
      0.1336  0.0179 CLASSIFICATION2="Class C/1"
      0.0859  0.0074 CLASSIFICATION1="Class 3"
     -0.0782  0.0061 JOBPOST="Graduates"
      0.0632  0.0040 SALARYSTRATA="[40K - 50K)"
     -0.0618  0.0038 AGEGROUP="[50 - 60)"
      0.0586  0.0034 GENDER="Male"
     -0.0585  0.0034 GENDER="Female"
     -0.0548  0.0030 CLASSIFICATION1="Class 4"
      0.0525  0.0028 JOBPOST="Clerical"
     -0.0509  0.0026 RACE="White"
      0.0510  0.0026 RACE="Ethnic"
```

"r-value" is the coefficients for linear correlation. Having 0 means no correlation. 1.0 is the perfect *positive* linear correlation and -1.0 is the perfect *negative* linear correlation. "r-square" is the squared values of "r-value". *Variables with high absolute "r-value" are good candidates for predictive modeling.*

3. Categorical Bar/Histogram Charts

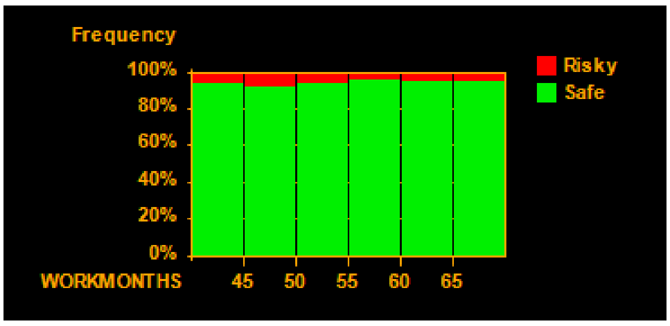
Relevant variables can be identified using CMSR hotspot drill-down and correlation analysis tools, as explained in the previous sections. In addition, CMSR categorical bar and histogram charts can be used to *identify* and to *verify* relevant variables. The following two bar charts show risk distribution by categorical items of two different categorical variables.





The left bar chart shows little variation in risk distribution amongst categorical items. This type of variables is not useful for predictive modeling. Note that predictive modeling relies on variations amongst different items. *The right bar chart, on the other hand, shows significant variations in risk distribution. This type of variables can be useful in predictive modeling and can be used in predictive models.*

Similar analysis can be performed on linear (= numerical) variables using categorical histogram charts. The following figure shows categorical distribution of risk over a numerical variable. Red parts show only some variations amongst different numerical value ranges. But close examination shows relative size differences of red proportions, which represent risk, are significantly large. So this variable *may* be used in predictive models.



Categorical bar and histogram charts can be used to *verify* suitability of variables identified using hotspot drill-down analysis and correlation analysis.

(It is important to note that your local anti-discrimination laws may bar to use certain demographic variables if your models are used in discriminatory purposes. Check your local anti-discrimination laws.)

Rosella

Rosella Software
www.roselladb.com