Paper 1601-2014

Graphs Useful For Variable Selection in Predictive Modeling

Robert Moore, Thrivent Financial, Minneapolis, MN

ABSTRACT

This paper illustrates some SAS® graphs that can be useful for variable selection in predictive modeling. Analysts are often confronted with hundreds of candidate variables available for use in predictive models, and this paper will illustrate some simple SAS graphs that are easy to create and are useful for visually evaluating candidate variables for inclusion or exclusion in predictive models. The graphs illustrated in this paper are bar charts with confidence intervals using the GCHART procedure and comparative histograms using the UNIVARIATE procedure. The graphs can be used for most combinations of categorical or continuous target variables with categorical or continuous input variables. This paper assumes the reader is familiar with the basic process of creating predictive models using multiple (linear or logistic) regression.

INTRODUCTION

Analysts developing predictive models often have hundreds of candidate variables for possible use in models. Regardless of the modeling method (regression, decision trees, neural networks, etc.) one step in the modeling process (after dealing with missing values and outliers) is to reduce the number of candidate variables for possible inclusion in the model. This variable reduction step requires understanding the association between the individual candidate variables and the target (or dependent) variable. The graphs discussed in this paper provide simple ways to visualize the associations between the individual candidate variables and the target variable. Both binary and continuous target variables will be considered with both categorical and continuous input variables. The variable names and values displayed in this paper were altered to protect company information.

When the predictive modeler is considering a large number of candidate variables for inclusion in a predictive model it is efficient to assign the list of categorical input variables to one macro variable, and the list of continuous (interval and ratio) variables to another macro variable. For simplicity, only short lists of variables will be illustrated in this paper, but the methods scale up for long lists of variables. By using ODS to direct the sequence of graphs to a PDF or HTML file, the graphs are captured for documentation and convenient viewing by the modeler.

BAR CHARTS WITH CONFIDENCE INTERVALS

When the target (dependent) variable is binary (taking just two possible values 0 or 1) some authors (see Rudd, 2001) recommend using PROC FREQ with the CHISQR option to evaluate the strength of association between the categorical input variables and the binary target variable. The candidate variables with the strongest association with the binary target as indicated by a Chi-square statistic above some threshold value are then selected for further investigation. This is an effective first step in reducing the number of candidate variables, and visualizing these associations can provide additional insight and reduce the time spent examining the resulting frequency tables.

CATEGORICAL INPUT VARIABLES

In cases where the target variable is binary and the input variable is categorical, bar charts with confidence intervals can be generated easily using PROC GCHART. The resulting bar charts will display a separate bar for each value of the categorical input variable with confidence intervals for the response proportion. These bar charts with confidence intervals can be used to visually judge whether the difference between the response proportions is statistically significant by examining the overlap between confidence intervals. In general, if the confidence intervals do not overlap the difference is judged significant, and if there is overlap, the difference is not judged significant. While this paper does not advocate this technique as a replacement for formal statistical significance testing, this technique can provide valuable insight into the strength of association between the input variable and the binary target variable. For a detailed discussion of the method of judging significant differences by examining overlapping confidence intervals see Schenker (2001).

Since horizontal bar charts can ease visual comparisons (see Friendly, 1991) those are illustrated here, but the same ideas apply to vertical bar charts. The resulting charts will display the input variable along the vertical y-axis, and the response along the horizontal x-axis. Additionally, the bars are displayed in descending order of the mean of the target variable for each category. This can help distinguish differences between the bars and overlap between

confidence intervals. Also, when ordinal input variables maintain their order in the display it can indicate a linear relationship with the target variable.

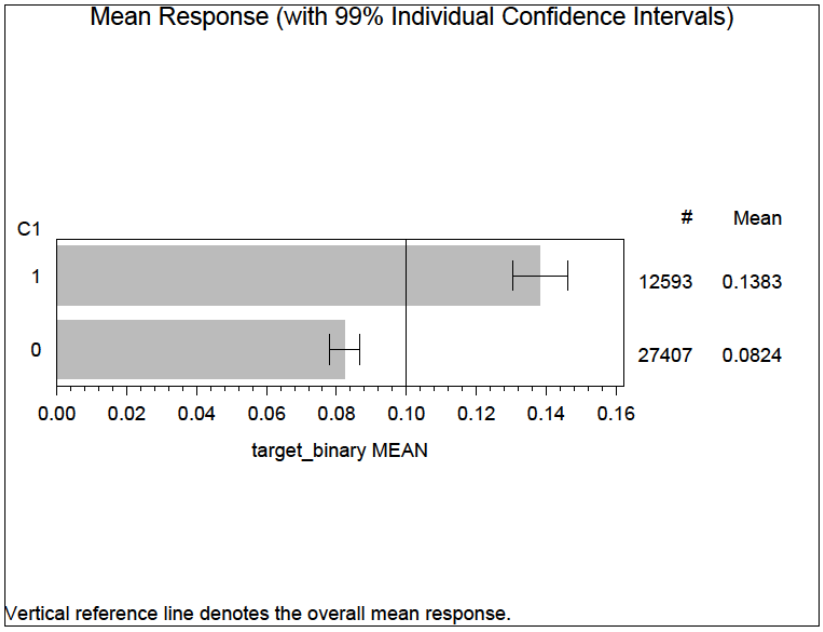

Figure 1 – Bar chart with confidence intervals for binary target and input variable (significant).

Figure 1 shows a horizontal bar chart with confidence intervals for a binary target with a categorical (binary) input variable called C1. The vertical reference line indicates the overall response rate of 10%, and the 99% confidence intervals on the bars definitely do not overlap each other or the reference line, so the difference would be judged highly significant, so this input would be selected as a potentially important candidate modeling variable. Typically, a wider separation between the non-overlapping confidence intervals indicates a stronger association between the input and the target variables.

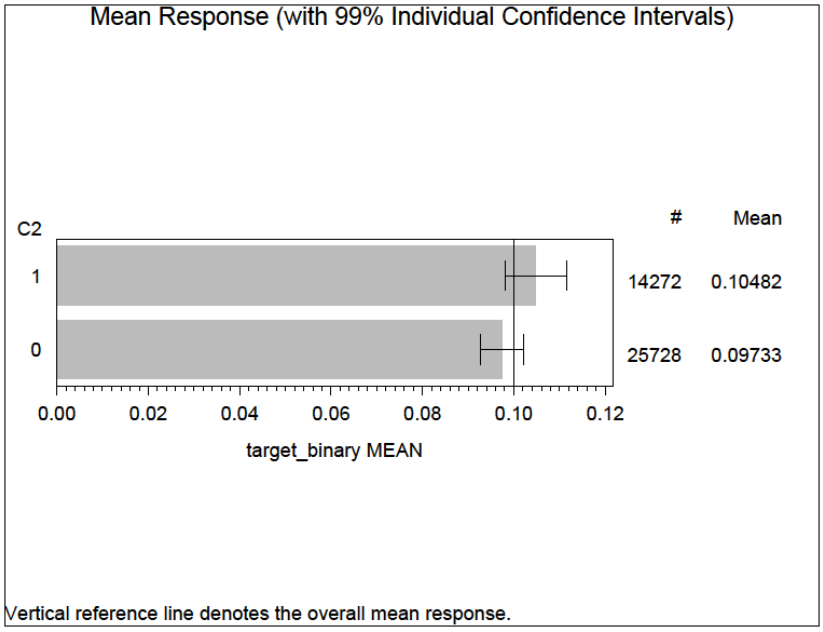

Figure 2 – Bar chart with confidence intervals for binary target and input variable (insignificant).

In figure 2, a similar graph for a different categorical (binary) input variable called C2 illustrates a case when the confidence intervals clearly overlap each other and the reference line, so this input would be rejected as a candidate modeling variable.

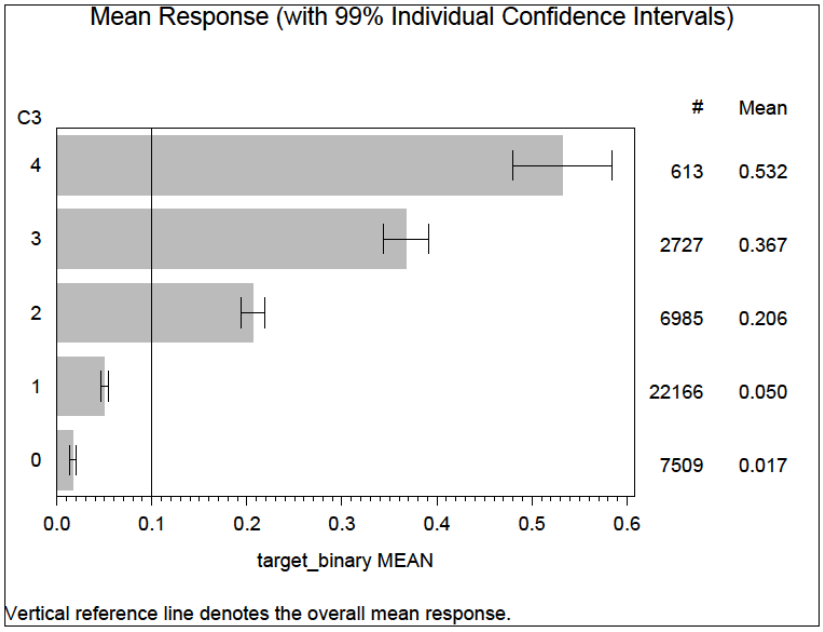

Figure 3 – Bar chart with confidence intervals for a binary target with ordinal input variable.

Figure 3 shows a similar chart with an ordinal input variable that takes the 5 consecutive values 0, 1, 2, 3, and 4. The Bonferroni principle (see Bickel, 1977) says that (1-α)100% individual confidence intervals for a categorical input with m distinct categories form approximate (1 – α·m)100% simultaneous confidence intervals. Since the individual confidence intervals in the chart are at the 99% level, they form approximate 95% simultaneous Bonferroni confidence intervals. Since the confidence intervals do not overlap, this is a potentially important candidate modeling variable. Also, since the order of the values along the vertical y-axis is maintained, it indicates a strong increasing relationship with the target variable.

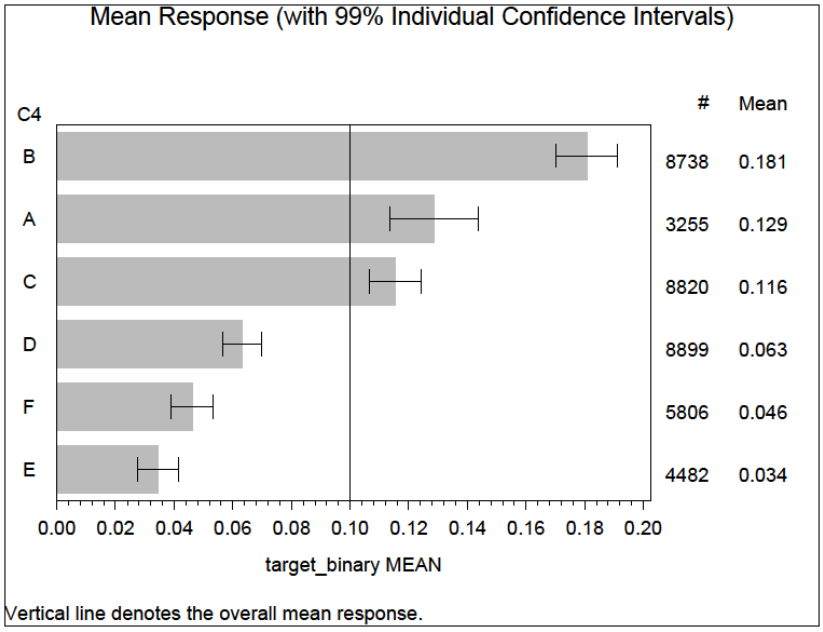

Figure 4 – Bar chart with confidence intervals for a binary target with nominal input variable.

Figure 4 shows a similar chart with a nominal input variable that takes 6 distinct values (where order has no meaning). Since the individual confidence intervals are at the 99% level, they form approximate 94% simultaneous Bonferroni confidence intervals. Judging from the confidence intervals, category B is significantly different from the other categories. Additionally, the overlapping confidence intervals for categories A and C suggest that they might be merged, and similarly categories E and F might also be merged. Again, this input would be selected as a potentially important candidate modeling variable.

The following code was used to generate the horizontal bar charts shown in figures 1 through 4. The data set 'moddata' contains the target variable and all the input variables. The first few lines are for convenience with line 1 assigning a list of 4 categorical variables named C1, C2, C3, and C4 to a macro variable named "catvars". The PROC SQL section on lines 3 through 5 captures the overall response rate in a macro variable named "rr_overall". Lines 7, 8, 23, and 24 are used to direct the charts to a PDF file using ODS output.

Lines 12 and 13 requests a horizontal bar chart for each of the categorical variables listed in the macro variable "catvars". A few of the many options available with PROC GCHART are used here after the slash on line 13 to make the charts more useful. The options on lines 13 and 14 specify the type of statistic will be the mean of the summary variable which is the binary target. For a binary target this mean will be the response percentage. The 'descending' option on line 17 is used to display the bars in descending order of the response rate. Line 19 displays a reference line at the overall response rate. Line 20 displays the confidence intervals with both upper and lower limits, and on line 21 the 'clm=' option specifies the confidence level of 99 percent.

```
0001 %let catvars = C1 C2 C3 C4;
0002
0003 proc sql noprint;
0004    select mean(target_binary) into: rr_overall from moddata;
0005 quit;
0006
0007 ods listing close;
0008 ods pdf file="C:\barchartCI.pdf";
0009 title j=c "Response Rates (with 99% Individual Confidence Intervals)";
0010 footnote j=l "Vertical line denotes the overall response rate.";
0011 pattern1 v=solid c=graybb;
0012 proc gchart data=moddata;
0013    hbar &catvars / type=mean
0014                    summvar= target_binary
0015                    freqlabel='#'
0016                    meanlabel='Mean'
0017                    descending
0018                    discrete
0019                    ref=&rr_overall
0020                    errorbar=both
0021                    clm=99;
0022 run;
0023 ods pdf close;
0024 ods listing;
```

The bar charts with confidence intervals illustrated here have been for a binary target variable, and they can also be used when the target variable is continuous. The code is almost the same with the type of statistic remaining the mean but the summary variable being the continuous target variable rather than the binary target variable. The interpretation of the bar charts with confidence intervals is similar.

CONTINUOUS INPUT VARIABLES

When the input variables are continuous, similar bar charts can be generated by partitioning the range of the variables into disjoint intervals. PROC GCHART will do this automatically and the code is almost the same as for the categorical variables. The macro variable "convars" on line 2 of the code fragment below is now a list of continuous variables. The 'descending' and 'discrete' options are replaced by the 'levels=' option on line 6. The 'levels=' option shown on line 6 is used to specify the input variable range is divided into 16 intervals of equal width. Without the 'levels=' option, the charts are generated with the default number of intervals for each variable. Details about how the PROC GCHART determines the default intervals widths can be found in the procedure documentation. The use of 16 intervals here is subjective, and it is usually necessary to experiment with a few different numbers to find which one provides the most understandable charts. If the number of bars is too large, the charts become difficult to read.

```
0001 proc gchart data=moddata;
0002    hbar &convars / type=mean
0003                    summvar= target_binary
0004                    freqlabel='#'
0005                    meanlabel='Mean'
0006                    levels=16
0007                    ref=&rr_overall
0008                    errorbar=both
0009                    clm=99;
0010 run;
```

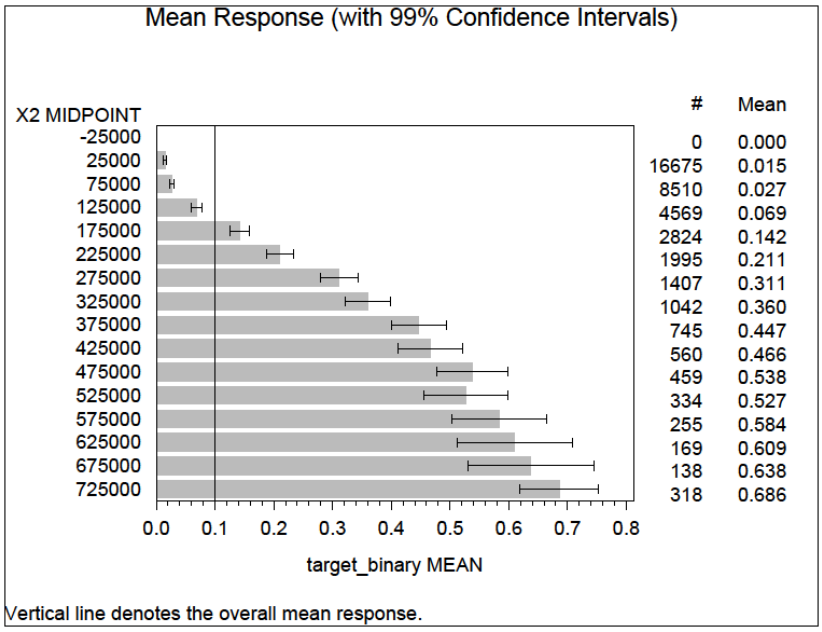

Figure 5 – Bar chart with confidence intervals for a binary target with continuous input variable.

Figure 5 shows a bar chart with a continuous input variable named X2 using the specified 16 intervals. Since the individual confidence intervals are at the 99% level, they form approximate 84% simultaneous Bonferroni confidence intervals. As with most statistics when applied to large data sets, care must be taken when interpreting the confidence intervals in these charts. As illustrated in this chart, the confidence intervals are narrow for intervals with a large number of observations, and wider for intervals with a smaller number of observations. Even though the confidence intervals for the larger values of this input variable overlap, the chart indicates a strong increasing relationship with the target variable.

Figure 6 shows below a similar chart with a continuous input variable named X3 again using the specified 16 intervals. This chart illustrates extremely wide confidence intervals due very small numbers of observations in the intervals at the large values of the input variable. When there are an adequate number of observations in each interval and separation between the confidence intervals, such a graph suggests a non-linear relationship and may provide guidance with respect to possible transformations to make the relationship more linear.

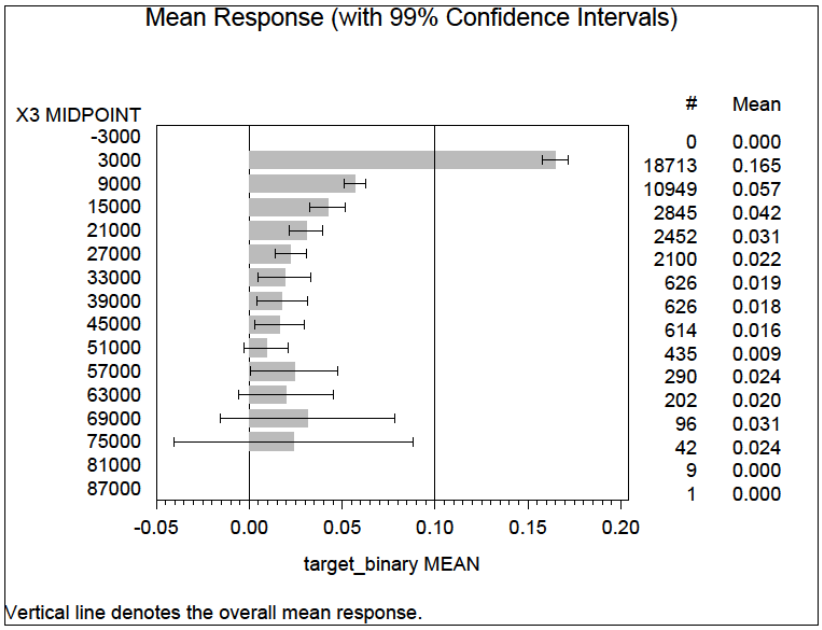

Figure 6 – Bar chart with confidence intervals for a binary target with continuous input variable.

COMPARATIVE HISTOGRAMS

In cases where the target variable is categorical and the input variable is continuous, comparative histograms can be generated using PROC UNIVARIATE. The charts will display histograms for the continuous input variable separately for each value of the categorical target. The histograms can easily be compared to visually judge whether the distribution of the continuous variable is different between the responders and non-responders.

The following code was used to generate the comparative histograms shown in figures 7 and 8. The first line is for convenience and assigns a list of 4 continuous variables named X1, X2, X3, and X4 to a macro variable named "convars". Lines 3, 4, 15, and 16 are used to direct the charts to a PDF file using ODS output. The 'noprint' option was used on line 7 to suppress the usual PROC UNIVARIATE printed output. On line 9, the 'class' statement is used to identify the binary target, and the histograms are requested on line 10. The 'inset' statement is used on line 11 to specify some statistics to be displayed in the histograms. The options following the slash on line 11 are used to specify where the inset box with the statistics is displayed in the histogram.

The number of bars displayed in the histogram and the points displayed along the horizontal axis were determined by the procedure default, but can be specified by using either the 'nmidpoints=' or 'midpoints=' options with the histogram statement on line 10. The 'nmidpoints=' option can be used to specify the number of intervals displayed in the histograms, and the 'midpoints=' options allows specification of a list of midpoints for the intervals.

```
0001 %let convars = X1 X2 X3 X4;
0002
0003 ods listing close;
0004 ods pdf file="C:\comphistograms.pdf";
0005 title j=c 'Comparative Histograms by Response Category';
0006 pattern1 v=solid color=grayee;
0007 proc univariate data=moddata noprint;
0008    var &convars;
0009    class target_binary;
0010    histogram &convars;
0011    inset mean std skewness kurtosis / height=1.5 pos=NE;
0012 run;
0013 title;
0014 quit;
0015 ods pdf close;
0016 ods listing;
```

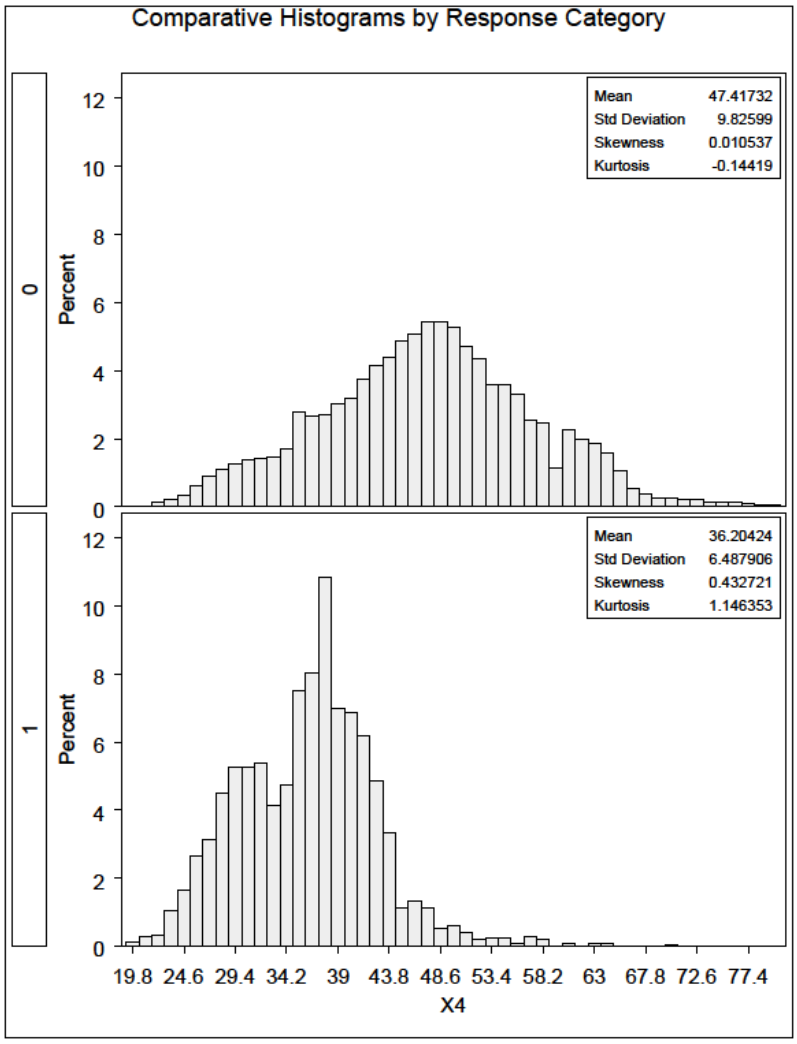

Figure 7 – Comparative histogram for a binary target with continuous input variable.

Figure 7 shows a comparative histogram for a binary target with a continuous input variable named X4. As indicated on the left vertical, the upper histogram is for the non-responders with the binary target variable equal to 0, and the lower histogram is for the responders with the binary target variable equal to 1. In this case the variable actually represents age, and the chart suggests that the younger ages are much more responsive than the older ages. This input would be selected as a potentially important candidate modeling variable.

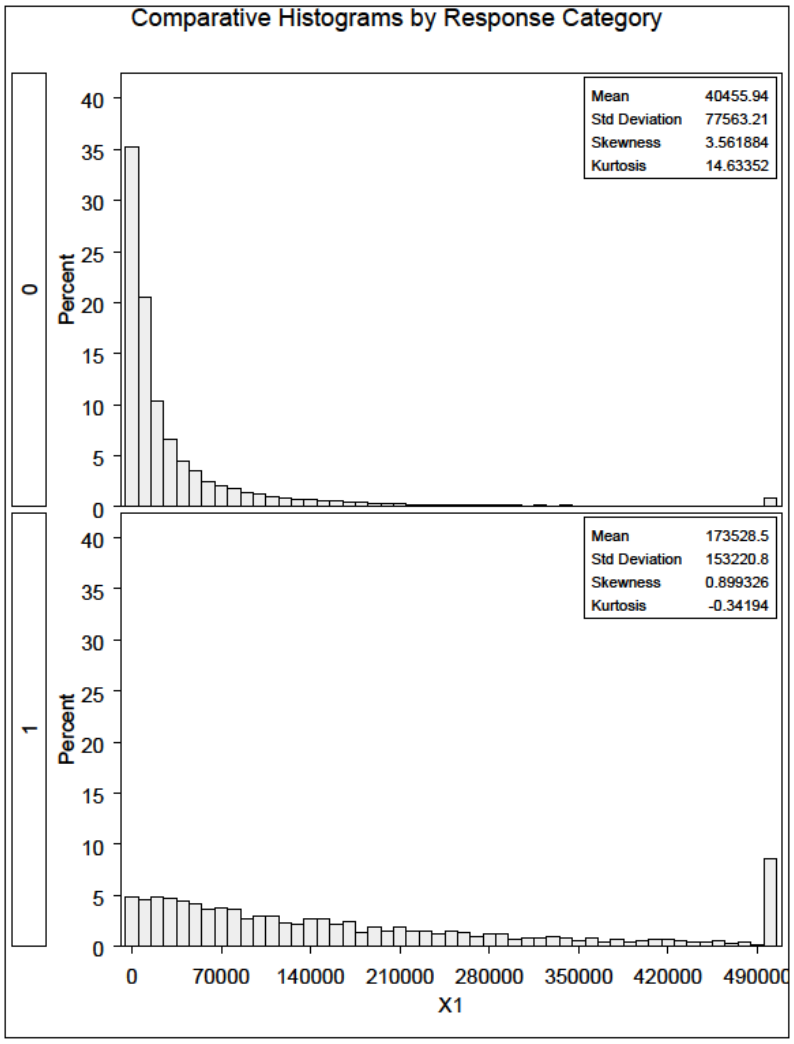

Figure 8 – Comparative histogram for a binary target with continuous input variable.

Figure 8 shows a comparative histogram for a binary target with a continuous input variable named X1. In the upper frame, the distribution of this variable for non-responders is skewed with a long tail at the higher values, and in the lower frame, the distribution for responders is much less skewed. In this case the variable actually represents amounts of money invested in a particular product, and the chart suggests that the responders tend to have larger amounts of money invested in the product than non-responders. Again, this input would be selected as a potentially important candidate modeling variable.

The comparative histograms illustrated here were for a binary target variable and continuous input variable, with the binary variable being used as the class variable to create the charts. The same code can be used for categorical target variables with more than 2 categories, but the charts become more cluttered the larger the number of categories. When the target variable is continuous, there is no corresponding display, but similar charts can be used to compare the continuous target variable distributions across values of categorical input variables. Unfortunately, the displays become less useful with more distinct values assumed by the categorical variables.

OTHER APPLICATIONS

The graphs illustrated here can be useful not just for variable selection, but for any exploratory data analysis as a way to visually examine a large number of variables for any association with a collection of other variables. For example, to explore which variables might be drivers of profit, lists of candidate driver variables can be compared to variables that measure profit considered as target variables.

SAS Enterprise Miner contains many variable selection features for use with predictive modeling, and the graphs illustrated here can be produced in SAS Enterprise Miner using a code node. Enterprise Miner automatically creates macro variables for the binary, nominal, ordinal, and continuous input variables which can be used in the code fragments above.

CONCLUSION

The graphical methods illustrated here provide a convenient way for analysts developing predictive models to gain visual insights into the relationships between candidate input variables and the target variable. This can assist the modeler in reducing the number of variables for possible inclusion in the model. The graphs can be used to understand the association between binary or continuous input variables with either binary or continuous target variables. By assigning long lists of variable names to macro variables, and directing the output to PDF or HTML files using ODS, the charts can be generated for a large number of variables with few lines of code.

REFERENCES

Bessler, L. (2003), "Easy, Elegant, and Effective SAS® Graphs", SUGI 28, Paper 68-28

Bickel, P. and Doksum, K. (1977) “Mathematical Statistics”, San Francisco, CA: Holden-Day

Friendly, M. (1991), “SAS System for Statistical Graphics”, First Edition, Cary, NC: SAS Institute Inc.

Rud, O. (2001), “Data Mining Cookbook: Modeling Data for Marketing, Risk and Customer Relationship Management”, New York, NY: John Wiley & Sons, Inc.

Schenker, N. and Gentleman, J. (2001), “On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals”, The American Statistician, Vol. 55, No. 3, p. 182-186

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Robert Moore
Thrivent Financial
625 Fourth Avenue South
Minneapolis, MN 55415
Work Phone: 612-844-4036
E-mail: robert.moore@thrivent.com