

Predictive Modeling Using SAS Visual Statistics: Beyond the Prediction

Xiangxiang Meng, Wayne Thompson, and Jennifer Ames, SAS Institute Inc.

ABSTRACT

Predictions, including regressions and classifications, are the predominant focus of many statistical and machine-learning models. However, in the era of Big Data, a predictive modeling process contains more than just making the final predictions. For instance, a large collection of data often represents a set of small, heterogeneous populations. Identification of these sub groups is therefore an important step in predictive modeling. Additionally, Big Data data sets are often complex, exhibiting high dimensionality. Consequently, variable selection, transformation, and outlier detection are integral steps. This paper provides working examples of these critical stages using SAS® Visual Statistics, including data segmentation (supervised and unsupervised), variable transformation, outlier detection, and filtering, in addition to building the final predictive model using methodology such as linear regressions, decision trees and logistic regressions. The illustration data were collected over 2010 to 2014 from vehicle emission testing results.

INTRODUCTION

Predictive modeling leverages statistical and machine learning models to predict or classify outcomes of interest. The concept of a predictive modeling pipeline is constantly evolving, driven by the generation of new analytic algorithms, new data sources, or both. In the era of Big Data, data sets that are derived from a distributed storage system, such as the Hadoop Distributed File System, not only have greater volume, but also higher complexity. A large collection of data often represents a set of small, heterogeneous populations, which results in the high dimensionality that is commonly observed.

This paper develops examples using SAS® Visual Statistics to illustrate some critical stages in a predictive modeling pipeline for handling complex data. Techniques used include supervised and unsupervised data segmentation; supervised variable transformation, such as grouping of high cardinality variables using decision trees; stratified (group by) modeling; outlier detection; interactive feature creation and data filtering; and post-model visualization.

SAS® Visual Statistics is an add-on modeling tool for SAS® Visual Analytics. More specifically, the modeling capabilities of SAS® Visual Statistics 7.2 are included within the user interface of SAS® Visual Analytics Explorer 7.2. Figure 1 highlights the icons in the user interface of SAS® Visual Analytics Explorer for the SAS® Visual Statistics add-on. This includes, from left to right, linear regression, logistic regression, generalized linear models, clustering, and model comparison. SAS® Visual Statistics also adds interactive decision tree capabilities (i.e., prune, split, and train), model assessment, and model score code generation to the core Decision Tree visualization of SAS® Visual Analytics Explorer. All SAS® Visual Statistics models are based on the in-memory LASR Analytic Server and can be scaled up to meet your Big Data computational needs on a distributed grid. All SAS® Visual Statistics models can be exported via score code, derived columns for predictive values, and PDF output. The examples and figures in this paper are created in SAS® Visual Statistics 7.2 and SAS® Visual Analytics Explorer 7.2.

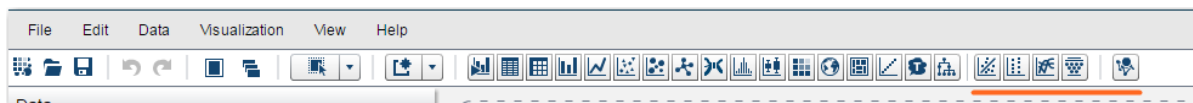


Figure 1. Add-ons to SAS® Visual Analytics Explorer 7.2 from SAS® Visual Statistics 7.2.

DATA DESCRIPTION

The examples in this paper are derived using a car emission data set that contains the emission and fuel economy testing results published by the United States Environmental Protection Agency (EPA). The data, which is aggregated and simplified, is based on the original data source available from the EPA website. The final SAS® data set used in this paper contains testing results from 2010 to 2014 on 16,468

vehicles from 55 vehicle makes of both cars and trucks. There are 26 columns in the SAS® data set related to vehicle information, testing procedures and testing results. The columns used through this paper are listed in Table 1.

Table 1. Variables Used in the Examples.

Variable Type	Variable Labels	Cardinality
Categorical	Vehicle Cylinders	9
	Vehicle Make	55
	Vehicle Type	2
	Vehicle Year	5
Continuous	Emission of Total Hydrocarbons (g/mi)	
	Vehicle MPG	
	Vehicle Weight	

Figure 2 provides a visualization of the relationship between Vehicle MPG and Vehicle Weight using the scatter plot in SAS® Visual Analytics Explorer. A linear curve is fit to predicted MPG using the weight of the vehicles.

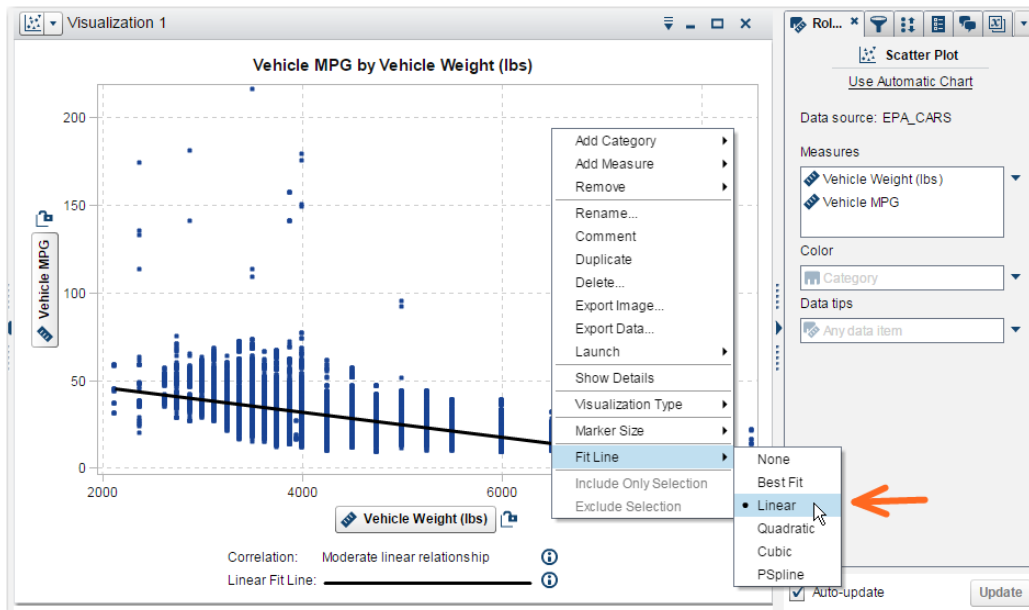


Figure 2. Linear fit of Vehicle MPG and Vehicle Weight.

STRATIFIED MODELING

LINEAR REGRESSIONS ON MULTIPLE SEGMENTS

Large data sets usually consist of a set of heterogeneous sub-populations. Many advanced statistical models such as random- and mixed-effect models have been proposed to account for variance heterogeneity. Instead, a simple but practical approach is to divide and conquer. This process entails dividing the entire data set into different segments and then building separate models for each segment. Stratified modeling is often driven by business needs to build models and make decisions separately for each customer or product subgroup. Common segmentations include age, income group, geographical region, product line, customer loyalty class, etc.

In SAS® Visual Statistics, you can identify potential heterogeneity in the data using a variety of visualization techniques, including scatter plots, heat maps, box plots, and geo maps. After you identify different segments based on one or more categorical effects, you can use the identified variables as a group-by in SAS® Visual Statistics to develop separate predictive models. For example, Figure 3 shows

the scatter plot of Vehicle MPG and Vehicle Weight, colored by Vehicle Type. It is clear that different distribution patterns exist in the CAR segment and the TRUCK segment of the data.

You can right click on the visualization to launch a linear regression model (**Launch > Linear Regression**). The color variable (Vehicle Type) in the scatter plot visualization is automatically transferred to the **Group By** role in the linear regression visualization, which builds two separate linear regressions for cars and trucks. The vertical-axis variable (Vehicle MPG) is used as the response and the horizontal axis variable (Vehicle Weight) is a predictor (Figure 4).

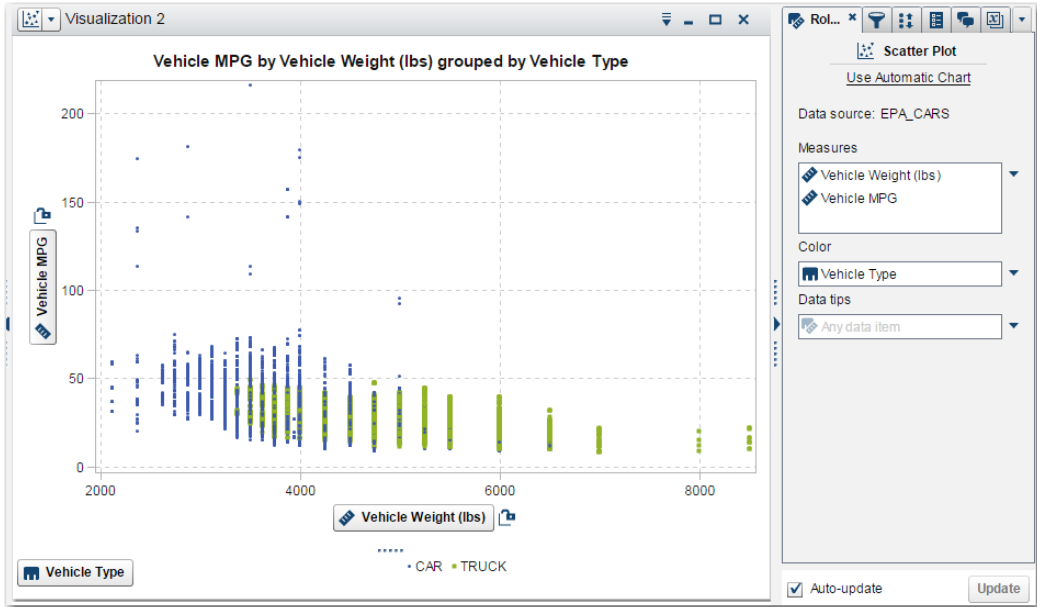


Figure 3. Scatter plot of Vehicle MPG and Weight colored by Vehicle Type.

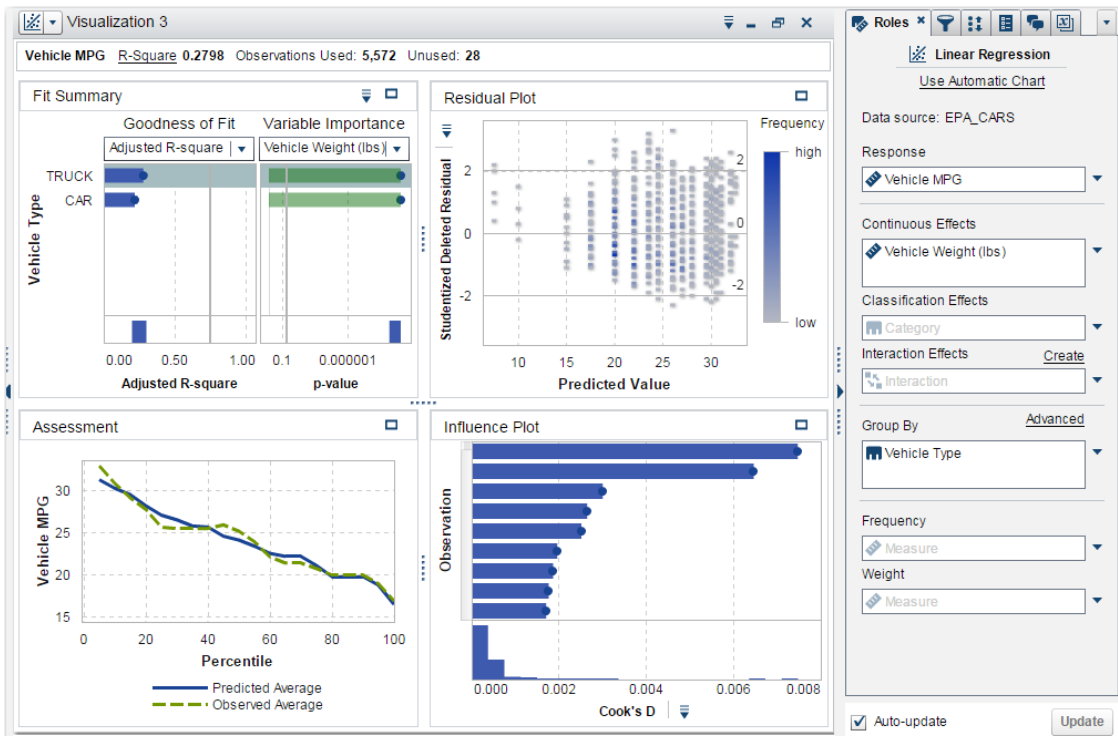


Figure 4. Group-By linear regression launched from the scatter plot displayed in Figure 3.

For each segment defined by the **Group By** variables, SAS® Visual Statistics includes a variable importance chart that contains the negative log of p-values from Type III tests, a residual plot, an influence plot with the top 2000 observations ranked by an influence statistics (Cook's D is the default for linear regression), and the model assessment plot that contains the predicted average versus observed average for binned predictions. The residual plot is converted from a scatter plot into a heat map when the data has a large number of observations (Figure 4). You can click on the bars in either of the two top left bar charts to select different segments and display the results from the different models. Summary tables such as ANOVA, Type III tests, and parameter estimates are available by clicking the hidden **Show Details** button (Figure 5). In the example, two simple regression models were fit, individually, for cars and trucks as follows:

$$\text{Vehicle MPG} = 48.68152 - 0.00482 * \text{Vehicle Weight, if Vehicle Type} = \text{'Truck'}$$

$$\text{Vehicle MPG} = 67.84679 - 0.00902 * \text{Vehicle Weight, if Vehicle Type} = \text{'Car'}$$

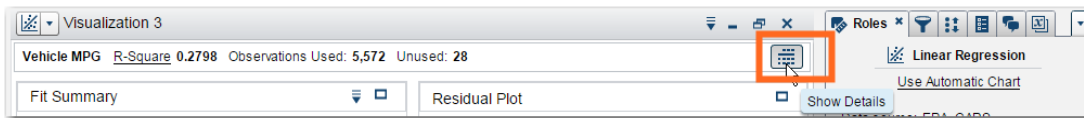


Figure 5. Show/Hide Details button for toggling the display of summary tables.

ADVANCED GROUPBY FEATURES

A common data wrangling strategy for stratified modeling is the selection or aggregation of the segments for a high cardinality variable. For example, you might want to exclude segments with few observations or manually rank the segments based on some business factors such as the overall revenue for each customer sub group. You can then filter down to the segments with revenue above a specific threshold. For the car emission example, consider building separate models for each vehicle make (55 makes in total), but include only those makes that have enough observations during the time period (2010 to 2014) with relatively low total hydrocarbon emissions. This is accomplished in SAS® Visual Statistics by using the *Rank* tab and the *Advanced Group By* dialogue.

In Figure 6, Vehicle Make is used to rank the makes by frequency, and only the top 25 makes are selected. In Figure 7, the Advanced Group By dialogue is used to further filter down the number of segments to 10, which corresponds to the 10 vehicle makes with the lowest average emission of total hydrocarbon out of the top 25 by frequency. Click **OK** to exit. The 10 linear regression models are trained simultaneously.

The Fit Summary plot shows that Vehicle Make is a significant predictor for all models except for the BMW Mini (Figure 8). You can quickly generate a scatter plot that shows Vehicle Weight and Vehicle MPG have no relationship in this segment (Figure 9). Note that this example applies a filter to the data in order to compare the segments with largest and smallest adjusted R-square values (Nissan versus MINI). The scatter plot indicates that the MINI and Nissan segments approximately follow the same linear relationship between Vehicle MPG and Vehicle Weight. However, the variance of the observations in the MINI segment is much larger than the variance of the observations in the Nissan segment. This actually validates the use of stratified linear regression because heterogeneous sub groups do exist.

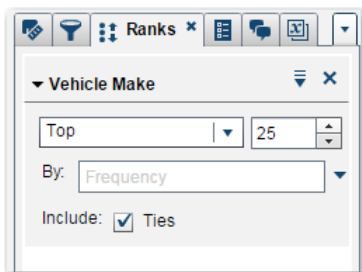


Figure 6. Using Rank to include the top 25 makes by frequency.

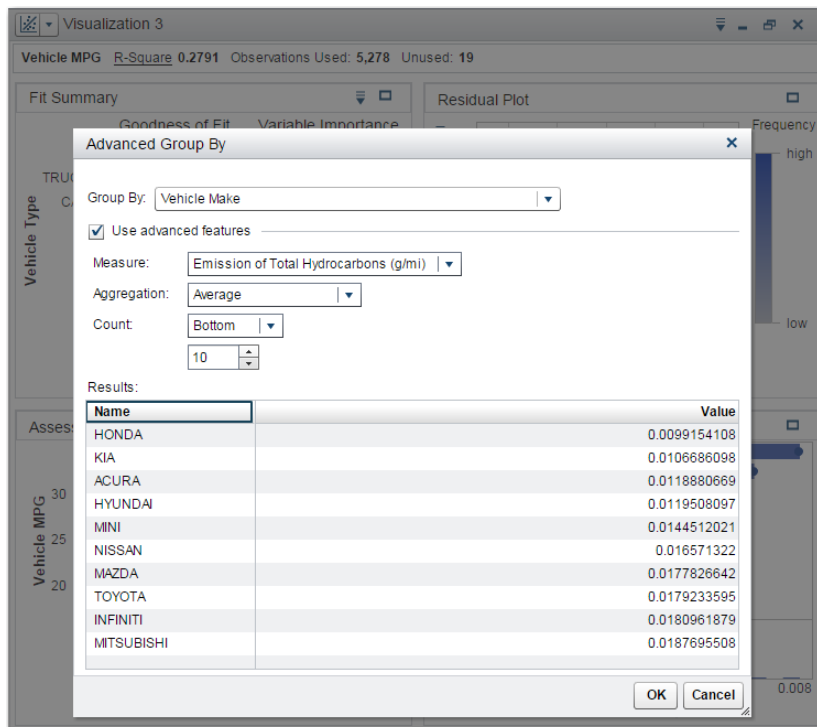


Figure 7. Using Advanced Group By dialogue to select the 10 makes with lowest average emission of total hydrocarbon (gram/mile).

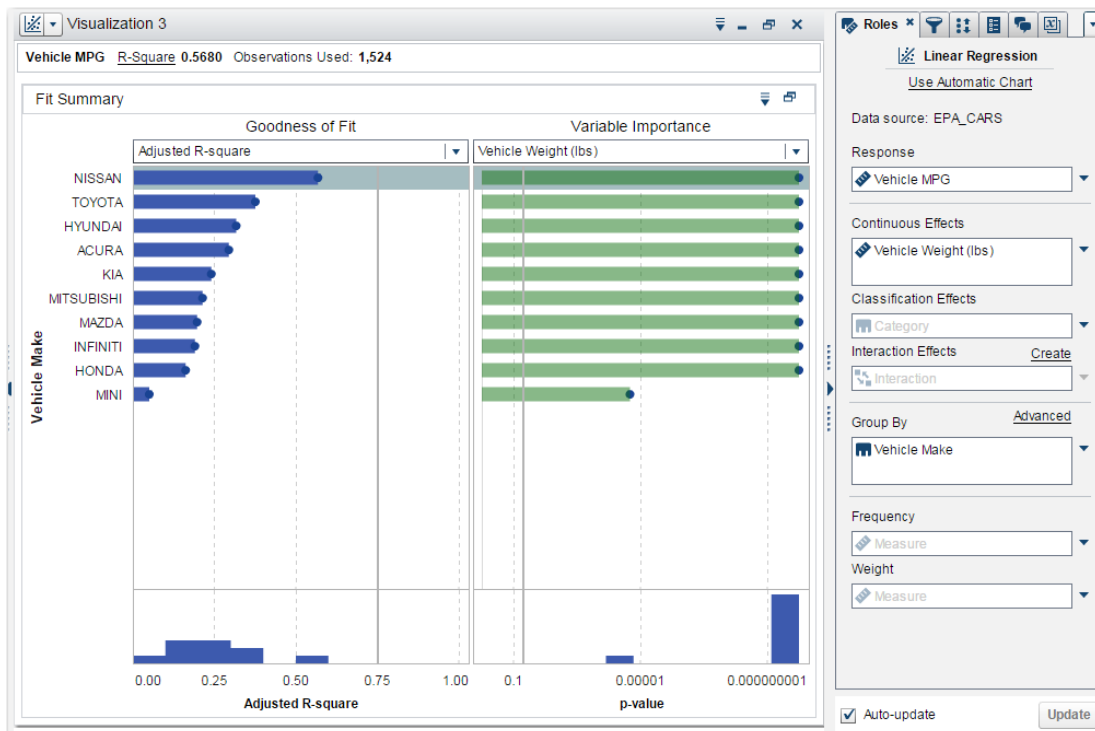


Figure 8. The Fit Summary tab shows the importance of the predictors and goodness of fit statistics across different segments.

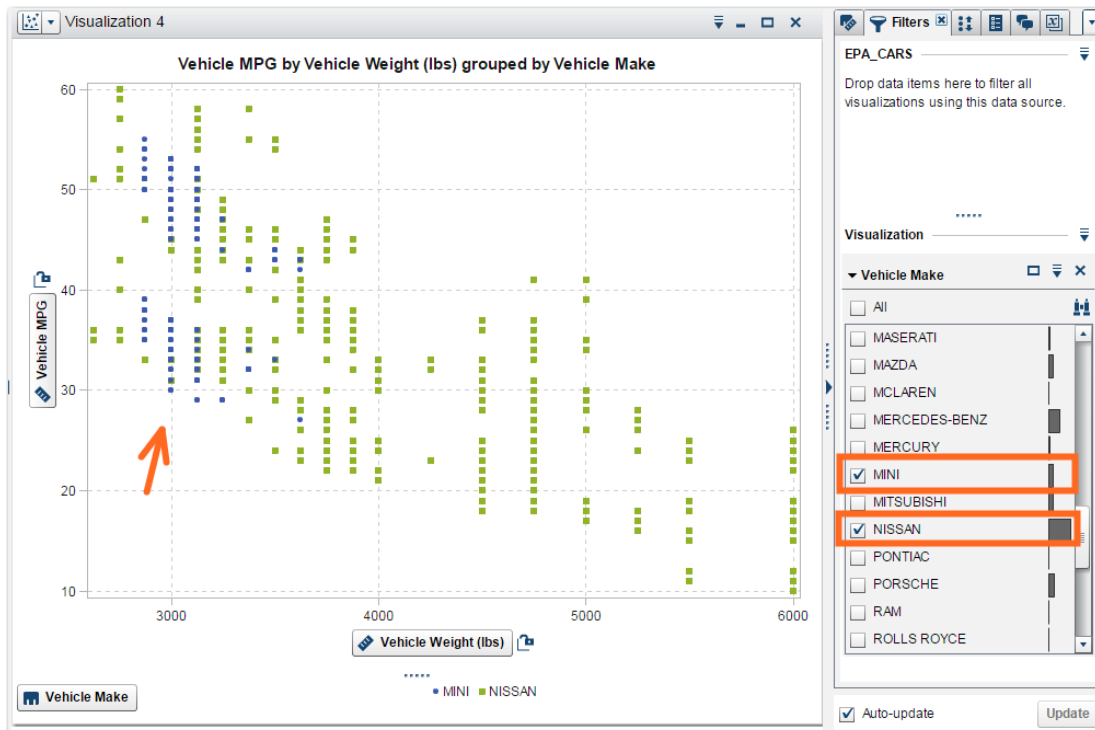


Figure 9. Use a scatter plot to confirm the finding in stratified linear regressions. Nissan and MINI vehicles follow the same trend but MINI segment has larger variance and thus less significant p-value.

DERIVING SEGMENTS FROM DECISION TREES OR CLUSTERING

In many applications, a pre-defined segmentation variable does not exist, and finding a reasonable segmentation of the data is an integral part of the predictive modeling process. Segmentation can be derived in SAS® Visual Statistics on the fly. Decision trees and k-means clustering models are available in SAS® Visual Statistics for building supervised or unsupervised segmentation. These are derived as terminal node (leaf) assignments in a decision tree or as a cluster membership assignment in k-means clustering. Following the car emission example in the previous section, rather than building one linear regression model for each vehicle make, you might prefer to build linear regression models stratified by the level of total hydrocarbon emissions (for instance, low, medium, and high), each containing vehicles from several makes. In predictive modeling, this dimension reduction step is often referred to as supervised binning or grouping.

The decision tree model is a popular tool used to derive supervised grouping of a high-cardinality variable. In this example, first you derive a calculated item that is a binary indicator of high total hydrocarbon emission (Figure 10). The cutoff value of 0.03 is chosen from the third quartile of the observed total hydrocarbon emissions. A calculated item in SAS® Visual Analytics is a temporary variable that is attached to the original data source without generating a new data copy in memory. When this temporary variable is involved in any visualization or model, it is computed on the fly.

Next, a decision tree model is built with the derived variable THC Indicator as the response, and Vehicle Make as a predictor. To produce a small tree with large leaves (segments), specify the following properties values: **Maximum Branches** to 4, **Maximum Levels** to 2, and **Leaf Size** to 1000. Decision tree models are widely used in many other applications, and each application requires different parameter settings. For example, when performing anomaly detection, it is advisable to use a deeper binary tree with small leaf size.

The decision tree model generates four leaves, where each leaf contains a subgroup of Vehicle Make based on the levels of total hydrocarbon emissions. Right click on the tree visualization and select **Derive**

a Leaf ID Variable. Name it *Segments of Make by THC*. The new variable is available in the Data Pane, and can be used in any other visualizations or models.

In this example, you open the previous linear regression visualization and replace the previous **Group By** variable (Vehicle Make) with the new leaf ID variable (Figure 12). Note that the previous Rank is no longer needed (Figure 6), so you can delete it from the **Ranks** tab.

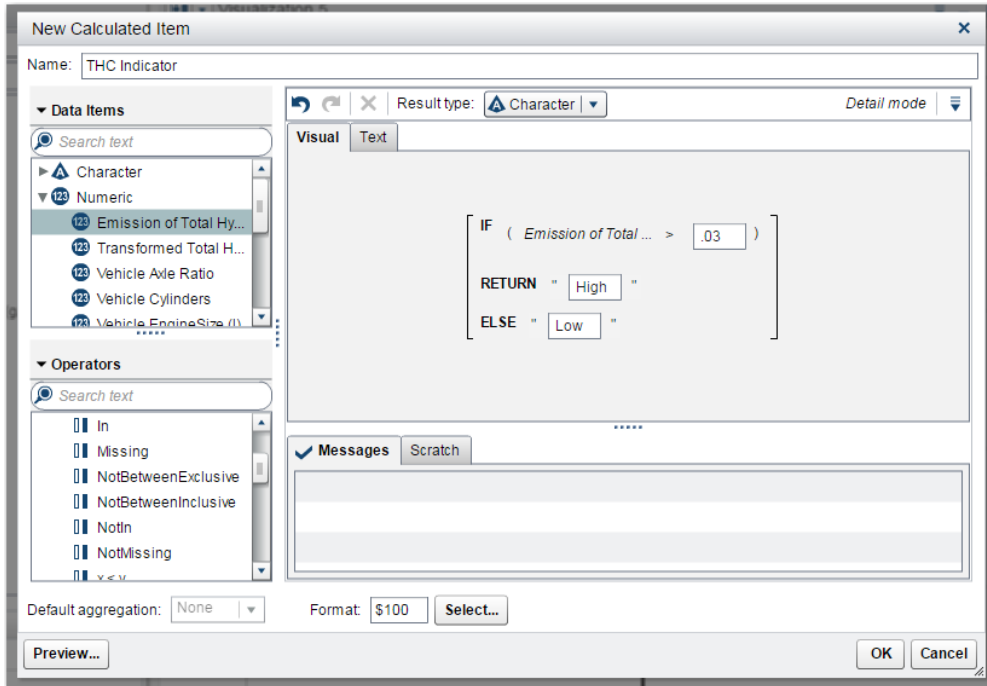


Figure 10. Derive a binary total hydrocarbon indicator based on the third sample quartile.

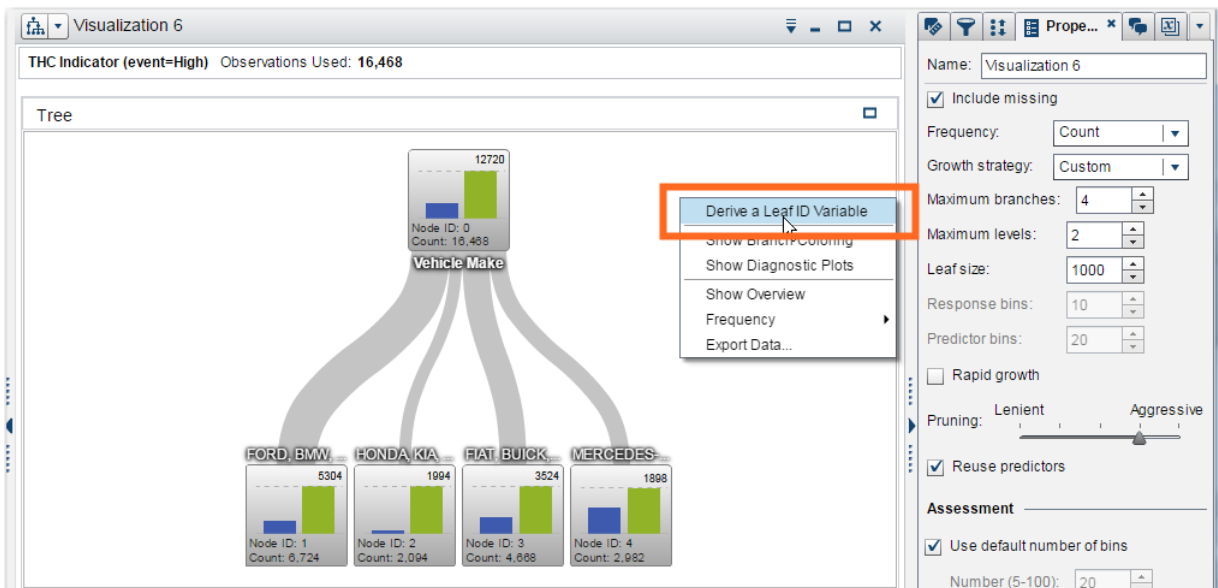


Figure 11. Derive a grouping of the Vehicle Makes using decision tree model. Figure 10 shows the derivation of the THC Indicator. In this example, we aggregate 55 makes into 4 different groups to represent different levels of total hydrocarbon emission.

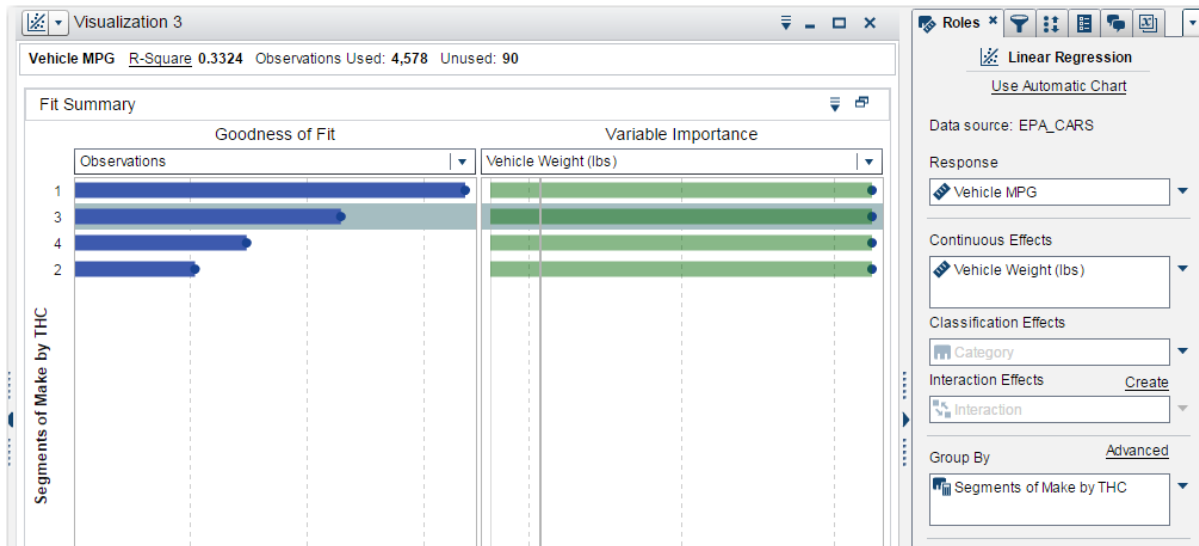


Figure 12. The fit summary output after changing the Group By variable to the new THC Indicator derived from decision tree. The four levels represents four different leaves from the decision tree models in Figure 11. The segment indices (1, 2, 3, and 4) in Figure 12 correspond to the Node ID indices in Figure 11.

OUTLIER DETECTION

Anomaly detection, otherwise known as outlier detection, refers to the search for observations that deviate from the rest of the observations. This procedure also includes steps to handle these anomalous data points.

Anomalies in a data source might come from a new sub-population that appears. For example, new types of vehicles with different fuel economy might be introduced in the prediction of vehicle MPG for all types of vehicles available in the market. Or, new types of customers are acquired from a recent expansion of a business or product line in predicting the sales for all stores in the country. Anomalous data can come from data quality considerations, or even malicious activity such as fraud, identify theft, or network intrusions. For example, an outlier network traffic pattern in a company's network system might be from hacker activity sending data to an unauthorized computer. Unlike data segmentation, which tends to divide the data into large portions, anomaly detection usually is performed to identify a small fragment of data that follows a significantly different pattern when compared with other data points.

Detection of outlier data points can be model-based and rule-based, or supervised and unsupervised, depending on whether anomalous data is available. Many visualization tools in SAS® Visual Analytics Explorer can be used to quickly identify some outliers in the data, such as histograms, scatter plots, box plots, and Sankey diagrams. For example, the observations with Vehicle MPG greater than 100 in Figure 2 are potential outliers to this data. In this application, the outliers are not from data contamination or malicious activity. Instead, they represents a small portion of higher fuel economy cars.

Linear and logistic regression models in SAS® Visual Statistics provide output that includes several influence statistics that rank the observations by their influence to the model fit. In many applications, higher value influence statistics indicate outliers. Continuing with the linear regression model on Vehicle MPG, in the model for the first segment (Segments of Make by THC = 1), if you lasso and select the first five bars in the influence statistics window (Figure 13), the corresponding data points are also highlighted in the Residual Plot. It is clear that these vehicles are underestimated (residuals from 50 to 200 mileage per gallon). Right-click on either of these two graphs and click **Show Selected** to examine the data (Figure 14). Alternately, you can use the **Exclude Selected** option in the right-click menu to exclude those observations, and the model will be refit.

Decision tree and clustering (including the parallel coordinates plot) models can be used to identify small fragments of the data that represent anomalous activities. The discussion on this topic is skipped in this paper.

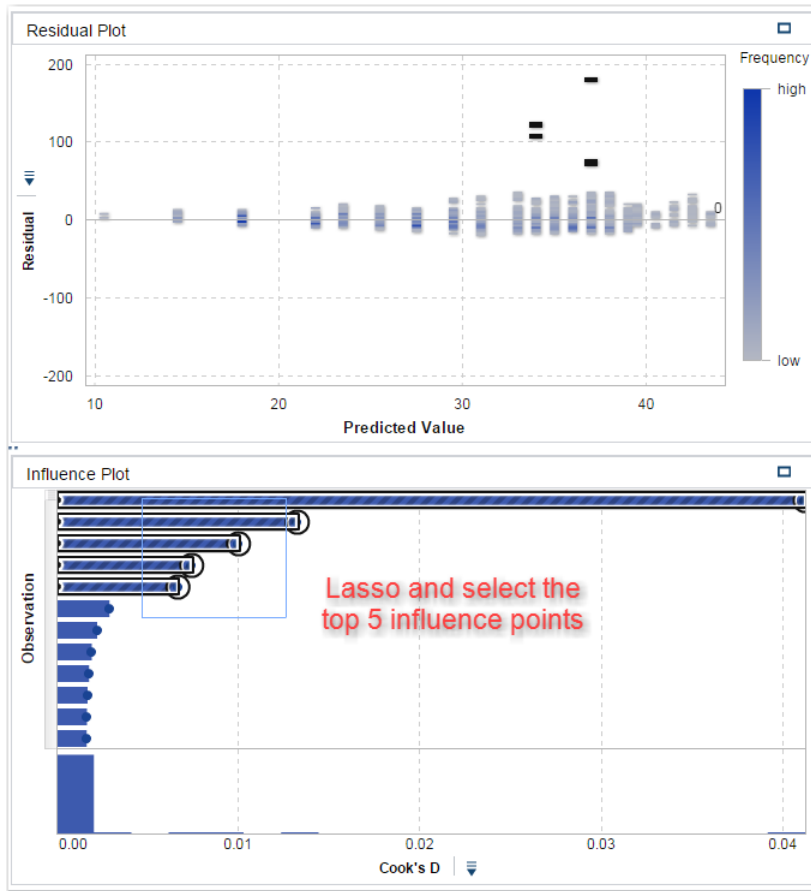


Figure 13. Using influence and residual plots to identify outliers. The selected observations can be examined by popping up a listing of the observations (Figure 14) or can be excluded in the model.

Predicted: MPG	Residual: MPG	Segments of Make by THC	Vehicle Make	Vehicle Model	Vehicle MPG	Vehicle Weight...	Cook's D
36.8780316	179.121968	1	TOYOTA	PRIUS PLUG-IN HYBRID	216	3500	0.041
34.0614715	122.938528	1	FORD	FOCUS	157	3875	0.013
34.0614715	122.938528	1	FORD	FOCUS	157	3875	0.013
34.0614715	122.938528	1	FORD	FOCUS	157	3875	0.013
34.0614715	106.938528	1	FORD	FOCUS	141	3875	0.010
34.0614715	106.938528	1	FORD	FOCUS	141	3875	0.010
34.0614715	106.938528	1	FORD	FOCUS	141	3875	0.010
36.8780316	76.1219684	1	TOYOTA	PRIUS PLUG-IN HYBRID	113	3500	0.007
36.8780316	72.1219684	1	TOYOTA	PRIUS PLUG-IN HYBRID	109	3500	0.007

Figure 14. The outlier observations selected in Figure 15. The observations represents a small fragment of fuel economy cars.

POST-MODELING VISUALIZATIONS

After you build a model in SAS® Visual Statistics, both the predicted values and the DATA Step score code of the model can be exported. The prediction outputs include cluster ID from clustering models, leaf ID from decision tree models, predicted values and residuals from linear regression and generalized linear regression models, and predicted probability of event and predicted values from logistic regression models. For the stratified model that you developed in the previous sections, the predicted value and residuals columns concatenate the outputs from five different linear regression models, and they are generated on the fly without reloading the original data source. The derived outputs can be used in any other visualization or models. Figure 11 and Figure 12 present two examples of how to use the output of one model in another visualization. Here, the leaf ID column is generated according to the terminal node assignment of a decision tree model and is then used in a linear regression visualization to stratify the data. Boosting is another example that uses the residual from one model as the response in the next model to further improve prediction accuracy.

In SAS® Visual Statistics you can also use the derived prediction outputs in a variety of visualizations provided by SAS® Visual Analytics Explorer to build predictive visualizations instead of exploratory visualizations. Examples includes variable profiling, score bands, and profit or loss calculations.

After completing the previous example, you can click on the dropdown menu (▾) of the linear regression visualization and derive the predicted values and residuals of Vehicle MPG. These new columns are available in the data pane. The last example in this paper (Figure 15) illustrates the use of Predicted MPG to profile two variables (Vehicle Cylinders and Vehicle Years) using bar charts and box plots. For simplicity, Vehicle Cylinders is filtered to a list of {4, 6, 8, 10, 12, and 16}

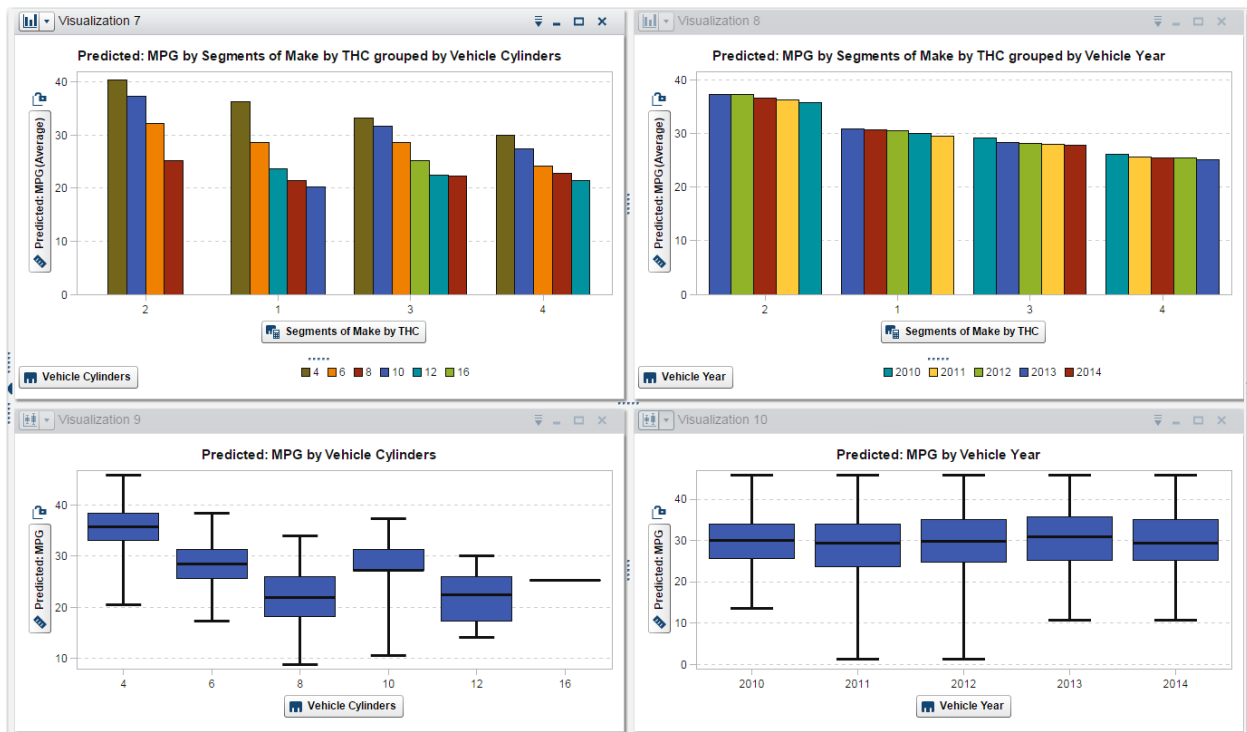
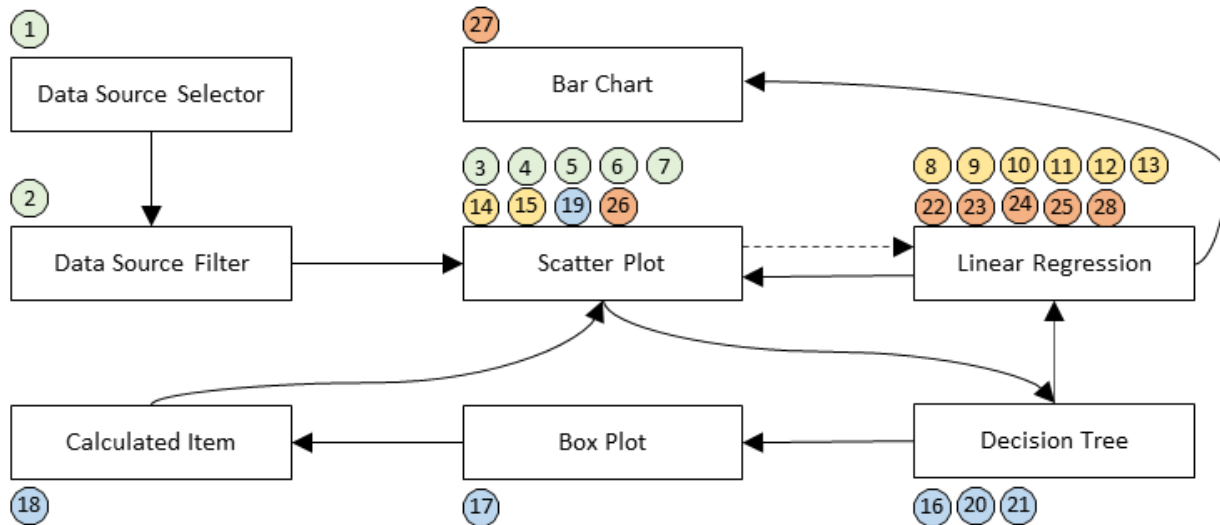


Figure 15. Variable profiling based on predicted MPG.

CONCLUSION

In conclusion, this paper uses SAS® Visual Statistics 7.2 and SAS® Visual Analytics Explorer 7.2 to form a working pipeline for predictive modeling. The flow chart below summarizes this pipeline with steps to repeat the pipeline in SAS® Visual Statistics.

Flow Chart 1. Summary of the predictive modeling pipeline developed in this paper.



Steps

1. Select the data source “EPA_CARS” In SAS® Visual Analytics Explorer.
2. Add a data source filter to only include Vehicle Type = ‘CAR’ or ‘TRUCK’.
3. Create a scatter plot, drag and drop Vehicle Weight and Vehicle MPG.
4. Fit a linear curve on the scatter plot.
5. Undo the previous action.
6. Drag and drop Vehicle Type as a **Color** variable.
7. Right click on the scatter plot to launch a linear regression. The **Color** variable is considered the **Group By** variable.
8. Check the results for both the CAR and the TRUCK segments.
9. Show hidden summary tables, then hide them again.
10. Turn off auto update.
11. Add Vehicle Make as a Rank to filter down to the top 25 vehicle makes by frequency.
12. Use the Advanced Group By window to filter down to the bottom 10 vehicle makes with lowest Emissions of Total Hydrocarbons.
13. Turn on auto update.
14. Go back to the scatter plot. Add a local filter on Vehicle Make that includes only MINI and NISSAN.
15. Replace Vehicle Type with Vehicle Make as the **Color** variable.
16. Build a decision tree model with Emission of Total Hydrocarbons and Vehicle Make. The response distribution is highly skewed. The remedy is to discretize the variable.
17. Use box plot to select a cutoff point. This example uses the third quantile of Emissions of Total Hydrocarbons).
18. Create an indicator variable THC Indicator. For Emissions of Total Hydrocarbons > 0.03 g/mi, return High, otherwise return Low.
19. Go back to the scatter plot, remove the local filter and replace Vehicle Make with THC Indicator.
20. Use THC Indicator as the **Response**, change **Max Levels** to 2, **Max Branches** to 4, **Leaf Size** to 1000.
21. Derive a Leaf ID variable named “Segment of Make by THC”.
22. Go back to the linear regression and remove the Rank variable.
23. Use Segment of Make by THC to replace the group by variable.
24. Lasso the outliers for a model and show selected outliers. Note that there is data brushing with the influence plot.
25. Derive predicted values from the dropdown menu of the linear regression.
26. Drag and drop predicted MPG and residual column to a new scatter plot, colored by Segment of Make by THC.

27. Start a new bar chart with Vehicle Type as a **Category** variable, Segment of Makes by THC as a **Group** variable, and Predicted: Vehicle MPG as a **Measure** variable. Use Average as the aggregation type.
28. Save the model score code from the dropdown menu of the linear regression.

REFERENCES

Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." *ACM Computing Surveys (CSUR)* 41.3 (2009): 15.

Cross, Glendon, Thompson, Wanye. "Understanding your customer: segmentation techniques for gaining customer insight and predicting risk in the telecom industry." SAS Global Forum 2008.

Neville, Padraic G. "Growing Trees for stratified modeling." *Computing Science and Statistics* (1998): 528-528.

SAS Visual Statistics. http://www.sas.com/en_us/software/analytics/visual-statistics.html

CONTACT INFORMATION <HEADING 1>

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

Xiangxiang Meng
SAS® Institute Inc.
Xiangxiang.Meng@sas.com

Wayne Thompson
SAS® Institute Inc.
Wayne.Thompson@sas.com

Jennifer Ames
SAS® Institute Inc.
Jennifer.Ames@sas.com

SAS® and all other SAS® Institute Inc. product or service names are registered trademarks or trademarks of SAS® Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.