Paper SAS5780-2016

SAS® Visual Statistics 8.1: The New Self-Service Easy Analytics Experience

Xiangxiang Meng, Cheryl LeSaint, Don Chapman, SAS Institute Inc.

ABSTRACT

In today's Business Intelligence world, self-service, which allows an everyday knowledge worker to explore data and personalize business reports without being tech-savvy, is a prerequisite. The new release of SAS® Visual Statistics introduces an HTML5-based, easy-to-use user interface that combines statistical modeling, business reporting, and mobile sharing into a one-stop self-service shop. The backbone analytic server of SAS Visual Statistics is also updated, allowing an end user to analyze data of various sizes in the cloud. The paper illustrates this new self-service modeling experience in SAS Visual Statistics using telecom churn data, including the steps of identifying distinct user subgroups using decision tree, building and tuning regression models, designing business reports for customer churn, and sharing the final modeling outcome on a mobile device.

INTRODUCTION

When analyzing data, it is important to be able to easily identify relationships between the variables. By identifying relationships, you are able to make predictions for variables of interest. A common variable of interest is a binary variable. Should a person be admitted to a program? Is this transaction fraudulent? This paper examines data on whether a customer cancels his or her subscription, which is also known as churn. The churn variable's relationship with other information about the customer's account is examined in a variety of ways throughout this paper. The exploratory data analysis and feature engineering all build up to identifying significant predictors of churn through a logistic regression. The combination of SAS® Visual Analytics and SAS® Visual Statistics in the 8.1 release brings data analysis and reporting to your fingertips. The 8.1 release of SAS Visual Analytics and SAS Visual Statistics was pre-production when this paper was authored; therefore, details are subject to change.

SAS VISUAL STATISTICS 8.1

UNIFICATION

The 8.1 release of SAS Visual Analytics is rebuilt from the ground up and combines SAS® Visual Analytics Explorer, SAS Visual Analytics Designer, and SAS Visual Statistics into a single user interface. Pie charts, histograms, and linear regressions meld together. Exploratory tasks, such as auto-charting and summary tables, sit alongside analytical tasks, such as linear regression and clustering. These tasks are seamlessly blended in classic reporting capabilities, such as a rich layout system and display rules. This unified experience allows SAS Visual Analytics and SAS Visual Statistics users to execute the Business Intelligence and Analytics continuum, as shown in Figure 1.

MODERN DESIGN

The modern user experience supports the modeling needs of statisticians and data scientists as well as the reporting needs of business analysts. The application is designed to run in your browser and is written entirely in HTML5. This opens up options for developing models away from the desktop.

The underlying infrastructure has been rewritten to meet the deployment demands of the future. It supports more versatile deployment options ranging from on-site to in-the-cloud deployments by leveraging microservices that meet the scaling and reliability needs expected in today's world. You can also take advantage of the next-generation in-memory analytics server from SAS® that unifies all the analytics procedures into a single server.

Data access is built into SAS Visual Statistics. You have full access to enterprise data stored in Hadoop or in the corporate database. You also have the ability to upload your own data stored in plain text files, Excel spreadsheets, or a SAS data set. This level of self-service, along with a rich set of data manipulation capabilities, enables the feature engineering expected by sophisticated users.

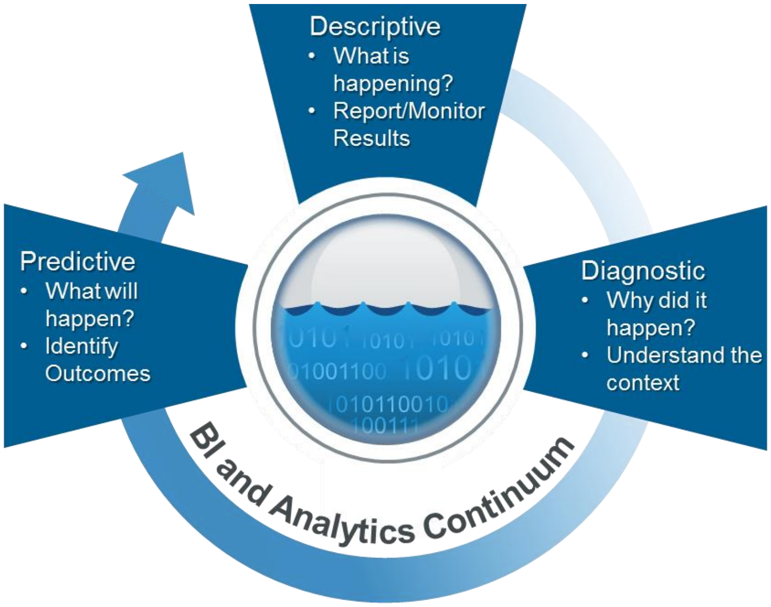
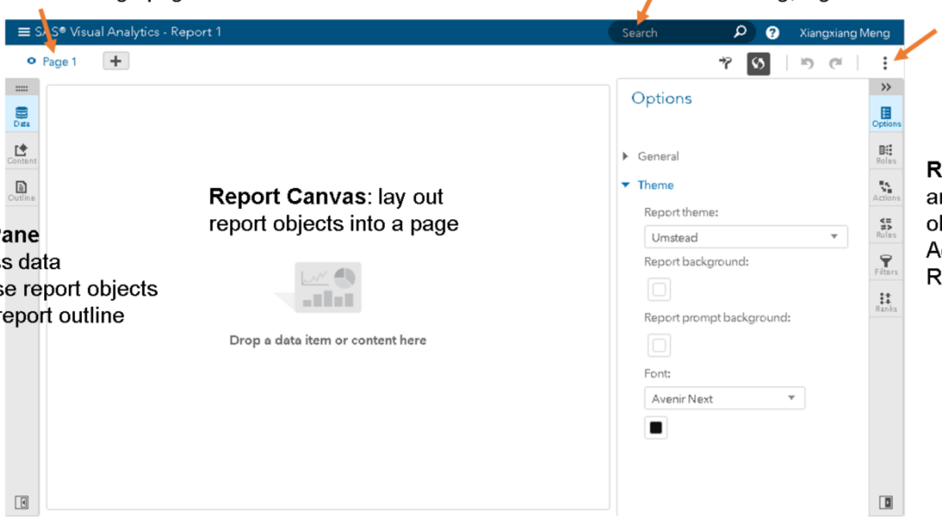


Figure 1. Business Intelligence and Analytics Continuum

USER EXPERIENCE

SAS Visual Analytics and SAS Visual Statistics has been seamlessly integrated, making all modeling work immediately accessible in a report. Documenting and sharing models are easy; you just save your work. This work can be viewed on mobile devices using the SAS Visual Analytics mobile viewers.

Often the generation of a report is secondary to interactively building models and generating score code. SAS Visual Statistics does not require report layout; it is there only if needed. It also provides the tools to compare two models side-by-side, or to evaluate them interactively using the model comparison task. Figure 2 provides a basic introduction to the layout of the user interface of SAS Visual Analytics 8.1. SAS Visual Statistics is an add-on to SAS Visual Analytics, which provides additional statistical modeling tasks in the **Content** menu.



Report theme: Umstead

Report background: □

Report prompt background: □

Font: Avenir Next

■

Page Navigation: navigate and manage pages

Banner: help, search, user setting, log out

Toolbar Menu: common application actions such as save and open

Report Canvas: lay out report objects into a page

Left Pane
Access data
Choose report objects
View report outline

Right Pane: Define and modify a report object (Options, Roles, Actions, Rules, Filters, Ranks)

Figure 2. User Interface of SAS Visual Analytics 8.1

CASE STUDY USING TELECOM CHURN DATA

DATA DESCRIPTION

The examples in this paper are derived using telecommunications data. The variable of interest is whether the customers cancel their service or not, also known as churn. The data set is available from the UCI Machine Learning Repository of databases. The data set contains 3,333 observations, where each row contains the information collected for a customer account. Table 1 provides a brief description about the variables in this table.

Variable Name	Description
Churn	Label of churn (Yes/No).
Day_Calls, Day_Charge, Day_Mins	Total day calls, charges, and minutes
Eve_Calls, Eve_Charge, Eve_Mins	Total evening calls, charges, and minutes
Night_Calls, Night_Charge, Night_Mins	Total night calls, charges, and minutes
Intl_Calls, Intl_Charge, Intl_Mins	Total international calls, charges, and minutes
Account_Length	Length of account before churn
CustServ_Calls	Total number of customer service calls
State	States of USA in which the customer account is registered
Intl_Plan	Indicator for international plan
VMail_Message	Number of voice mail messages
VMail_Plan	Indicator for voice message plan

Table 1. Description of the Telecom Churn Data

EXPLORING THE DATA

SAS Visual Analytics 8.1 provides a variety of tools for data visualization and exploration. For the churn data, we are interested in exploring the data and identifying factors that influence churn. Figure 3 shows a collection of several visualizations created in SAS Visual Analytics 8.1.

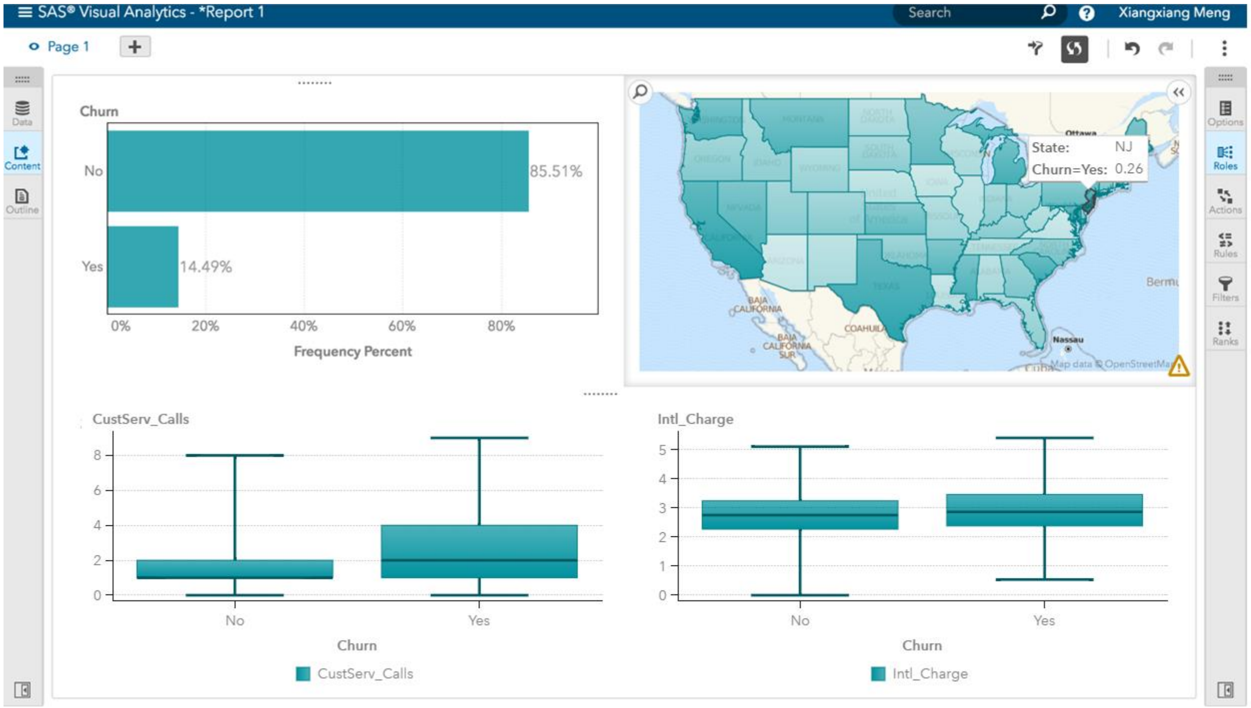


Figure 3. Data Exploration Using SAS Visual Analytics 8.1

The bar chart shows that 14.5 percent of the customer accounts are churners. The geo map shows the percentage of churning across different states. New Jersey has the highest churn rate (26 percent). Note that the color column used in the geo map is a new numeric calculated item CHURN=YES. We explain how to create this calculated item in the next section.

You can also look at the distribution of a data column in SAS Visual Analytics. For example, the box plots in Figure 3 compare the distribution of total numbers of international calls and customer service calls for churners and non-churners. It is interesting to see that churners have more customer service calls.

FEATURE ENGINEERING

Feature engineering is an important step to improve the performance of a statistical or machine learning model, and it is recognized as the most manual and time-consuming effort in a learning process. SAS Visual Analytics 8.1 provides a few tools to create new features from existing columns. These features are created on-demand and do not require additional disk and memory footprint. Within SAS Visual Analytics 8.1, you can do the following tasks:

- Determine whether a variable should be used as categorical or measure.
- Create a new hierarchy using a set of categorical variables.
- Create a custom category that represents a grouping of the levels of a categorical variable with high cardinality.
- Create a calculated item using user-specified formulas.

Most of these options are available in the **Add** menu of the Data pane, as shown in Figure 4.

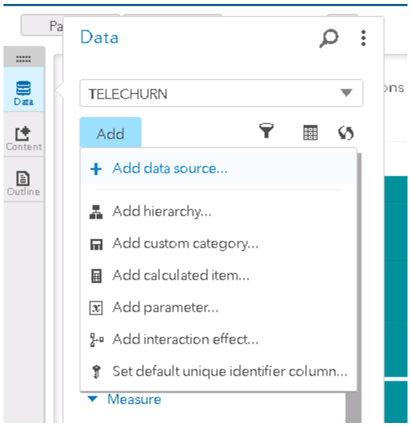
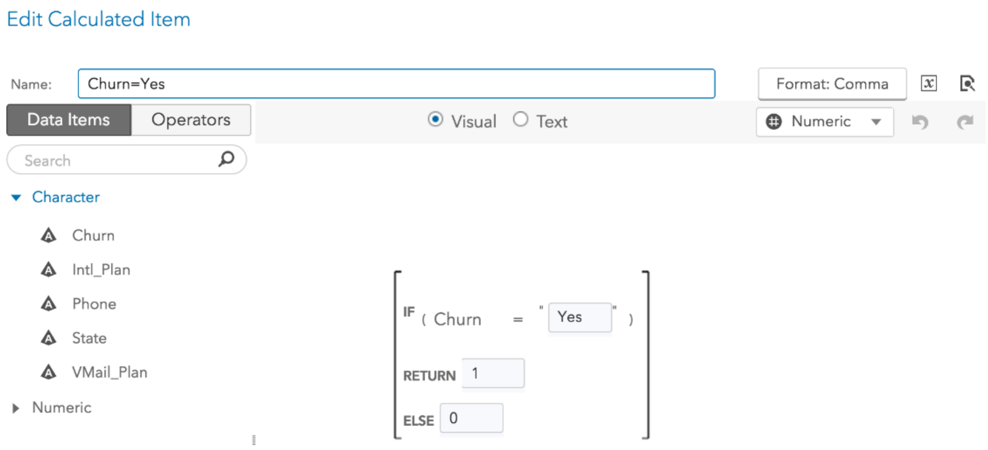


Figure 4. Add Menu for Creating New Columns

The CHURN column contains character values YES and NO, which cannot be accepted by a few visualizations that require numeric aggregators. A workaround is to create dummy indicators using a calculated item for CHURN = Yes and CHURN = No, as shown in Figure 5.



Edit Calculated Item

Name: Churn=Yes

Format: Comma

● Visual ○ Text

Numeric

IF (Churn = "Yes")
RETURN 1
ELSE 0

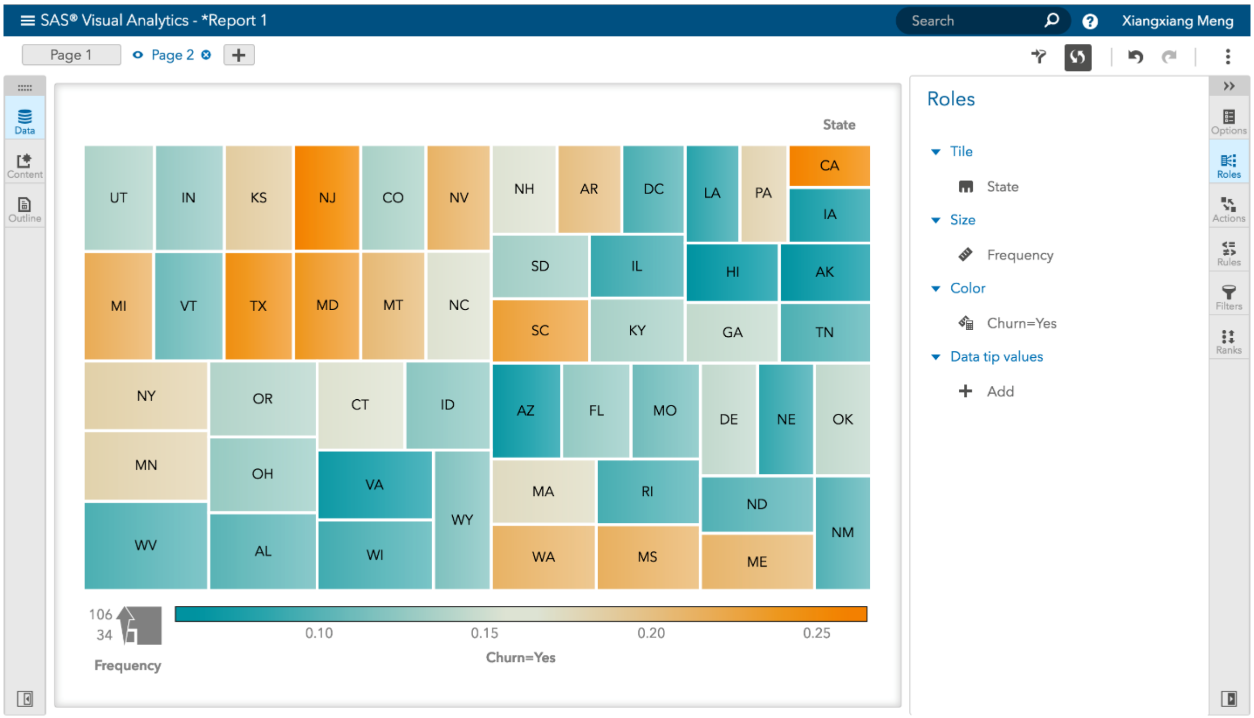
Figure 5. Create Dummy Variables in SAS Visual Analytics

You can drag and drop to create a calculated item using the data items and the operators provided in the Edit Calculated Item window shown in Figure 5, with the output column as either numeric or character. Alternatively, you can also use the text editor to create a new item. For example, the above dummy variable can be created using the following code:

```
    IF ( 'Churn'n = 'Yes' )
    RETURN 1
    ELSE 0
```

Measure data items have a default aggregation type of Sum in SAS Visual Analytics. For the calculated item of Churn=Yes, a more natural aggregation type is Average. This change can be made in the Data pane by editing the properties of Churn=Yes.

You can now use the Churn=Yes column in a treemap to compare average churn rates across the states, as shown in Figure 6. The size of each rectangle represents the number of accounts for that state.



SAS® Visual Analytics - *Report 1

Roles

Tile: State

Size: Frequency

Color: Churn=Yes

Data tip values: Add

Figure 6. Treemap of Churn Rates across States Using the Calculated Churn=Yes Column

The data contains only total charges for each type of call (Day, Evening, Night, and International). You can easily derive other features such as average charge per call, total numbers of domestic calls, total domestic charge, and so on:

```
'Day_Charge'n   / 'Day_Calls'n                /* Day_Avg_Charge */
'Eve_Charge'n   / 'Eve_Calls'n                /* Eve_Avg_Charge */
'Night_Charge'n / 'Night_Calls'n              /* Night_Avg_Charge */
'Intl_Charge'n  / 'Intl_Calls'n               /* Intl_Avg_Charge */
'Day_Calls'n + 'Eve_Calls'n + 'Night_Calls'n  /* Total_Domestic_Call*/
'Day_Charge'n + 'Eve_Charge'n + 'Night_Charge'n /* Total_Domestic_Charge*/
```

Note that deriving multiple calculated items or deriving calculated items multiple times does not require additional disk storage or data passes. The definitions of the calculated items are attached to the in-memory data source and are computed only when they are used by a visualization or model. Figure 7 shows a scatter plot of the two calculated items DAY_AVG_CHARGE and EVE_AVG_CHARGE. You can easily identify data anomalies: one account has an extremely high average evening charge and a few accounts have high average day charges, the majority of which are churners.

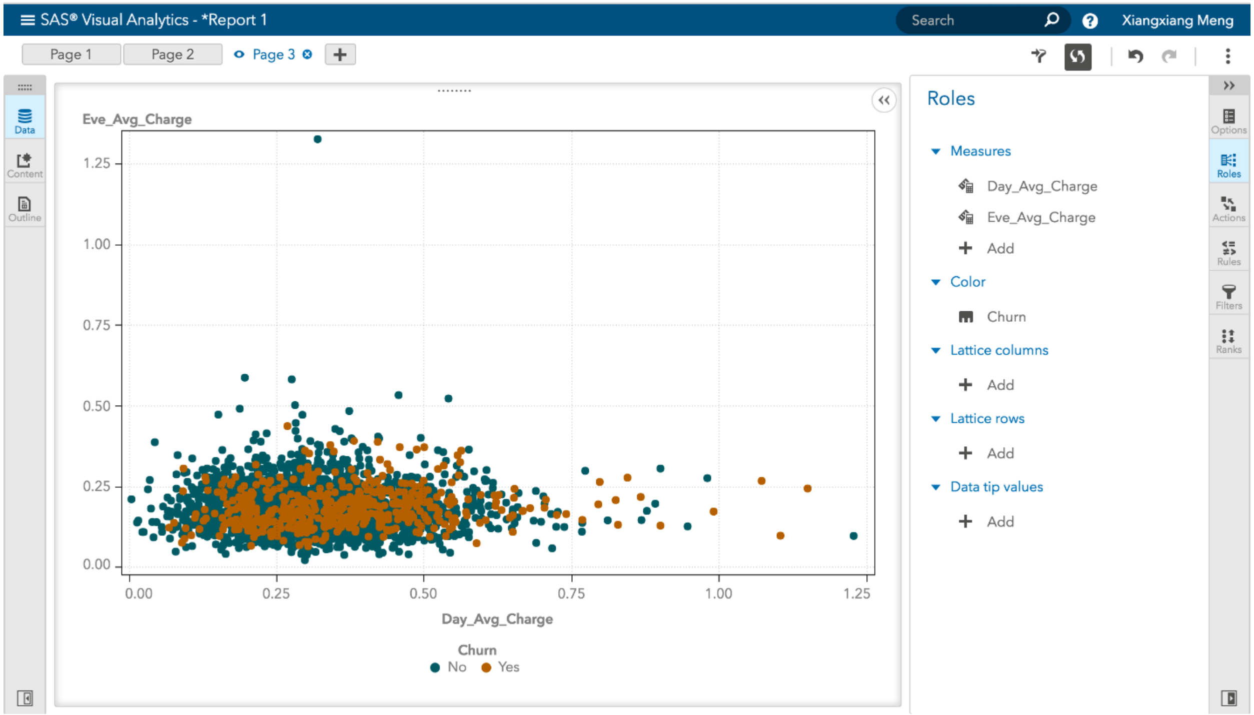


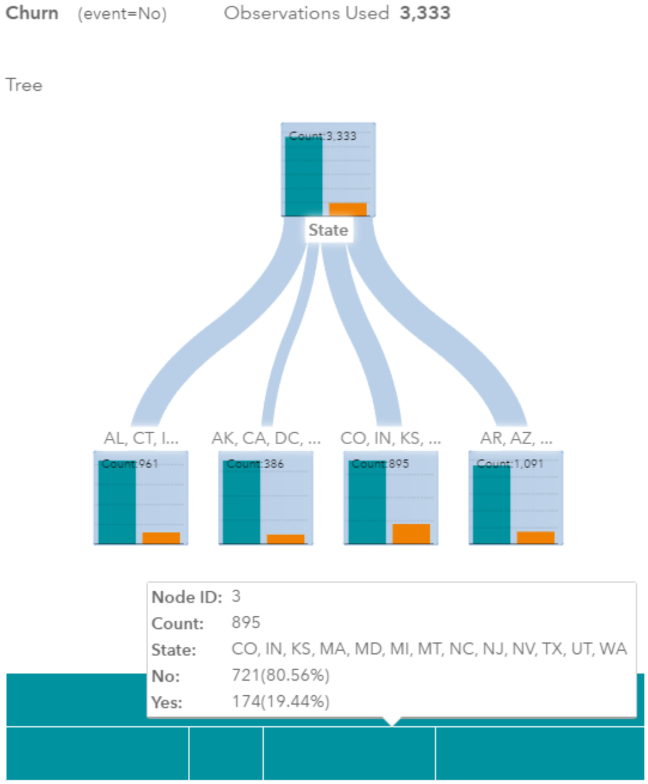
Figure 7. Scatter Plot of the Calculated Average Day and Evening Charges

DATA SEGMENTATION

The geo map in Figure 3 and the treemap in Figure 6 show that churn rate varies across states. This implies STATE is a significant factor for predicting churns. However, the STATE column is a high-cardinality variable with 51 levels and you might not want to use it in a model directly. SAS Visual Statistics provides several methods for dimension reduction and data segmentation. For example, you can use the decision tree model in SAS Visual Statistics to group the levels of a high-cardinality variable into several leaves, based on a target variable, as shown in Figure 8.

In this example, a decision tree model is built with response variable CHURN and only one predictor STATE. Decision trees are widely used in many applications such as predictive modeling, data segmentation, and outlier detection. Each application requires different tree parameter settings. For data segmentation, you often need a smaller tree to ensure each leaf of the tree has enough observations. Figure 8 shows a two-level decision tree with four branches. The tooltip shows that the third node (Node

ID = 3) is the data segment with highest churn rate (19.44 percent) and contains the following states: CO, IN, KS, MA, MD, MI, MT, NC, NJ, NV, TX, UT, and WA.



Churn (event=No) Observations Used 3,333

Tree

Node ID: 3
Count: 895
State: CO, IN, KS, MA, MD, MI, MT, NC, NJ, NV, TX, UT, WA
No: 721(80.56%)
Yes: 174(19.44%)

Figure 8. A Two-level Four-branch Decision Tree for Data Segmentation

It is often desirable to use the data segmentation from a decision tree model in other models. With SAS Visual Statistics, you can derive a leaf ID column that represents the leaf assignment of the observations. Figure 9 shows the right-click menu (on a mobile device, hold to pop up this menu) to derive a leaf ID variable. For this use case, the leaf ID contains four values (1, 2, 3, 4) that represent the grouping of 51 states into four segments with different levels of churn rates (CHURN = Yes).

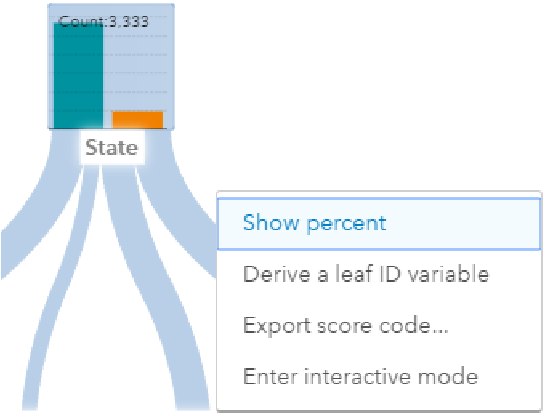
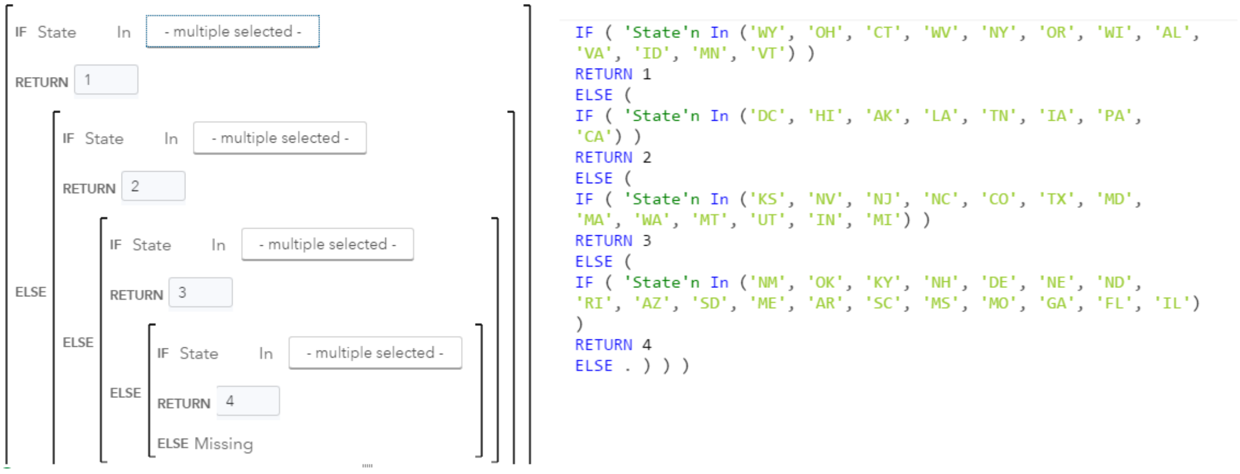


Figure 9. Derive Segmentations (Leaves) Using Decision Tree

You can edit the new derived column as well. Figure 10 shows both the visual and text edit windows for the column derived from the decision tree model. You can override the definition of the leaf IDs by a pre-determined business rule.


IF State In - multiple selected -
RETURN 1
ELSE
IF State In - multiple selected -
RETURN 2
ELSE
IF State In - multiple selected -
RETURN 3
ELSE
IF State In - multiple selected -
RETURN 4
ELSE Missing

```
IF ( 'State'n In ('WY', 'OH', 'CT', 'WV', 'NY', 'OR', 'WI', 'AL',
'VA', 'ID', 'MN', 'VT') )
RETURN 1
ELSE (
IF ( 'State'n In ('DC', 'HI', 'AK', 'LA', 'TN', 'IA', 'PA',
'CA') )
RETURN 2
ELSE (
IF ( 'State'n In ('KS', 'NV', 'NJ', 'NC', 'CO', 'TX', 'MD',
'MA', 'WA', 'MT', 'UT', 'IN', 'MI') )
RETURN 3
ELSE (
IF ( 'State'n In ('NM', 'OK', 'KY', 'NH', 'DE', 'NE', 'ND',
'RI', 'AZ', 'SD', 'ME', 'AR', 'SC', 'MS', 'MO', 'GA', 'FL', 'IL')
)
RETURN 4
ELSE . ) ) )
```

Figure 10. Editable Tree-based Segmentation

Deriving a leaf ID variable is one way of saving the results of a model. All models in SAS Visual Statistics allow you to save the analytical contents for later use. You can do the following tasks can do the following tasks:

- Save the model as part of the report. If you open a saved model and the underlying data source has been updated, the model is automatically retrained.
- Save a footprint of the model as SAS DATA step code (score code). You can use the score code to score a new data source for either prediction or validation purposes.
- Derive new columns from the model. These columns are attached to the currently loaded table and can be further used in any other models or report objects.

BUILDING LOGISTIC REGRESSION

Connectivity between different modules is also important for a self-service analytical platform. SAS Visual Statistics 8.1 allows you to derive predicted outcomes from a model and use them to build a report or another model. You can also link a model to a control or a visualization. In this section we demonstrate a group-by logistic regression use case using the data segmentation derived from the decision tree model. First let us build the basic report layout. Figure 11 shows a layout of a report page that contains a button control (top), a bar chart (middle left), and a logistic regression model (middle right). Variables have not been assigned to the components, and therefore each component displays sample output with a red warning icon to indicate that it is still under construction.

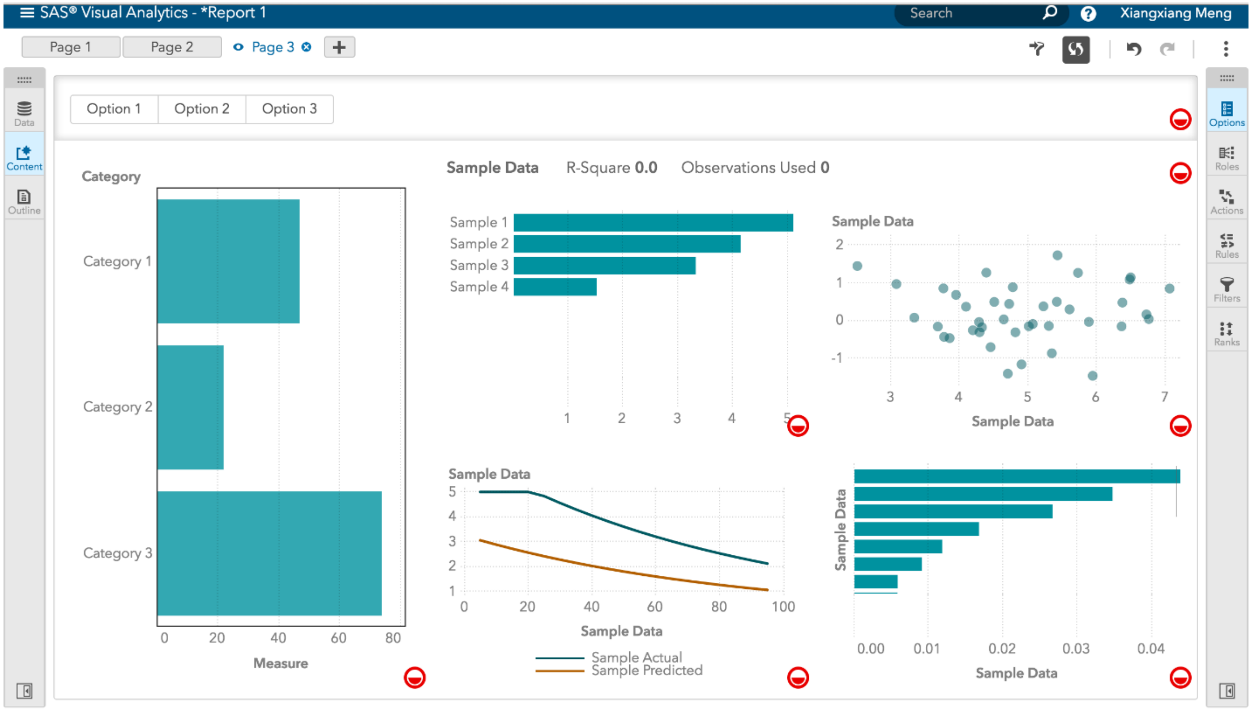
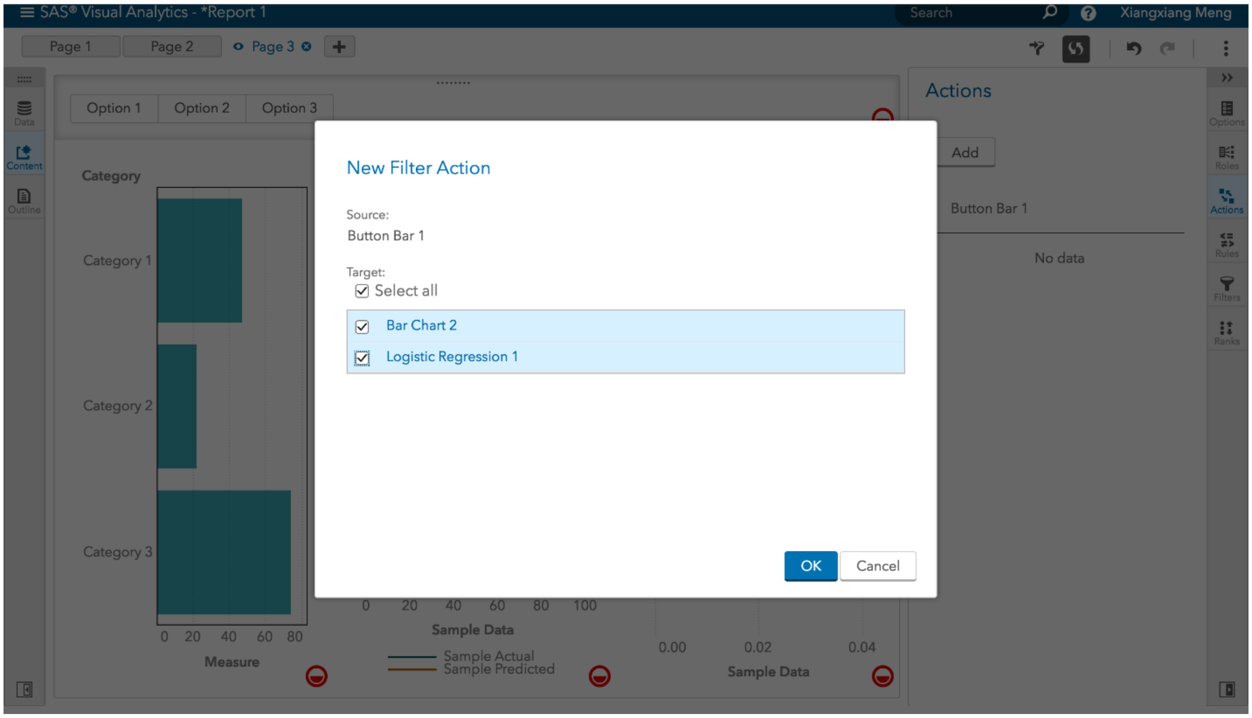


Figure 11. Layout of Various Report Objects and a Logistic Regression without Filling in Actual Data

Second, we link both the bar chart and the logistic regression to the button bar control. This is done by creating a new Filter Action from the button bar, as shown in Figure 12.



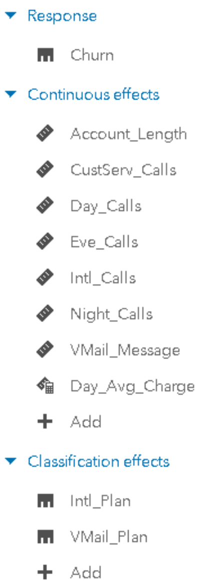
New Filter Action

Source:
Button Bar 1

Target:
☑ Select all
☑ Bar Chart 2
☑ Logistic Regression 1

Figure 12. Creating New Filter Action

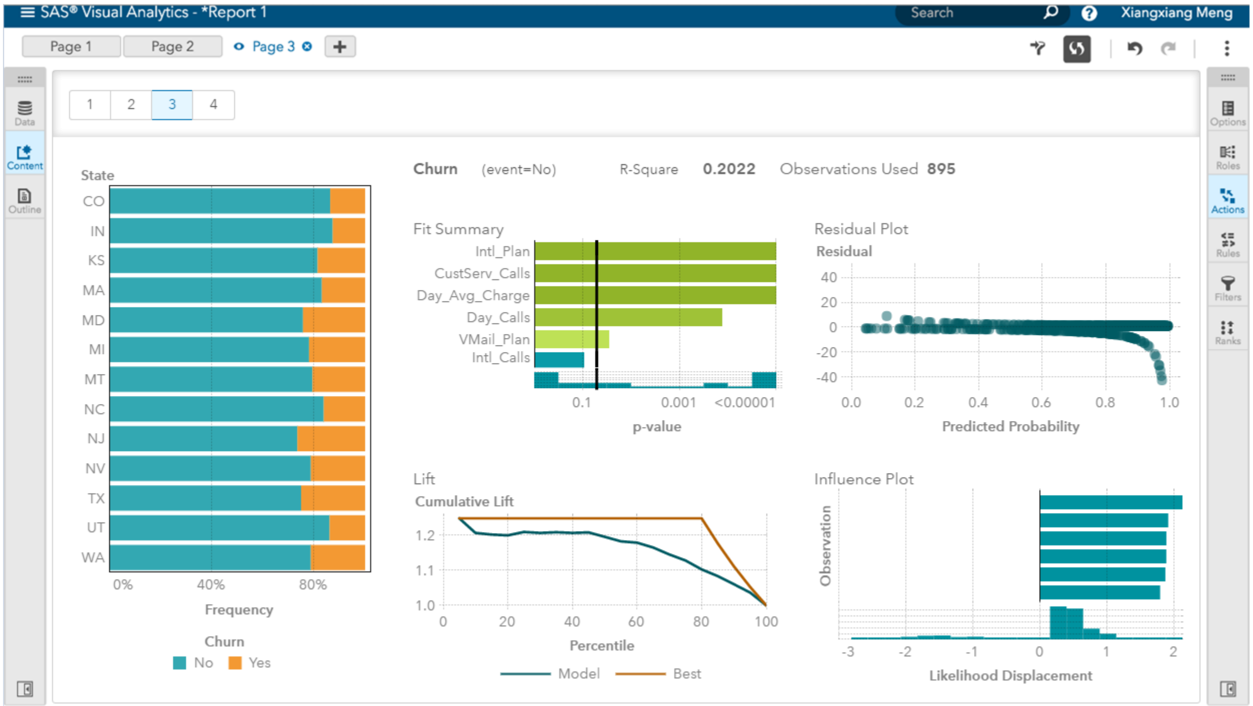
The last step is to assign data roles to each visualization. For example, you can assign the new LEAF ID to the button bar, assign STATE and CHURN to the bar chart, assign CHURN as the response variable, and assign various explanatory effects to the logistic regression model, as shown in Figure 13.



- Response
  - Churn
- Continuous effects
  - Account_Length
  - CustServ_Calls
  - Day_Calls
  - Eve_Calls
  - Intl_Calls
  - Night_Calls
  - VMail_Message
  - Day_Avg_Charge
  - Add
- Classification effects
  - Intl_Plan
  - VMail_Plan
  - Add

Figure 13. Response and Effects Used in the Logistic Regression

By default, the button bar control is not active and the bar chart and the logistic regression are built based on the entire data (assuming no missing values). If you click a value on the button bar control, the bar chart and the logistic regression model are updated to use only the data with a specific level of the button bar variable. Figure 14 shows a report of the churn distribution across the states and the logistic regression model built for the leaf ID = 3 data segmentation. Note that the model is trained and the visualizations are rendered only the first time you click a value on the button bar. It will be cached afterward and the model won't be refit when you switch the views.



Churn (event=No) R-Square 0.2022 Observations Used 895

Figure 14. Response Distribution and Logistic Regression Model for the Leaf ID = 3 Data Segment

Looking at the results for the logistic regression, you can see in the Fit Summary plot the variables that are significant at predicting whether a customer will cancel (churn). Here you can see that whether they have an international plan (Intl_Plan), the total number of customer service calls they make

(CustServ_Calls), the average charge for their calls during the day (Day_Avg_Charge), the total number of calls during the day (Day_Calls), and whether they have a voice message plan (VMail_Plan) are all significant. This aligns with some of the exploratory data analysis. The box plots showed more separation between churners and non-churners for CustServ_Calls than Intl_Call. The fact that Day_Avg_Charge is significant at predicting churn was seen in the scatter plot, which showed large values of this variable are associated with customers that churn. These significant predictor variables would be good to focus on when attempting to reduce customer churn.

MOBILE VIEWER

Exploratory data analysis, model construction, and report building can all be done through SAS Visual Analytics and SAS Visual Statistics using a web browser from a desktop client or mobile device. Once a final report has been settled on that summarizes the findings of your analysis, the report can be viewed by many. A saved report can be shared simply be opening it in SAS Visual Analytics Viewer. From SAS Visual Analytics Viewer, the user can view and interact with all pages in the report, email the report to others, and print interesting results.

CONCLUSION

In conclusion, SAS Visual Analytics and SAS Visual Statistics 8.1 provide a unified platform for your analytic journey. The paper uses churn data to demonstrate this new self-service experience and provides a working example of exploring, manipulating, modeling, and building business reports of the churn data, all in a single user interface.

CONTACT INFORMATION

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

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

Cheryl LeSaint
SAS Institute Inc.
Cheryl.LeSaint@sas.com

Don Chapman
SAS Institute Inc.
Don.Chapman@sas.com