SAS® Visual Analytics 8.1: Getting Started with Analytical Models

Using This Book

Audience

This book covers the basics of building, comparing, and exploring analytical models in SAS Visual Analytics. The example in this book leverages the model-building capabilities of SAS Visual Statistics and SAS Visual Data Mining and Machine Learning, in addition to the reporting and exploration features of SAS Visual Analytics.

Although the example in this book includes feature engineering, data exploration, model building, and model comparison, the emphasis is on familiarizing yourself with the SAS Visual Analytics user interface. This book is designed primarily for those with some analytic experience, but who are unfamiliar with the SAS Visual Analytics interface. It is also appropriate for those who are familiar with the reporting and exploration features of SAS Visual Analytics and want to learn more about the analytical features.

Requirements

To perform the tasks in this book, you need the following software, information, and privileges:

- a link to a working deployment of SAS Visual Analytics 8.1
- a supported web browser (see the SAS support site for supported web browsers and versions)
- an account that can log on to the working deployment
- the input data provided for this book

Introduction

About SAS Visual Statistics and SAS Visual Data Mining and Machine Learning

SAS Visual Statistics and SAS Visual Data Mining and Machine Learning add modeling tools to SAS Visual Analytics. SAS Visual Statistics adds:

- linear regression
- logistic regression
- generalized linear model
- clustering

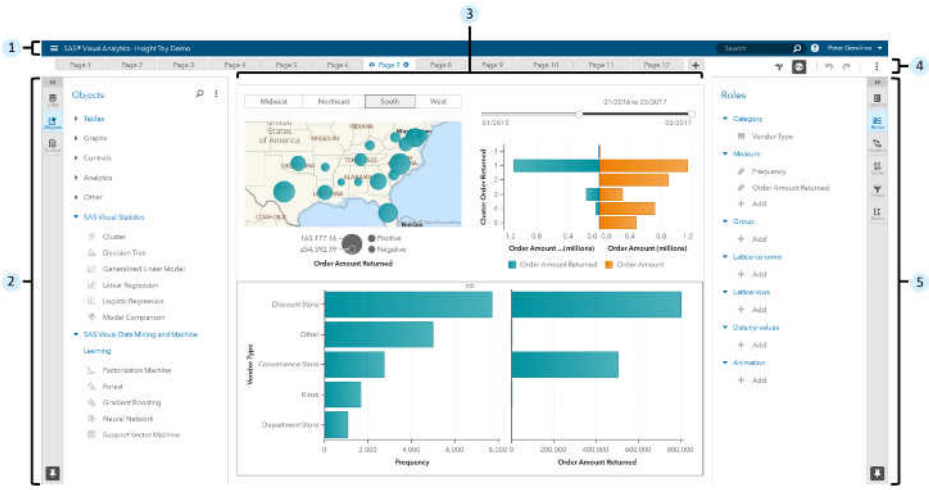
In addition, you gain access to a more powerful decision tree that includes the ability to interactively train the tree by selecting which branches to prune, split, or train. SAS Visual Data Mining and Machine Learning adds:

- factorization machine
- forest
- gradient boosting
- neural network
- support vector machine

Models that you create in SAS Visual Analytics can be exported and used to score new data. Although the exact export method varies by model, exported information can be in the form of DATA step code, an ASTORE table, and segment ID. For more information about what is exported for each model, see the SAS Visual Statistics and SAS Visual Data Mining and Machine Learning documentation within the SAS Visual Analytics documentation.

With SAS Visual Statistics and SAS Visual Data Mining and Machine Learning, you can perform a variety of tasks critical to creating good predictive models. These tasks include data segmentation (supervised and unsupervised), supervised variable transformation, stratified modeling, outlier detection, interactive feature creation, data filters, and post-model visualization. The example provided in this book covers many of these tasks. However, not all tasks available in SAS Visual Statistics or SAS Visual Data Mining and Machine Learning are covered in this book.

The SAS Visual Analytics Interface



1 The application bar at the top enables you to access other SAS applications. The name of your report is displayed in the application bar. You can search for items, access help, update your settings, and sign out of SAS Visual Analytics.

2 The left pane enables you to work with data, report objects, and the report outline.

3 The canvas is the workspace for building a report. The appearance of the canvas is affected by the report theme.

4 The menu bar enables you to create pages, show or hide report and prompt areas, undo, redo, and access other menu options.

5 The right pane enables you to work with details about the report and report objects.

About the Example in This Book

The example in this book is intended to mimic the process a data scientist might follow when solving a specific business problem. Tasks might include acquiring and preparing data, exploring the data, generating and comparing models, and documenting the insights gained from a model to act on that information. In this book, the specific business problem the data scientist is trying to solve is increasing company profits at Insight Toy Company by 5%, compared to last year.

Suppose that you are the data scientist at Insight Toy Company. Sales to your vendors have been slowly declining. Your manager wants to investigate ways to increase profits. Fortunately, your IT department has gathered two years of data on all aspects of your business, and it is readily available. This data includes information about what products are sold, to which vendors they are sold, the associated costs, and some metrics about the sales reps and the vendors.

Your plan is to first review the data provided by your IT department. You want to perform feature engineering to create necessary columns. You want to visualize the data to get a better understanding of its most important features. You want to create and compare several models to determine a champion model. Finally, you want to apply that model to determine potential solutions to increase company profits.

After completing the example, you are encouraged to repeat any of the steps with a different set of inputs, different model types, or different visualizations.

Feature Engineering

About the Tasks in This Chapter

Feature engineering is the process of determining and creating the inputs that are used in a machine learning model. This process typically involves both a general knowledge of the models to be used and a domain-specific knowledge about the problem to be solved. In this chapter, you start the process by creating some features that are essential to solving the problem of increasing profits. In the final section of this chapter, you do some data exploration before creating a new feature.

Getting Started

1. Save the insightToyDemo.sas7bdat example data located on the SAS Visual Analytics documentation page. Note the location where you saved the data.
2. Sign in to SAS Home.
3. In SAS Home, click the **SAS Visual Analytics** tile to launch SAS Visual Analytics.
4. Click the **Data** button in the Welcome window to load your data. A window appears that enables you to open the data source for this project.
5. Click **Import** in the Open Data Source window, and click **Choose Files**.
6. Navigate to the location where you saved the insightToyDemo.sas7bdat data set, and select **insightToyDemo.sas7bdat**. Click **Open**.
7. Click the **Import Item** button.
8. Click **OK**.

Calculate Total Costs

For this example, the total cost of an order is the sum of its distribution, marketing, product, and sales costs. To create the **Total Cost** measure, complete the following steps:

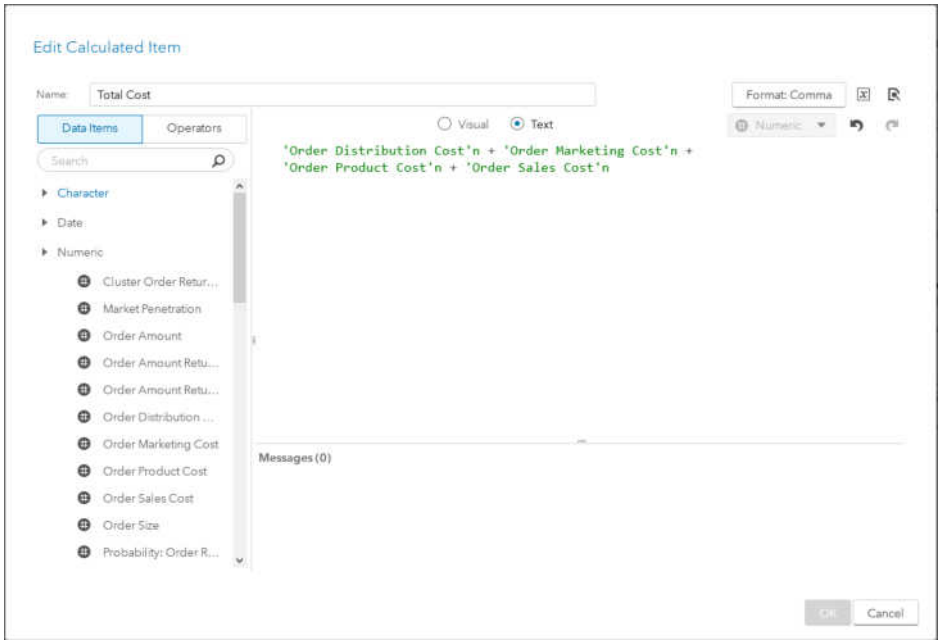
1. In the **Data** pane, click **Add**, and select **Add calculated item**.
2. In the **Name** field, enter `Total Cost`.

4

By default, the calculated item type is set to **Numeric** with **Format: Comma**. These are correct for this new item.

3 Select **Text** in the upper center of the window.

4 Replace any text in the field with the following expression into the code field: `'Order Distribution Cost'n + 'Order Marketing Cost'n + 'Order Product Cost'n + 'Order Sales Cost'n`, and click **OK**.



The **Total Cost** item appears in the **Measure** list.

5 Click the 💾 icon to save the project. Because this is the first time saving your project, the Save As window appears. Navigate to a folder where you have Write permission. In the **Name** field, enter **`Insight Toy Demo`**. Click **Save**.

Calculate the Order Profit Measure

Order profit is defined as the amount that you received for an order minus the amount returned and the total cost. To create the **Order Profit** measure, complete the following steps:

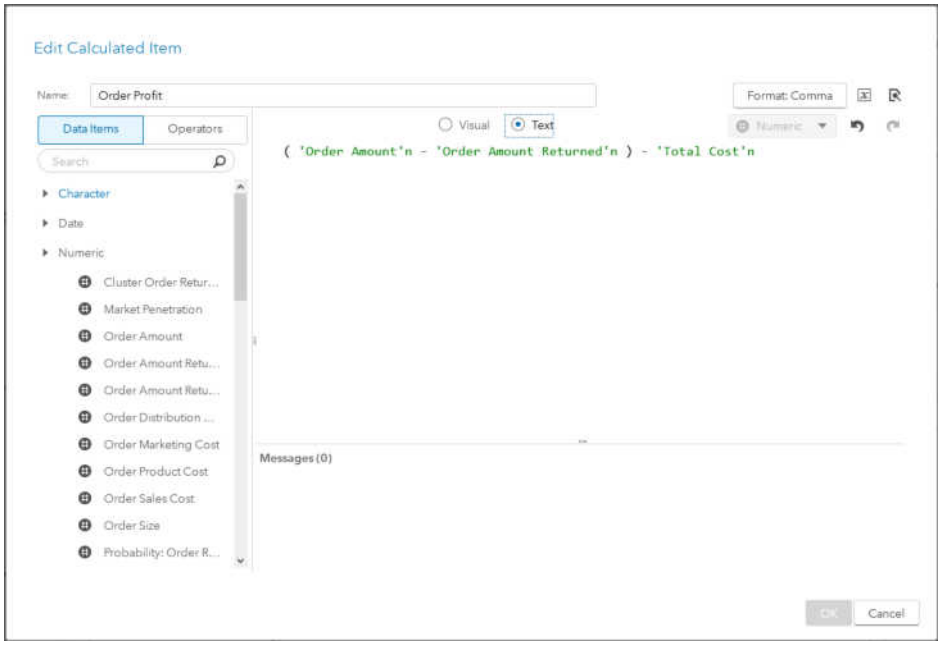
1 In the **Data** pane, click **Add**, and select **Add calculated item**.

2 In the **Name** field, enter **`Order Profit`**.

By default, the calculated item type is set to **Numeric** with **Format: Comma**. These are correct for this new item.

3 Select **Text** in the upper center of the window.

4 Enter the following expression into the code field: `('Order Amount'n - 'Order Amount Returned'n) - 'Total Cost'n`, and click **OK**.



The **Order Profit** item appears in the **Measure** list.

5 Save the project.

Create a Monthly Time Variable

To create the monthly time variable, complete the following steps:

1 In the **Data** pane, locate the **Order Date** category.
2 Right-click **Order Date**, and select **Duplicate**. This creates the category **Order Date (1)**.
3 Select **Order Date (1)**, and click .
4 In the **Name** field, enter `Order Date Month`.
5 For **Format**, click the **MMDDYYYY** button to change the category's format. In the Format window, select **MMYYYY**. Click **OK**.
6 Save the project.

Create a Vendor Region Category

Although your data contains the state-level location data for each vendor, it is sometimes better to group vendors by census region instead. To create a custom category, complete the following steps:

1 In the **Data** pane, click **Add**, and select **Add custom category**.
2 In the **Name** field, enter `Vendor Region`.
3 In the **Based on** field, select **Vendor State/Province**.
4 In **Values of Vendor State/Province**, select the following values: **Alabama**, **Arkansas**, **Delaware**, **District of Columbia**, **Florida**, **Georgia**, **Kentucky**, **Louisiana**, **Maryland**, **Mississippi**, **North Carolina**, **Oklahoma**, **South Carolina**, **Tennessee**, **Texas**, **Virginia**, and **West Virginia**.

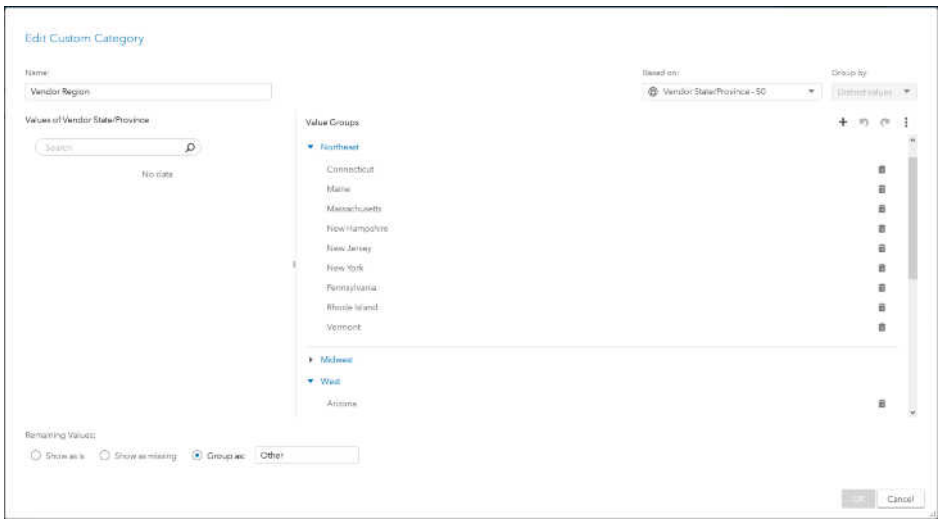
   Drag these items into **Value Group 1**.
5 Right-click **Value Group 1**, and select **Edit group name**. Enter `South` for **Name** in the Edit Group Name window, and click **OK**.
6 In **Values of Vendor State/Province**, select the following values: **Connecticut**, **Maine**, **Massachusetts**, **New Hampshire**, **New Jersey**, **New York**, **Pennsylvania**, **Rhode Island**, and **Vermont**.

   Drag these items into a new value group. This new group is automatically assigned the name **Value Group 1**.

7 Right-click **Value Group 1**, and select **Edit group name**. Enter `Northeast` for **Name** in the Edit Group Name window, and click **OK**.
8 In **Values of Vendor State/Province**, select the following values: **Illinois**, **Indiana**, **Iowa**, **Kansas**, **Michigan**, **Minnesota**, **Missouri**, **Nebraska**, **North Dakota**, **Ohio**, **South Dakota**, and **Wisconsin**.

   Drag these items into a new value group. This new group is automatically assigned the name **Value Group 1**.
9 Right-click **Value Group 1**, and select **Edit group name**. Enter `Midwest` for **Name** in the Edit Group Name window, and click **OK**.
10 In **Values of Vendor State/Province**, select the remaining 13 values: **Arizona**,**California**,**Colorado**,**Hawaii**,**Idaho**,**Montana**,**Nevada**,**New Mexico**,**Oregon**, **Utah**, **Washington**, **Wyoming**.

   Drag these items into a new value group. This new group is automatically assigned the name **Value Group 1**.
11 Right-click **Value Group 1**, and select **Edit group name**. Enter `West` for **Name** in the Edit Group Name window, and click **OK**.



12 Click **OK**.
13 Save the project.

Transform the Return Amount Measure

Because some of the models that you want to use assume that the data is normally distributed, you need to explore the **Order Amount Returned** measure.

1 From the **Data** pane, drag the **Order Amount Returned** measure onto the canvas.

   By default, a histogram is created. As you can see, the data is not normal, and it is skewed right.
2 From the **Objects** pane, drag a **Box Plot** onto the canvas below the histogram.
3 In the **Roles** pane for the box plot, specify **Vendor Type** for **Category** and **Order Amount Returned** for the **Measures**.
4 To enhance this visualization, make the following changes in the **Options** pane:

   - Change the value of **Box direction** to ▬→.
   - Change the value of **Outliers** to **Show Outliers**.

   From this plot, you notice that it is primarily the convenience stores and discount stores that are skewing your data with a lot of outliers. Department stores, kiosks, and all others are also skewed right, but not necessarily to the same degree. From this, you decide that a log transformation is appropriate to reduce overall skewness.
5 From the **Data** pane, click **Add**, and select **Add calculated item**.

6 In the **Name** field, enter `Order Amount Returned (Log)`.

By default, the calculated item type is set to **Numeric** with **Format: Comma**. These are correct for this new item.

7 In the Add Calculated Item window, select **Text** in the upper center of the window.

8 Enter the following expression into the code field: `'Order Amount Returned'n Log 10`, and click **OK**.

The **Order Amount Returned (Log)** item appears in the **Measure** list.

9 From the **Data** pane, drag the **Order Amount Returned (Log)** measure onto the right side of the canvas.

By default, a histogram is created. As you can see, the data is now much closer to a normal distribution

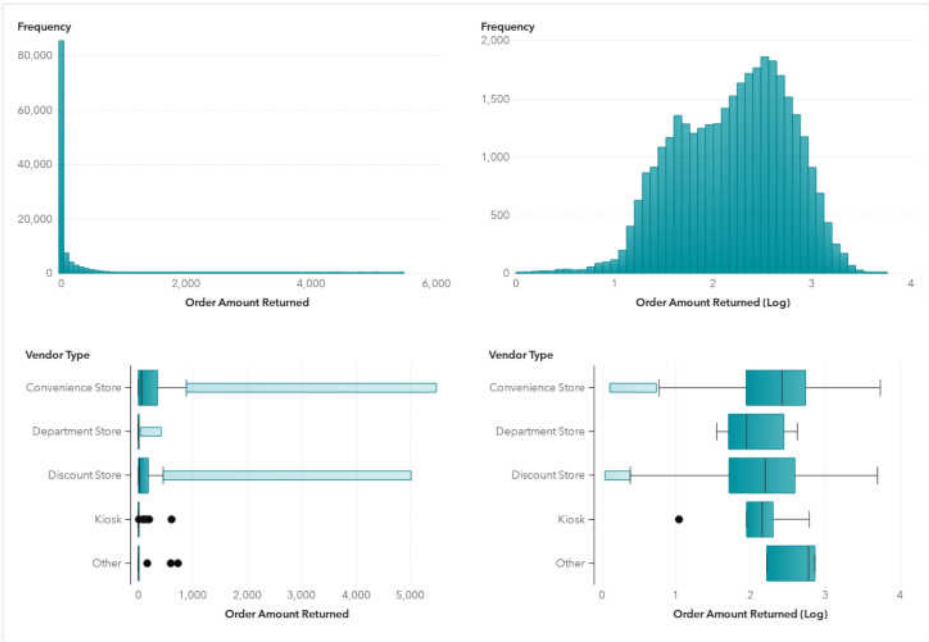
10 From the **Objects** pane, drag a **Box Plot** onto the canvas below the newest histogram.

11 In the **Roles** pane for the box plot, specify **Vendor Type** for **Category** and **Order Amount Returned (Log)** for the **Measures**.

12 To enhance this visualization, make the following changes in the **Options** pane:

- Change the value of **Box direction** to ▬→.
- Change the value of **Outliers** to **Show Outliers**.

These settings should match the previous box plot. Notice that the data is much closer to normal than before. Convenience stores and discount stores still present the most variance.



13 Save the project.

Data Exploration

About the Tasks in This Chapter

In this chapter, you explore the data with various explorations. The knowledge gained in these explorations helps you determine which inputs are going to be the best in your models.

Correlation Matrix

A correlation matrix, as the name implies, displays the correlation statistic either between all items in a set or between items in two separate sets. Complete the following steps to create a correlation matrix:

1. Click + to add a new page to the canvas.
2. From the **Objects** pane, drag a **Correlation Matrix** onto the canvas.
3. In the **Options** pane, set the following properties
   - Set **Show correlations** to **Between two sets of measures**.
   - Enable **Rotate axis labels**.
4. In the **Roles** pane, click **Add** under the **X axis** role, and select **Market Penetration**, **Order Amount**, **Order Amount Returned**, **Order Amount Returned (Log)**, **Order Distribution Cost**, **Order Marketing Cost**, **Order Product Cost**, **Order Sales Cost**, **Order Size**, **Sales Rep Rating**, **Sales Rep Vendors**, **Total Cost**, **Vendor Distance**, **Vendor Rating**, and **Vendor Satisfaction**. Click **OK**.
5. For the **Y axis** role, click **Add**, and select **Order Profit**. Click **OK**.
6. Notice which variables have the strongest correlation with **Order Profit**. For example, the strongest correlation is **Order Amount** at **0.7120**, and the weakest correlation is **Market Penetration** at **0.0100**.
7. Save the project.

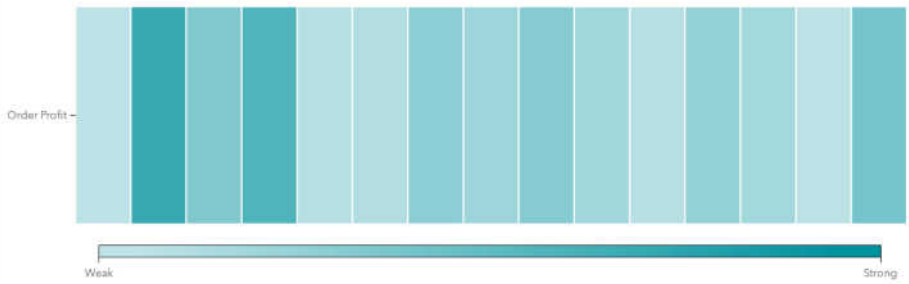
List Table

A list table lets you view aggregated information across many different measures at the same time. To create a list table, complete the following steps:

1. From the **Objects** pane, drag a **List Table** onto the canvas below the correlation matrix.
2. In the **Roles** pane, click **Add**, and select **Vendor Region**, **Vendor Type**, **Product Line**, **Order Amount**, **Order Amount Returned**, and **Vendor Satisfaction**.

   Notice that the **Vendor Satisfaction** values do not really make sense. The default aggregation statistic is sum. But, **Vendor Satisfaction** should use the mean instead.
3. In the **Data** pane, find **Vendor Satisfaction**. Click next to **Vendor Satisfaction**. Set the value of **Aggregation** to **Average**. Now, the list table values should make more sense.
4. Sort the list table descending by **Order Amount Returned**. Notice that the top of the list includes mostly discount stores with relatively low satisfaction ratings. At a glance, **Vendor Region**, **Product Line**, and **Order Amount** do not appear to display any pattern or bias.

   Based on this information, you want to further investigate **Vendor Type** and **Vendor Satisfaction**.



Vendor Region	Vendor Type	Product Line	Order Amount	Order Amount Returned ▾	Vendor Satisfaction
West	Discount Store	Game	1,064,183.03	483,422.39	52%
West	Discount Store	Figure	954,057.69	459,534.81	52%
Midwest	Discount Store	Game	668,793.71	374,175.51	50%
Midwest	Discount Store	Figure	655,808.31	372,328.66	50%
Northeast	Discount Store	Figure	725,401.44	367,406.67	52%
Northeast	Discount Store	Game	738,511.95	347,149.39	52%
South	Discount Store	Game	526,502.82	323,783.35	49%
Northeast	Discount Store	Bead	596,938.03	311,515.93	52%
West	Convenience Store	Bead	511,971.25	300,134.52	51%
South	Discount Store	Figure	455,787.88	280,970.97	49%
Midwest	Discount Store	Bead	530,830.69	278,476.63	50%
West	Discount Store	Plush	577,146.91	275,089.81	53%
Northeast	Convenience Store	Bead	545,988.09	263,600.02	53%
South	Discount Store	Bead	415,360.04	258,363.31	49%

5 Save the project.

Forecasting

About the Tasks in This Chapter

In this chapter, you create a time series plot of your company's profits and explore how they might change in the upcoming year.

Create a Time Series Forecast

1 Click + to add a new page to the canvas.

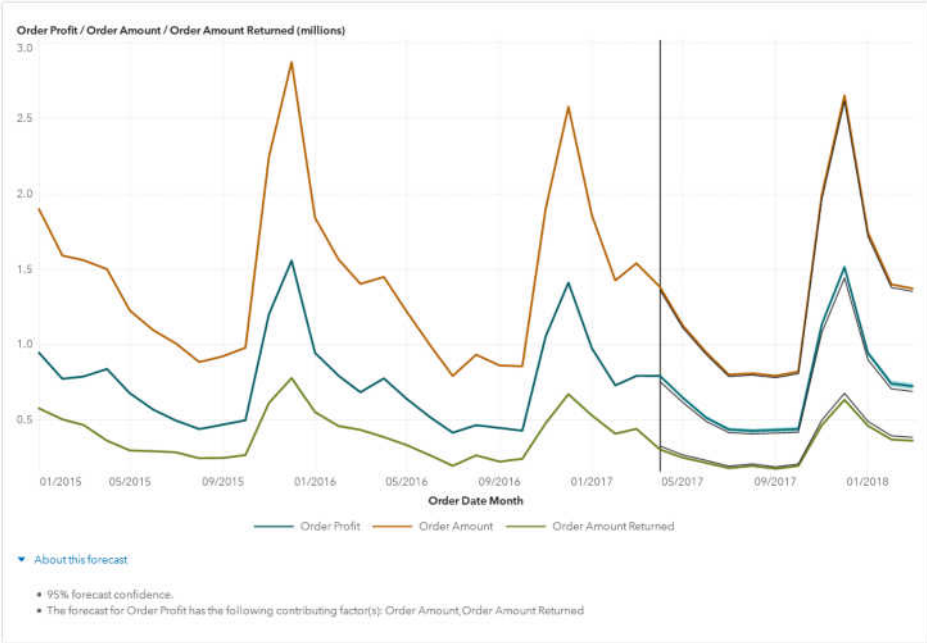
2 From the **Objects** pane, drag a **Time Series Plot** onto the canvas below the correlation matrix.

3 In the **Roles** pane, make the following changes:

- Under **Time axis**, click **Add**, and select **Order Date Month**.
- Under **Measure**, click **Add**, and select **Order Profit**. Click **OK**.
- Under **Measure**, right-click **Frequency**, and select **Remove Frequency**.
- Under **Forecast**, enable **Show forecast**.
- Under **Underlying factors**, click **Add**, and select **Order Amount** and **Order Amount Returned**. Click **OK**.

4 In the **Options** pane, for the **Forecast horizon** property, set the value to `12`.

The default forecast horizon is **6**, but because the data is aggregated monthly, and you want a full-year forecast, you should increase it to 12. Note that the time series takes into account the seasonality of your products.



5 Save the project.

What If Analysis / Goal Seeking

1 Make sure that the time series plot is selected. In the **Roles** pane, under **Forecast**, click the **What If** button.

A what-if analysis helps determine how the underlying factors need to change to reach your 5% growth goal.

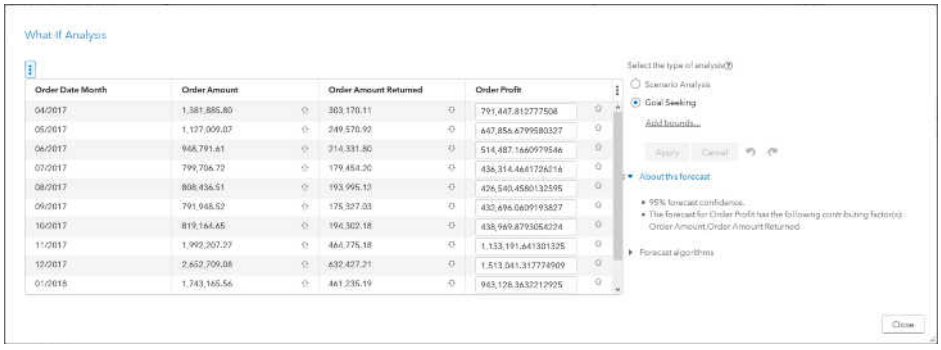
2 In the What-If Analysis window, change the type of analysis to **Goal Seeking**.

3 Click the ⋮ icon, and select **Set series values**.

The Set Series Values window enables you to adjust the forecast series based on a specified factor.

4 In the Set Series Values window, select **By percentage**, and enter 5 in the field. Click **OK**.

5 When you are back in the What-If Analysis window, click **Apply** under **Goal Seeking**. Note that **Order Amount** was adjusted up and **Order Amount Returned** was adjusted down to generate the 5% increased **Order Profit** prediction.



6 Click **Close**.

7 Save the project.

Modeling

About the Tasks in This Chapter

In this chapter, you create four statistical models to determine whether an order was returned. Two of the models are available from SAS Visual Statistics—the logistic regression model and the decision tree model. The other two models are available from SAS Visual Data Mining and Machine Learning—the forest model and the gradient boosting model. You compare these four models to determine a champion.

Specify the Partition Column

Your data set contains the **z Partition** category. This category is intended as a partition variable and contains only two values, **T** for training observations and **V** for validation observations. To create a partition column based on this category, complete the following steps:

1. In the **Data** pane, right-click on the **z Partition** category, and select **Set as partition column**.
2. In the Edit Partition Column window, ensure that the value of **Training data value** is **T** and **Validation data value** is **V**.
3. Click **OK**.

Create a Logistic Regression

Now that you have completed the feature engineering, data exploration, and forecasting tasks, you are ready to begin modeling your data. The first model that you create is a logistic regression on whether an order is returned.

1. Click + to add a new page to the canvas.
2. From the **Objects** pane, drag a **Logistic Regression** onto the canvas.
3. In the **Roles** pane of the logistic regression, make the following assignments:
   - Under **Response**, click **Add**, and select **Order Returned**.
   - Under **Continuous effects**, click **Add**, and select **Order Distribution Cost**, **Order Marketing Cost**, **Order Product Cost**, **Order Sales Cost**, **Vendor Distance**, **Vendor Rating**, and **Vendor Satisfaction**. Click **OK**.
   - Under **Classification effects**, click **Add**, and select **Product Line** and **Vendor Type**. Click **OK**.
4. In the **Options** pane of the logistic regression, enable **Use validation partition**. After doing so, the default model statistic is changed from the R-Square value to the Validation Misclassification rate.
5. To view the misclassification plot, right-click in the **Validation Lift** plot, and select **Switch to validation misclassification**. The overall validation misclassification rate is approximately 5.5%.
6. Right-click in the **Validation misclassification** plot, and select **Switch to validation ROC**. The ROC chart for this model rises steeply, and then quickly flattens.
7. In the **Options** pane, locate **Event level**, and click the **Change** button.
8. In the Select Event Level window, select **Y**, and click **OK**.
9. Notice the **Fit Summary** plot. This plot ranks the variables by their relative importance as measured by p-values. You can see that the most important variables in the model are **Vendor Satisfaction**, **Vendor Type**, and **Product Line**.

10 Save the project.

Create a Decision Tree

1 Click + to add a new page to the canvas.

2 From the **Objects** pane, drag a **Decision Tree** onto the canvas.

3 In the **Roles** pane of the decision tree, make the following assignments:

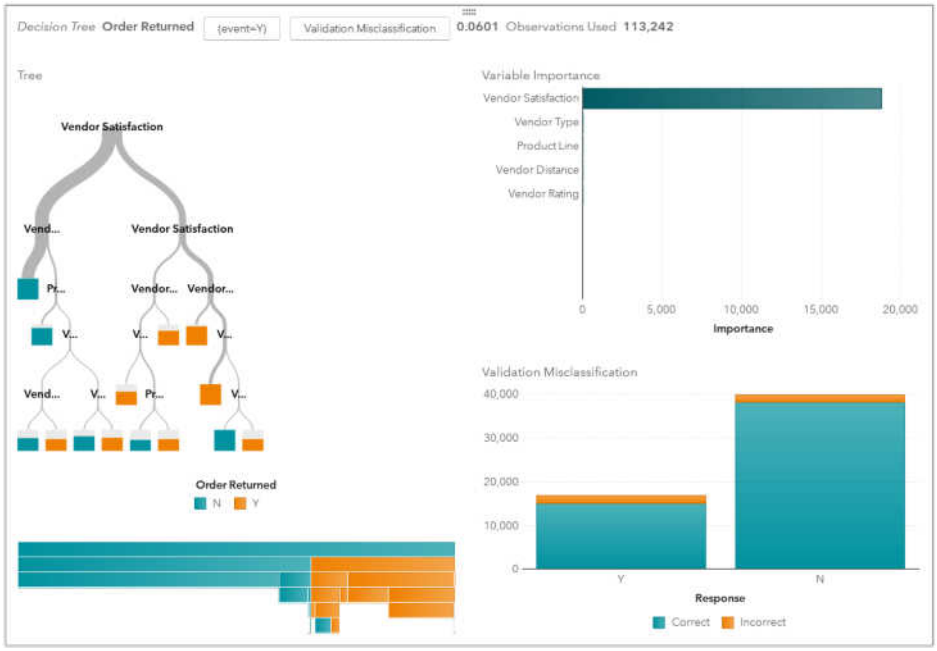
- Under **Response**, click **Add**, and select **Order Returned**.
- Under **Predictors**, click **Add**, and select **Product Line**, **Vendor Type**, **Order Distribution Cost**, **Order Marketing Cost**, **Order Product Cost**, **Order Sales Cost**, **Vendor Distance**, **Vendor Rating**, and **Vendor Satisfaction**. Click **OK**.

4 In the **Options** pane, enable **Use validation partition**. The overall validation misclassification rate is approximately 6%.

5 In the **Options** pane, locate **Event level**, and click the **Change** button.

6 In the Select Event Level window, select **Y**, and click **OK**.

7 In the **Variable Importance** plot, notice that **Vendor Satisfaction** is the most important predictor in the tree.



8 Save the project.

Create a Forest

A forest is a predictive model that combines the results of many different decision trees based on a random subset of the data and a random subset of predictors. To determine the predicted value of an observation, each tree in the forest gets to vote, and the majority vote is chosen. To create a forest, complete the following steps:

1 Click + to add a new page to the canvas.

2 From the **Objects** pane, drag a **Forest** onto the canvas.

3 In the **Roles** pane of the forest, make the following assignments:

- Under **Response**, click **Add**, and select **Order Returned**.
- Under **Predictors**, click **Add**, and select **Product Line**, **Vendor Type**, **Order Distribution Cost**, **Order Marketing Cost**, **Order Product Cost**, **Order Sales Cost**, **Vendor Distance**, **Vendor Rating**, and **Vendor Satisfaction**. Click **OK**.

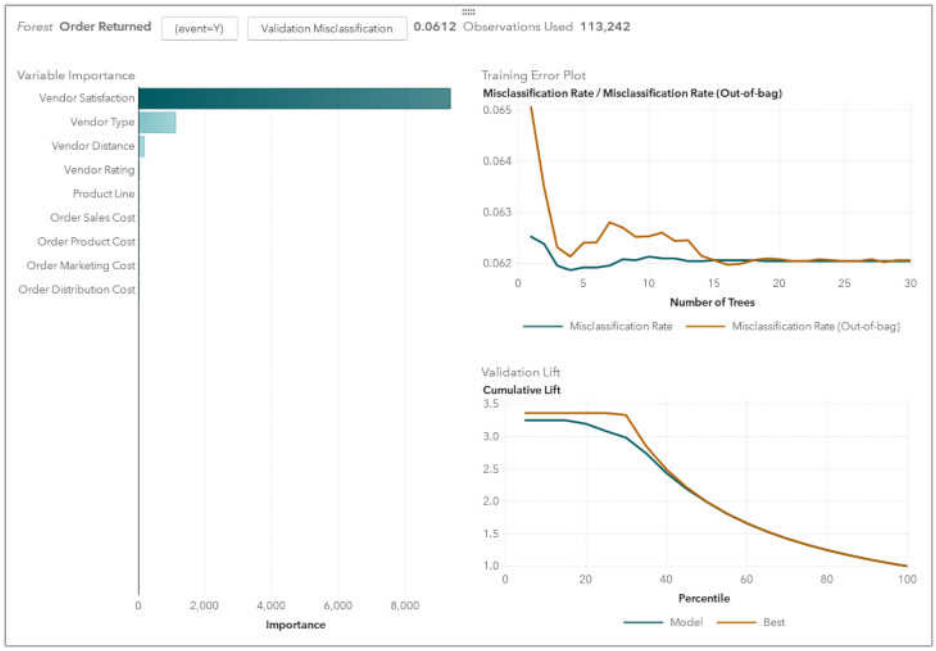
4 In the **Options** pane, enable **Use validation partition**. The overall validation misclassification rate is approximately 19%.

5 Notice that the **Training Error Plot** converges fairly quickly, enabling you to train fewer trees for similar results. In the **Options** pane, set the value of **Number of trees** to `30`.

6 In the **Options** pane, locate **Event level**, and click the **Change** button.

7 In the Select Event Level window, select **Y**, and click **OK**.

8 In the **Variable Importance** plot, notice that **Vendor Satisfaction** is the most important predictor in the tree.



9 Save the project.

Create a Gradient Boosting

Like a forest, gradient boosting combines the results of several decision trees. The key difference is that when a new decision tree in a gradient boosting is created, it uses the results of the previous decision trees to better tune the results. To create a gradient boosting model, complete the following steps:

1 Click + to add a new page to the canvas.

2 From the **Objects** pane, drag a **Gradient Boosting** onto the canvas.

3 In the **Roles** pane of the gradient boosting, make the following assignments:

- Under **Response**, click **Add**, and select **Order Returned**.
- Under **Predictors**, click **Add**, and select **Product Line**, **Vendor Type**, **Order Distribution Cost**, **Order Marketing Cost**, **Order Product Cost**, **Order Sales Cost**, **Vendor Distance**, **Vendor Rating**, and **Vendor Satisfaction**. Click **OK**.

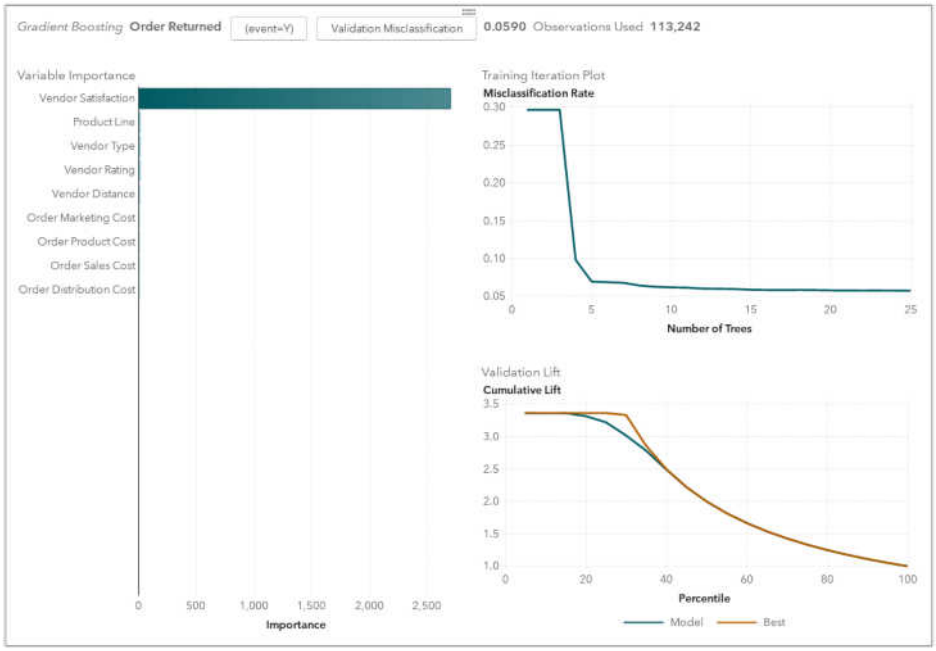
4 In the **Options** pane, enable **Use validation partition**. The overall validation misclassification rate is approximately 17%.

5 Notice that the **Training Iteration Plot** has not necessarily converged. In the **Options** pane, set the value of **Number of trees** to `25`.

6 In the **Options** pane, locate **Event level**, and click the **Change** button.

7 In the Select Event Level window, select **Y**, and click **OK**.

8 In the **Variable Importance** plot, notice that **Vendor Satisfaction** is the most important predictor in the tree.



9 Save the project.

Create a Model Comparison

Now that you have four competing predictive models, you can add a model comparison to determine a champion model.

1 Click + to add a new page to the canvas.

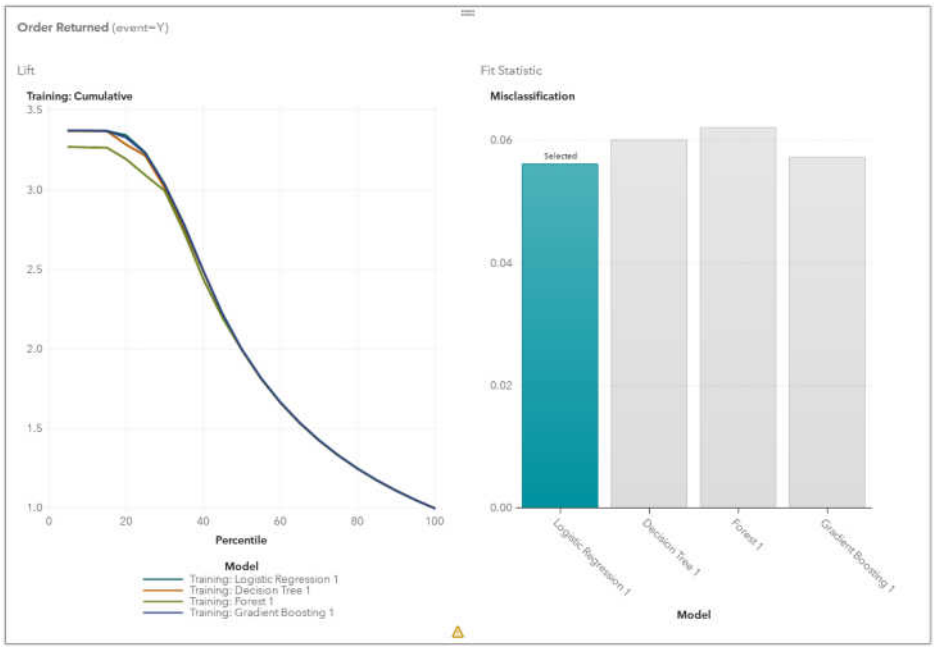
2 From the **Objects** pane, drag a **Model Comparison** onto the canvas.

3 In the Add Model Comparison window, review the following fields:

- **Partition** should contain **Use validation partition**.
- **Response** should contain **Order Returned**.
- **Event level** should contain **Y**.

4 Under **Available models**, you should see the four models that you previously created. Enable **Select all**, and click **OK**.

5 Because you modeled a category response, the default fit statistic is the misclassification rate. In the **Fit Statistic** plot, notice that **Logistic Regression 1** has the best misclassification rate. This is your champion model.



6 Save the project.

Results

About the Tasks in This Chapter

In this chapter, you export the champion model from the previous chapter. Then, you further explore the results of that model to determine where you can reduce returned orders and increase company profits.

Export the Champion Model

1 Navigate back to the page that contains the logistic regression.
2 To export the model, right-click in the **Fit Summary** window, and select **Export model**.
3 Review the code in the Export Model window, and click **Export**. If your browser asks you whether to open or save the exported file, save the file.

   This file contains the SAS code necessary to score new data based on the logistic regression model that you just created.
4 Save the project.

Derive Predicted Values

1 To derive the predicted values for the model, right-click in the **Fit Summary** window, and select **Derive predicted values**.
2 In the New Prediction Variables window, review the **Name** field for both **Predicted values** and **Probability values**. Click **OK**.

   The value of the **Predicted values** variable contains the model's prediction for each observation of whether the order was returned. Possible values of this variable are **Yes**, **No**, and missing. Similarly, the value of the **Probability values** variable is the probability assigned by the model that the given observation is a returned order. Possible values of this variable are real numbers between 0 and 1.

3 In the **Data** pane, there are two new items, **Predicted: Order Returned** and **Probability: Order Returned=Y**. These items are available as inputs to other models and as data items in report visualizations.
4 Save the project.

Create the Count Variables

1 In the **Data** pane, right-click **Order Returned**, and select **Create calculation for data item**.
2 For **Type**, select **Count**.
3 In the Create Calculation window, enter `Order Returned (#)` in the **Name** field.
4 Click **OK**.
5 In the **Data** pane, right-click **Predicted: Order Returned**, and select **Create calculation for data item**.
6 For **Type**, select **Count**.
7 In the Create Calculation window, enter `Predicted Order Returned (#)` in the **Name** field.
8 Click **OK**.
9 Save the project.

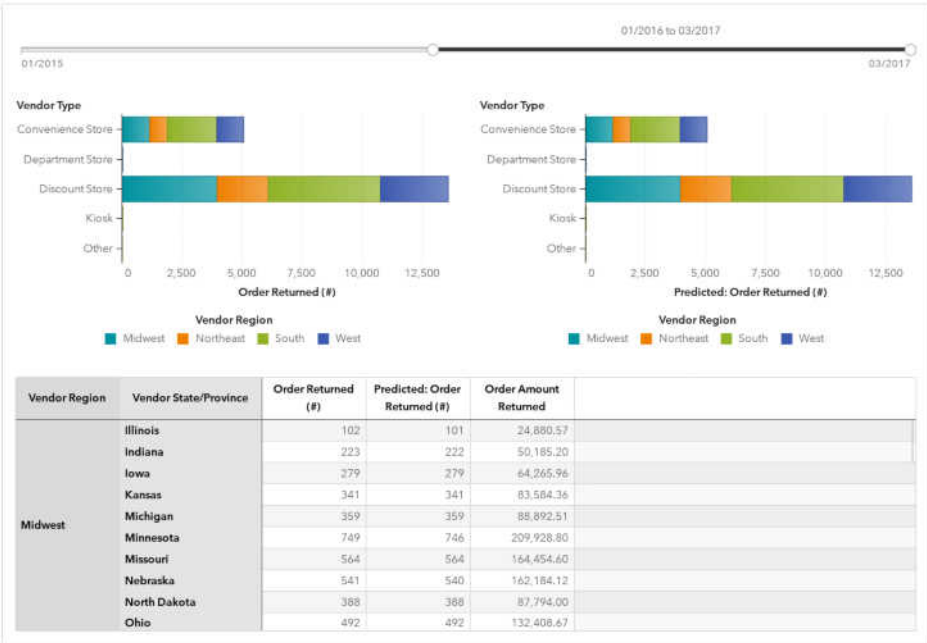
Review the Model Results

1 Click + to add a new page to the canvas.
2 From the **Objects** pane, drag a **Slider** onto the canvas.
3 In the **Roles** pane of the slider, under **Measure/Date**, click **Add**, and select **Order Date Month**.

   You will use this slider in later steps to interactively filter the data displayed in other objects.
4 From the **Objects** pane, drag a **Bar Chart** onto the canvas below the slider.
5 In the **Roles** pane of the bar chart, make the following assignments:
   - Under **Category**, click **Add**, and select **Vendor Type**.
   - Under **Measure**, click **Add**, and select **Order Returned (#)**. Click **OK**. Right-click **Frequency**, and select **Remove Frequency**.
   - Under **Group**, click **Add**, and select **Vendor Region**.
6 In the **Options** pane, under Grouping style, click ..
7 In the **Filters** pane of this bar chart, click **Add**, and select **Order Returned**.
8 Under **Order Returned**, deselect **N**.
9 From the **Objects** pane, drag a second **Bar Chart** onto the canvas below the slider and to the right of the first bar chart.
10 In the **Roles** pane of this new bar chart, make the following assignments:
   - Under **Category**, click **Add**, and select **Vendor Type**.
   - Under **Measure**, click **Add**, and select **Predicted: Order Returned (#)**. Click **OK**. Right-click **Frequency**, and select **Remove Frequency**.
   - Under **Group**, click **Add**, and select **Vendor Region**.
11 In the **Options** pane, under Grouping style, click ..
12 In the **Filters** pane of this bar chart, click **Add**, and select **Order Returned**.
13 Under **Order Returned**, deselect **N**.
14 From the **Objects** pane, drag a **Crosstab** onto the canvas below the two bar charts.
15 In the **Roles** pane of the crosstab, make the following assignments:
   - Under **Rows**, click **Add**, and select **Vendor Region** and **Vendor State/Province**. Click **OK**.
   - Under **Measures**, click **Add**, and select **Order Returned (#)**, **Predicted: Order Returned (#)**, and **Order Amount Returned**. Click **OK**.
16 In the **Filters** pane of the crosstab, click **Add**, and select **Order Returned**.
17 Under **Order Returned**, deselect **N**.
18 Now, select the slider at the top of the canvas. In the **Actions** pane, click **Add**, and select **Add filter**.

19 In the Add Filter Action window, select **Select all**. Click **OK**.

On this page, you can compare how many orders the logistic regression model predicts as returned against the true number of returns over a custom time range. After reviewing these results, you are confident in your logistic regression model's ability to classify whether an order is returned.



Vendor Region	Vendor State/Province	Order Returned (#)	Predicted: Order Returned (#)	Order Amount Returned
Midwest	Illinois	102	101	24,880.57
	Indiana	223	222	50,185.20
	Iowa	279	279	64,265.96
	Kansas	341	341	83,584.36
	Michigan	359	359	88,892.51
	Minnesota	749	746	209,928.80
	Missouri	564	564	164,454.60
	Nebraska	541	540	162,184.12
	North Dakota	388	388	87,794.00
	Ohio	492	492	132,408.67

20 Save the project.

Cluster the Model Results

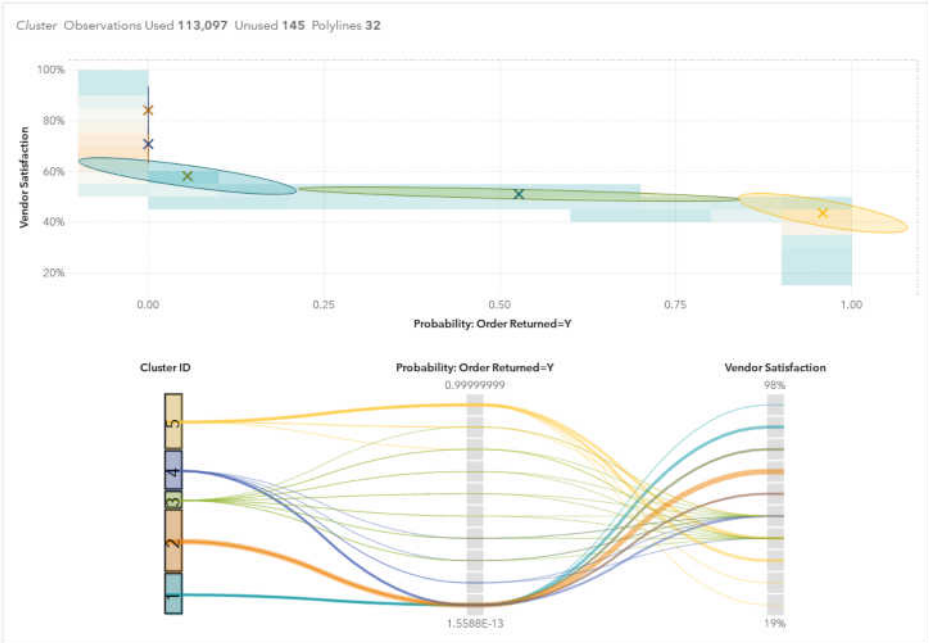
Now that you are confident in your model's ability to predict whether an order was returned, you want to cluster the results to further identify what types of orders are returned. SAS Visual Analytics provides two primary tools for model segmentation—k-means clustering and decision trees.

1 Click + to add a new page to the canvas.

2 From the **Objects** pane, drag a **Cluster** onto the canvas.

3 In the **Roles** pane, under **Variables**, click **Add**, and select **Probability: Order Returned=Y** and **Vendor Satisfaction**. Click **OK**.

4 Notice that one of the clusters includes observations where the value of **Probability: Order Returned=Y** is very high and the value of **Vendor Satisfaction** is very low. These are the observations that you want to target to reduce returns.



5 Right-click in the **Cluster** window, and select **Derive a cluster ID variable**.
6 In the New Cluster ID window, enter `Cluster Order Returned`. Click **OK**.

Explore the Clustered Results

In this section, you create a visualization that enables you to interactively view and filter your results by vendor region.

1 Click + to add a new page to the canvas.
2 From the **Objects** pane, drag a **Button Bar** onto the canvas.
3 In the **Roles** pane, under **Category**, click **Add**, and select **Vendor Region**.
4 From the **Objects** pane, drag a **Slider** onto the canvas to the right of the button bar.
5 In the **Roles** pane for the slider, under **Measure/Date**, click **Add**, and select **Order Date Month**.
6 In the **Data** pane, select **Vendor State/Province**, and click .
7 Under **Classification**, select **Geography**.
8 In the Geography Classification for Vendor State/Province window, select **US State Names** in the **Geography** field. Click **OK**.
9 From the **Objects** pane, drag a **Geomap** onto the canvas below the button bar and the slider.
10 In the **Roles** pane for the geomap, under **Category**, click **Add**, and select **Vendor State/Province**.
11 Under **Size**, click **Frequency**, and replace it with **Order Amount Returned**.
12 From the **Objects** pane, drag a **Butterfly Chart** to the right of the geomap.
13 In the **Roles** pane, under **Category**, click **Add**, and select **Cluster Order Returned**.
14 Under **Measure (bar)**, click **Add**, and select **Order Amount Returned**.
15 Under **Measure (bar 2)**, click **Add**, and select **Order Amount**.
16 In the **Objects** pane, drag a **Bar Chart** onto the canvas below the geomap and the butterfly chart.
17 In the **Roles** pane of the bar chart, make the following assignments:

- Under **Category**, click **Add**, and select **Vendor Type**.
- Under **Measure**, click **Add**, and select **Frequency** and **Order Amount Returned**.

18 On the canvas, select the button bar. In the **Actions** pane of the button bar, click **Add**, and select **Add filter**.
19 In the Add Filter Action window, select **Geo Map 1**, **Butterfly Chart 1**, and **Bar Chart 3**. Click **OK**.
20 On the canvas, select the slider. In the **Actions** pane of the slider, click **Add**, and select **Add filter**.
21 In the Add Filter Action window, select **Geo Map 1**, **Butterfly Chart 1**, and **Bar Chart 3**. Click **OK**.
22 On the canvas, select the butterfly chart. In the **Actions** pane of the butterfly chart, click **Add**, and select **Add linked selection**.

23 In the Add Linked Selection Action window, select **Select all**. Click **OK**.
24 In the butterfly chart, select the cluster that contains the most returned orders. The corresponding data values are selected in both the geomap and the bar chart. Note the cluster number with the most returned orders. You will need this cluster number in a future step.

Review the charts for each region. Notice that in every region, the amount returned is dominated by discount stores and convenience stores, and that a significant portion of these returns is contained in a single cluster. The vendors in this cluster are those that you want to target to reduce returns.

Armed with this information, you can work with your sales team to target those vendors that are not satisfied and are contributing to a financial loss.
25 Save the project.

Identify the Vendors for Investigation

1 Click + to add a new page to the canvas.
2 From the **Objects** tab, drag a **Crosstab** onto the canvas.
3 In the **Roles** pane, make the following assignments:

- Under **Columns**, click **Add**, and select **Cluster Order Returned**.
- Under **Rows**, click **Add**, and select **Vendor** and **Vendor Region**. Click **OK**.
- Under **Measures**, click **Add**, and select **Order Amount Returned**, **Order Amount**, **Order Profit**, and **Vendor Satisfaction**. Click **OK**.

4 On the **Filters** tab, click **Add**, and select **Cluster Order Returned**.
5 Select the cluster that you noted earlier as containing the most returned orders.
6 Use the crosstab to identify and investigate individual vendors and their contributions to your profits. Ideally, you can take this list to your sales team and work with them to better engage and satisfy these vendors, thus leading to fewer returned orders.

Cluster Order Returned		1			
Vendor	Sales Rep	Order Amount Returned	Order Amount	Order Profit	Vendor Satisfaction
vend:000008C1	srep:0000004Z	2,926.90	2,926.90	-647.20	38%
vend:000008C5	srep:0000004Z	2,139.81	2,139.81	-520.85	38%
vend:000008C7	srep:0000004Z	3,030.90	3,030.90	-677.14	43%
vend:000008C8	srep:0000004Z	2,240.70	2,240.70	-510.67	36%
vend:000008CA	srep:0000004Z	2,368.90	2,368.90	-397.04	50%
vend:000008Z6	srep:0000005B	8,132.33	9,369.50	-1,236.01	47%
vend:000008Z8	srep:0000005B	5,420.93	5,420.93	-1,171.98	44%
vend:000008ZO	srep:0000005B	6,694.29	6,694.29	-1,364.86	40%
vend:000008ZQ	srep:0000005B	4,694.68	4,694.68	-703.89	41%
vend:000008Z5	srep:0000005B	5,233.21	5,233.21	-851.05	41%
vend:00000934	srep:0000005B	3,886.77	3,886.77	-1,018.09	44%
vend:00000951	srep:0000005E	250.39	745.85	284.00	48%
vend:00000955	srep:0000005E	2,512.96	2,512.96	-564.18	46%
vend:00000956	srep:0000005E	2,454.73	2,566.53	-251.88	48%
vend:00000957	srep:0000005E	3,571.39	3,622.45	-515.06	48%
vend:00000958	srep:0000005E	1,719.46	1,719.46	-375.17	38%
vend:0000095A	srep:0000005E	2,042.45	2,044.04	-422.89	48%
vend:0000095J	srep:0000005E	2,252.97	2,252.97	-766.67	40%
vend:0000095N	srep:0000005E	2,815.17	2,815.17	-614.08	42%
vend:00000961	srep:0000005E	473.53	473.53	-196.60	41%
vend:00000963	srep:0000005E	1,511.43	1,511.43	-571.36	46%
vend:00000967	srep:0000005E	6,091.98	6,091.98	-1,451.75	41%
vend:00000968	srep:0000005E	154.66	154.66	-48.46	48%
vend:0000096C	srep:0000005E	3,419.78	3,419.78	-828.69	43%
vend:0000096F	srep:0000005E	4,808.13	4,808.13	-1,063.80	45%
vend:0000096G	srep:0000005E	3,017.55	3,017.55	-918.16	42%

7 Save the project.