

Interactive Modeling in SAS® Visual Analytics

Don Chapman, SAS Institute Inc.

ABSTRACT

SAS® Visual Analytics has two add-on offerings, SAS® Visual Statistics and SAS® Visual Data Mining and Machine Learning, that provide knowledge workers and data scientists an interactive interface for data partition, data exploration, feature engineering, and rapid modeling. These offerings are powered by the SAS® Viya™ platform, thus enabling big data and big analytic problems to be solved. This paper focuses on the steps a user would perform during an interactive modeling session.

INTRODUCTION

This paper illustrates how the use of the highly interactive, visual SAS Visual Data Mining and Machine Learning offering will not only make your data problems manageable but also engaging. This offering is composed of capabilities that range from data preparation to programmatic access to advanced machine learning in your language of choice. Each capability in the offering would require its own paper to do it justice, so this paper focuses on integrated data exploration, reporting, and analytical modeling. SAS® Visual Analytics allows collaboration between business analysts, citizen data scientists, and data scientists. This is important because the data scientists apply analytical methods to business data to create insights that drive the business direction.

This paper and the associated presentation will focus on the case study of a day in the life of a data scientist who needs to solve a business problem quickly. How do they acquire the data and get it prepared for modeling? How do they explore the data to understand its characteristics? How do they generate and compare models? How do they document those insights and apply them to solving a business problem?

THE BUSINESS PROBLEM

Picture yourself pulling into Starbucks on the way to work. Your manager calls to tell you that she has to pitch a plan to increase profits by 5% this year. She needs you to put together her presentation for an executive meeting tomorrow afternoon. This unfortunately is how too many of your days start. An unplanned request just became your top priority.

Time for a little background. You are a data scientist at the Insight Toy Company who works closely with a vice president of sales and marketing. Sales to our vendors have been slowly declining for the last couple of years and your manager has been asked to increase profits in her organization by 5% this year. It's time to get Insight Toy back on track.

THE PLAN

The first thing you need to do is come up with a plan for tackling this challenging task. Fortunately for you, Insight Toy has been collecting data for several years on all aspects of the business. They also have a great IT department who prepares the data for its analysts and data scientists. A quick inventory of the corporate data shows that you have access to the last two years of sales data. This data includes information on what products are sold to which vendors, the costs associated with the order, along with some metrics about the sales representative and the vendor.

Step one, come up with a plan. You decide to follow the tried-and-true strategy of:

1. Review the data and make a quick exploratory pass over the most relevant variables to understand their characteristics and relationships
2. Feature engineering
3. Start generating models and reviewing their results

4. Compare your models, and come up with a champion
5. Validate your model and apply it
6. Come up with a couple of potential solutions and present them to your manager

REVIEW AND EXPLORE THE DATA

Exploring the data is an important step in understanding the relationships within. A quick pass over the data and you see that you have the entire order history for every vendor dating back to January 1st, 2015. The first task you tackle is to look at the shape and characteristics of Order Profit, your response variable.

Next you want to see if there are any linear correlations between Order Profit and other variables you think contribute to Insight Toy's profit. You create a page and add a correlation matrix, as shown in Figure 1, to investigate the relationships. It reveals that Order Amount has a strong correlation as you would expect. Two other variables, Amount Returned and Vendor Satisfaction have a moderate correlation. You immediately document your findings by adding a comment to the report stating your observations.

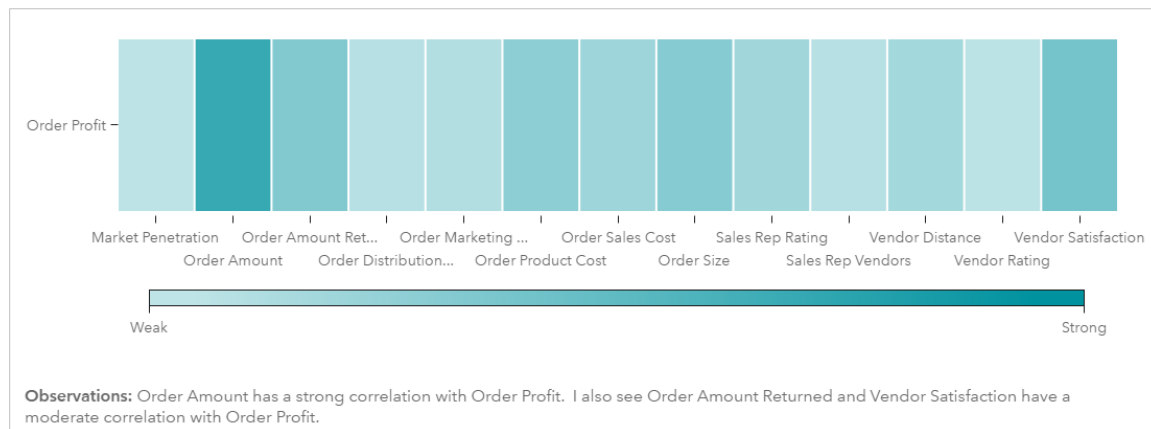


Figure 1. Correlation of Order Profit to Key Variables

The advantage you have in solving today's challenge is that you are using SAS Visual Analytics. This application has integrated data exploration, modeling, and reporting capabilities in a highly visual and interactive user interface. What else can the data tell you?

One the same page you add a List Table. A trusty table can convey a lot of information, especially when it aggregates the data for you. Figure 2 shows the aggregated list table you created.

Vendor Region	Vendor Type	Product Line	Order Amount	Order Amount Returned ▼	Vendor Satisfaction
West	Discount Store	Game	\$1,064,183	\$483,422	52%
West	Discount Store	Figure	\$954,058	\$459,535	52%
Midwest	Discount Store	Game	\$668,794	\$374,176	50%
Midwest	Discount Store	Figure	\$655,808	\$372,329	50%
Northeast	Discount Store	Figure	\$725,401	\$367,407	52%
Northeast	Discount Store	Game	\$738,512	\$347,149	52%
South	Discount Store	Game	\$526,503	\$323,783	49%
Northeast	Discount Store	Bead	\$596,938	\$311,516	52%
West	Convenience Store	Bead	\$511,971	\$300,135	51%
South	Discount Store	Figure	\$455,788	\$280,971	49%
Midwest	Discount Store	Bead	\$530,831	\$278,477	50%
West	Discount Store	Plush	\$577,147	\$275,090	53%

Observation: When you sort Order Amount Returned descending you see Discount Store vendors with a low Vendor Satisfaction are at the top of the list.

Figure 2. Returns List

You quickly see that the Vendor Type and Vendor Satisfaction need further investigation. The eyeball test shows that Discount Stores have the highest dollar amount for returns and you also see that vendor satisfaction is low for the vendors making returns.

On the next page you pull together several charts to quickly visualize the data. These charts, as seen in Figure 3, show you that convenience stores are also troublesome with respect to returns. They also show you that Product Line does not appear to be related to the returns.

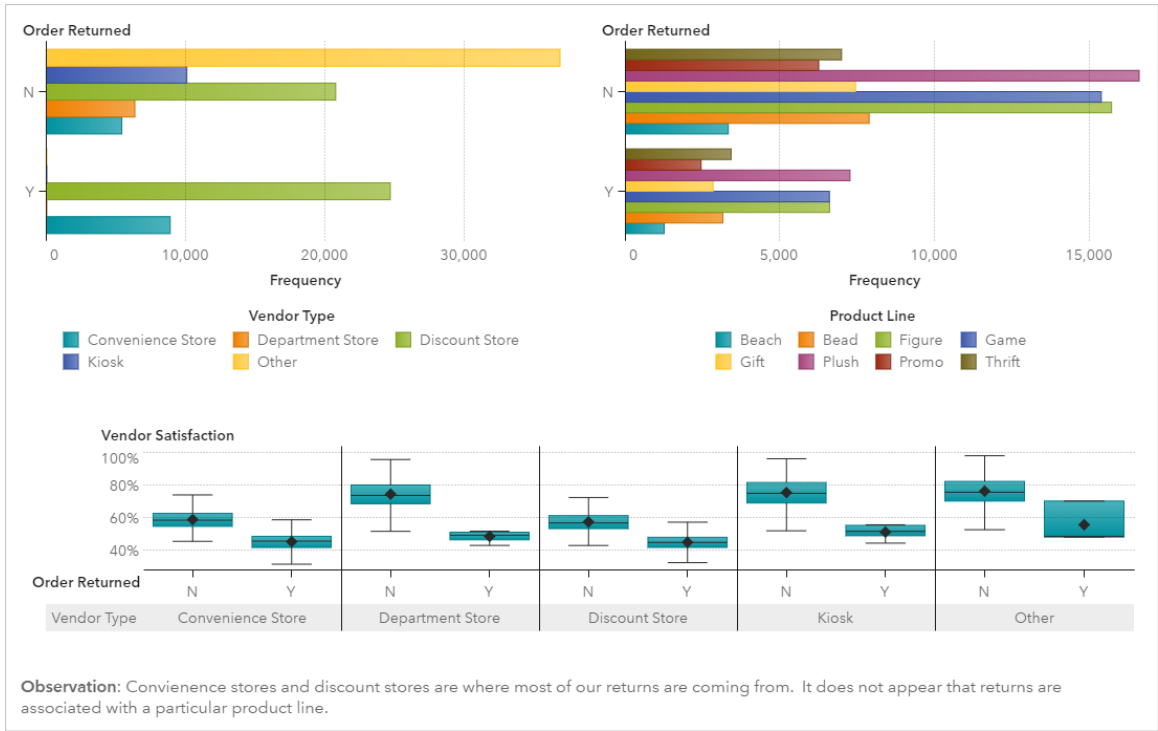


Figure 3. Returns Charts

You now understand the data better and have a good idea that decreasing the number of returned orders will help the bottom line.

Next you want to see what it will take to increase profits by 5%. A quick forecast, as shown in Figure 4, shows you that going after those returns will help Insight Toy's bottom line. You are happy to see the forecasting algorithm takes into account the seasonality of your products.

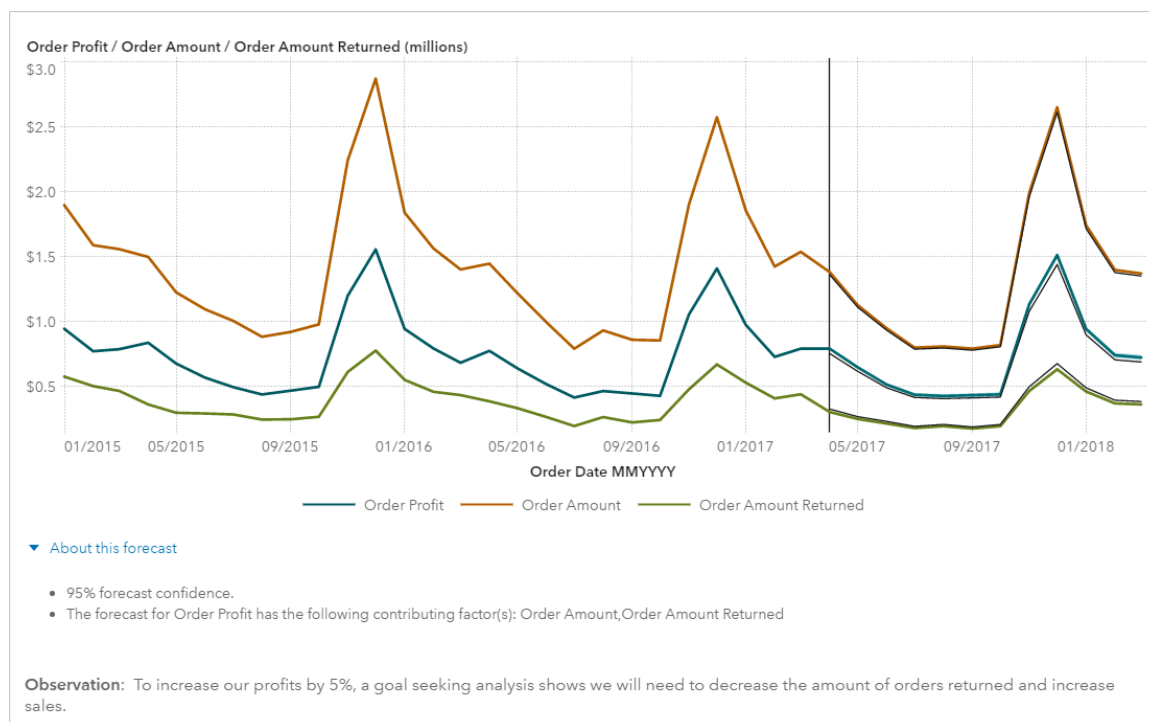


Figure 4. Profit Forecast

You used what-if analysis, specifically goal seeking as shown in Figure 5, to create this forecast. You can clearly see that Order Amount, which is how much our sales team is selling, needs to increase slightly in addition to the decrease in the amount of orders returned.

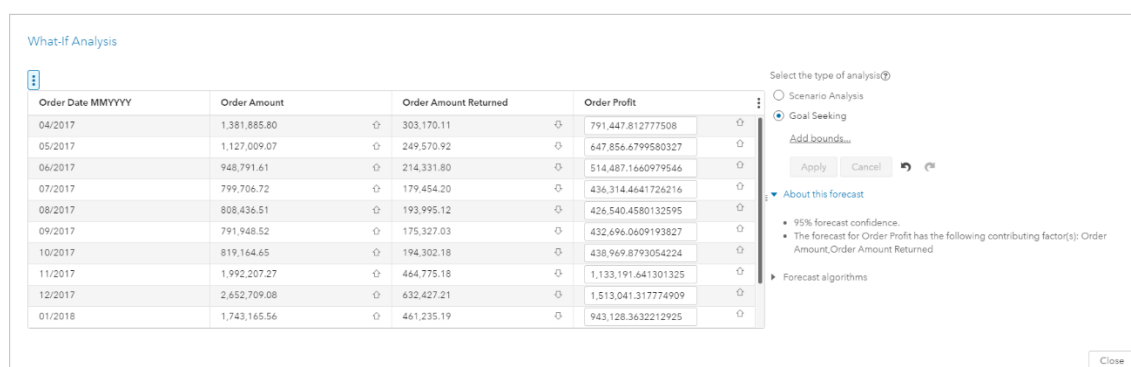


Figure 5. Goal Seeking

FEATURE ENGINEERING

As a data scientist you need to engineer features using your domain knowledge of Insight Toy and the problem at hand. SAS Visual Statistics and SAS Visual Data Mining and Machine Learning support traditional feature engineering such as segmentation. Calculations and custom categories are two features you can interactively create using a drag-and-drop interface or by editing code. You can create calculations based on simple math, for example here is the code for Order Profit:

```
( 'Order Amount'n - 'Order Amount Returned'n ) - 'Order Total Cost'n
```

You can also create calculations with conditional statements, for example Figure 6 shows the Vendor Active calculation in the calculation editor:

The screenshot shows the 'Edit Calculated Item' dialog box. The 'Name' field is set to 'Vendor Active'. The 'Format' is set to '\$'. The 'Visual' radio button is selected. The 'Character' dropdown is set to 'Character'. The 'Operators' tab is active, and the 'Boolean' category is expanded. The 'IF...ELSE' operator is selected. The calculation logic is displayed as follows:

```
IF Vendor Date Ended Missing  
RETURN " Y "  
ELSE " N "
```

The 'Messages (0)' section is empty. The 'OK' and 'Cancel' buttons are at the bottom right.

Figure 6. Calculation Editor

You also have the power to create ad-hoc hierarchies on-the-fly, duplicate a variable and change its format or aggregation, and even convert a measure to a category. The Order Date MMYYYY data item shown in Figure 7 was created by duplicating the Order Date data item and changing its format.

The screenshot shows the 'Data Item' editor. The 'Name' field is set to 'Order Date MMYYYY'. The 'Format' is set to 'MMYYYY'. The 'Distinct values' are 27. The 'Data source name' is 'OrderDate'. The 'Used by' section lists the following items:

- Time Series Plot 1
- Slider 1
- Slider 2
- Button Bar 2
- Slider 3

The 'Order Date MMYYYY' data item is selected. Other data items listed include 'Order Day of Week - 5', 'Order Month of Year - 12', 'Order Note - 89K', and 'Order Returned - 2'.

Figure 7. Data Item

Some algorithms, such as a linear regression, assume the data comes from a normal distribution. You take a quick look at the shape and distribution of your data since you are interested in Order Amount Returned. The two graphs on the left side of Figure 8 show that the Order Amount Returned variable is right-skewed.

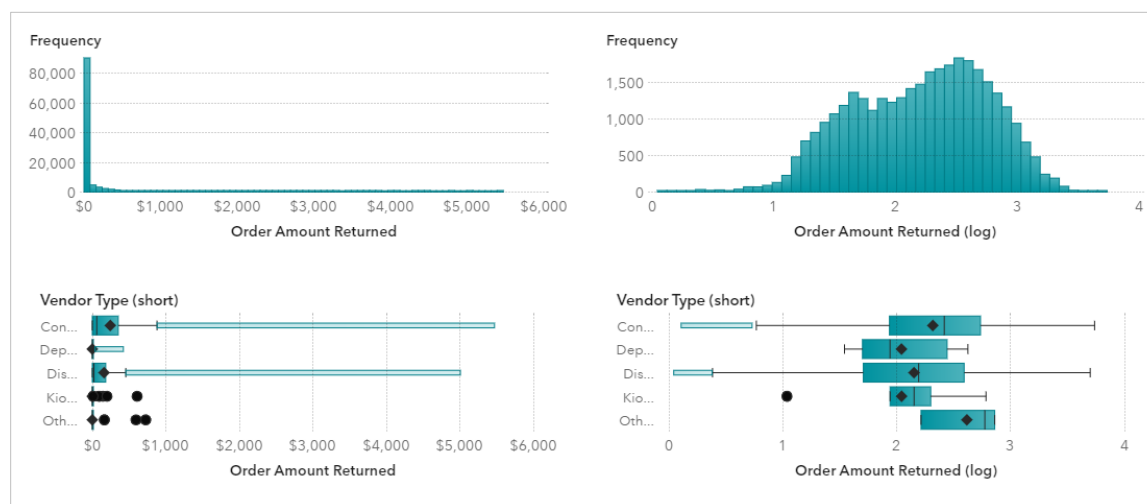


Figure 8. Variable Transformation

A log transformation can easily be created to reduce skewness. The code for the Order Amount Returned (log) calculation is:

```
( 'Order Amount Returned'n Log 10 )
```

The two graphs on the right side of Figure 8 show that the Order Amount Returned (log) variable follows a more normal distribution.

All calculations are dynamically constructed when the report is opened. This means you do not need to save a copy of the data for every report you create.

MODEL, MODEL, MODEL

Now it is time to start modeling. You know you have a binary response of either Y(es) or N(o) for Order Returned, so good candidate models are logistic regression, decision tree, forest, gradient boosting, neural network, and support vector machine. You want to model the vendors who have an event of Y; they are the ones returning orders.

The corporate data source has a partition variable that will allow you to train your models against a subset of the data and validate it against the rest of the data. By using partitioning, you will generate the best models without overtraining.

You decide to start out with a tried-and-true logistic regression model using data on the four costs associated with the order, information about the vendor, and the product line as your effect variables. In less than a minute you interactively add the model to the report, assign the response and effect variables, and configure modeling options.

The model shown in Figure 9 looks good. You can see a summary of how well the effect variables fit, the distribution of your residuals, and the model's misclassification chart. While the default statistic is the validation misclassification, you can also look at a number of other statistics such as AIC or R-Square.

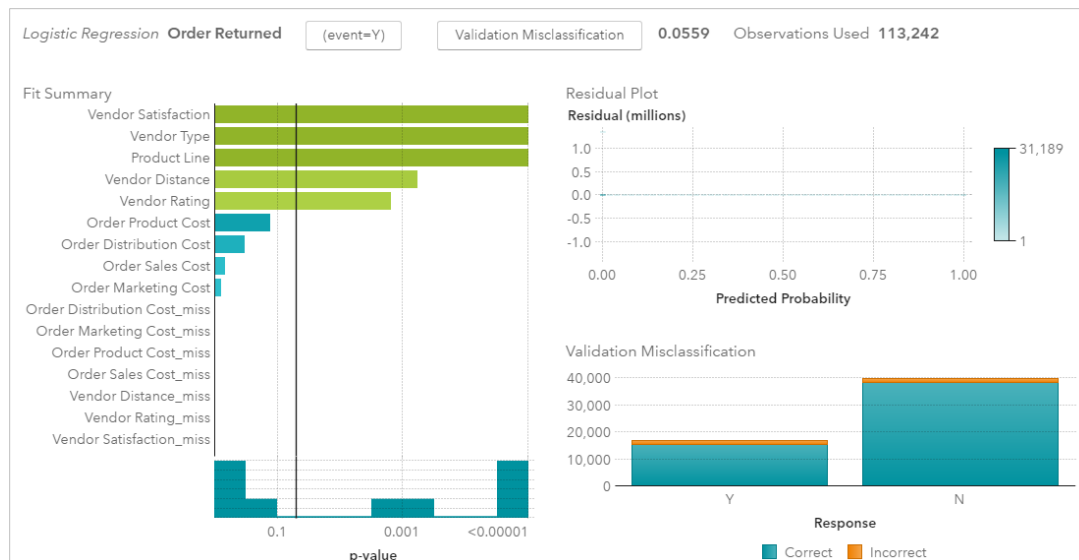


Figure 9. Logistic Regression

With a click of the mouse, you easily switch the validation misclassification chart to a validation lift chart, as shown in Figure 10, and then a validation ROC chart, as shown in Figure 11.

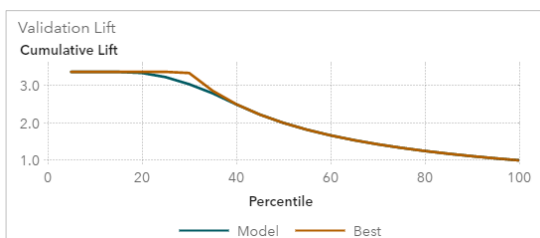


Figure 10. Logistic Regression Lift Chart

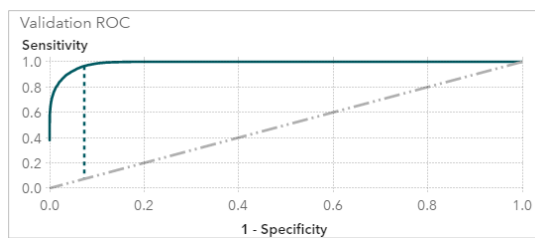


Figure 11. Logistic Regression ROC Chart

You use the same response and effects / target variables for creating additional models. On individual pages you create a decision tree model, forest model, gradient boosting model, neural network model, and support vector machine model. Each of these models is helping you predicted whether a vendor will return an order. The pages containing each of these models are shown in Figure 12 - Figure 16.

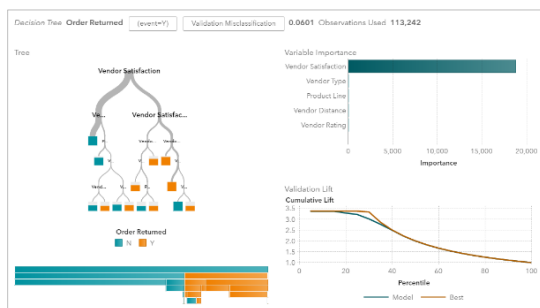


Figure 12. Decision Tree

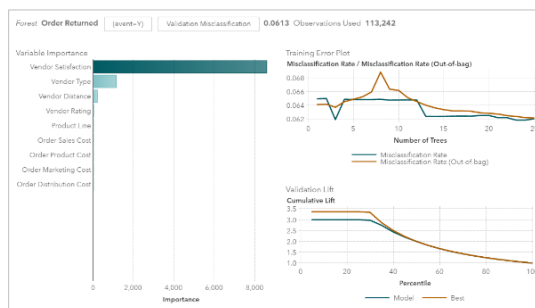


Figure 13. Forest

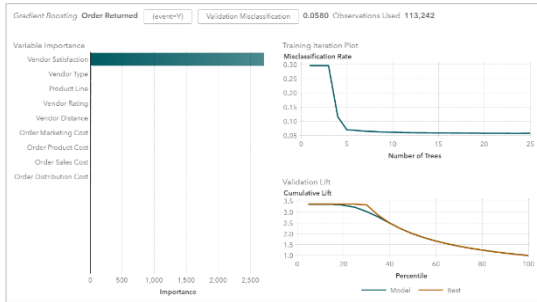


Figure 14. Gradient Boosting

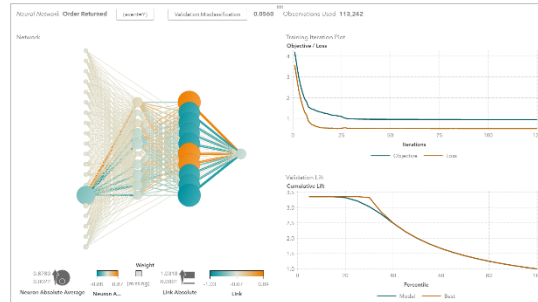


Figure 15. Neural Network

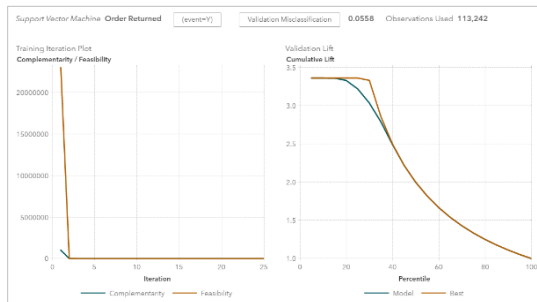


Figure 16. Support Vector Machine

As you create each model, you notice that Vendor Satisfaction is consistently one of the most important predictors for when an order is returned. You make a mental note of this observation.

MODEL COMPARISON

Once all the models are created, you can quickly compare all six in the model comparison visualization shown in Figure 17. There are fourteen different fit statistics that can be used to help you determine the champion model. The application guides you through this process by displaying the selected / best model for the active fit statistic.

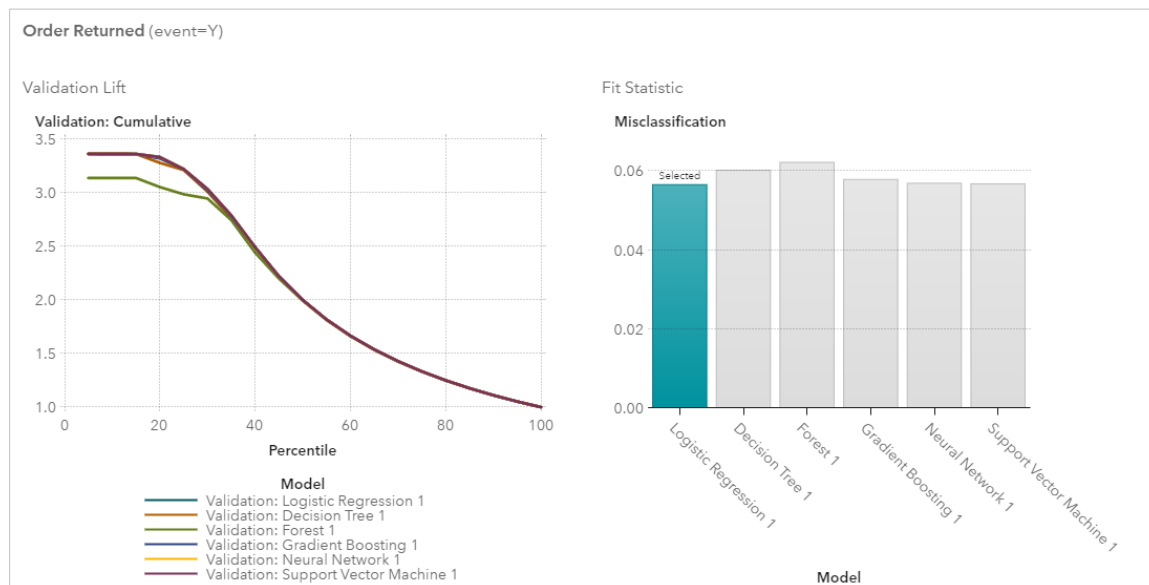


Figure 17. Model Comparison

After reviewing several of the fit statistics, you decide the logistic regression model is your champion. Don't forget to annotate the model comparison page with information on how and why you choose this model as the champion.

INTERACTIVELY REVIEW THE MODEL

This is where the power of having an interactive modeling tool pays dividends. You flip back to the page with your champion model, the logistic regression, and with a click of the mouse you derive the predicted value and probability value for the model. These values, Probability: Order Returned=Y and Predicted: Order Returned, are now available as new data items in the report. This allows you to use them as inputs to other models or as data items in report visuals. The data items for the logistic regression are stored as score code in the report. You also have the option to export the model for use in other applications such as SAS® Studio or to place it in your corporate analytics process.

Time to review our objective: you need to come up with a plan to increase profits by 5% this year. You have narrowed down options to reducing the amount of orders returned by Insight Toy vendors. You have modeled the vendors who have returned orders. Next you review the model to see if you can segment the vendors for a campaign targeted at the vendors that are returning the most orders.

You decide to review the model's prediction for Order Returned. You create a page, see Figure 18, that allows you to visualize the actual number of orders returned and the predicted number of orders returned based on the Probability Order Returned=Y. This page includes a parameter to dynamically control the prediction cutoff to review different scenarios.

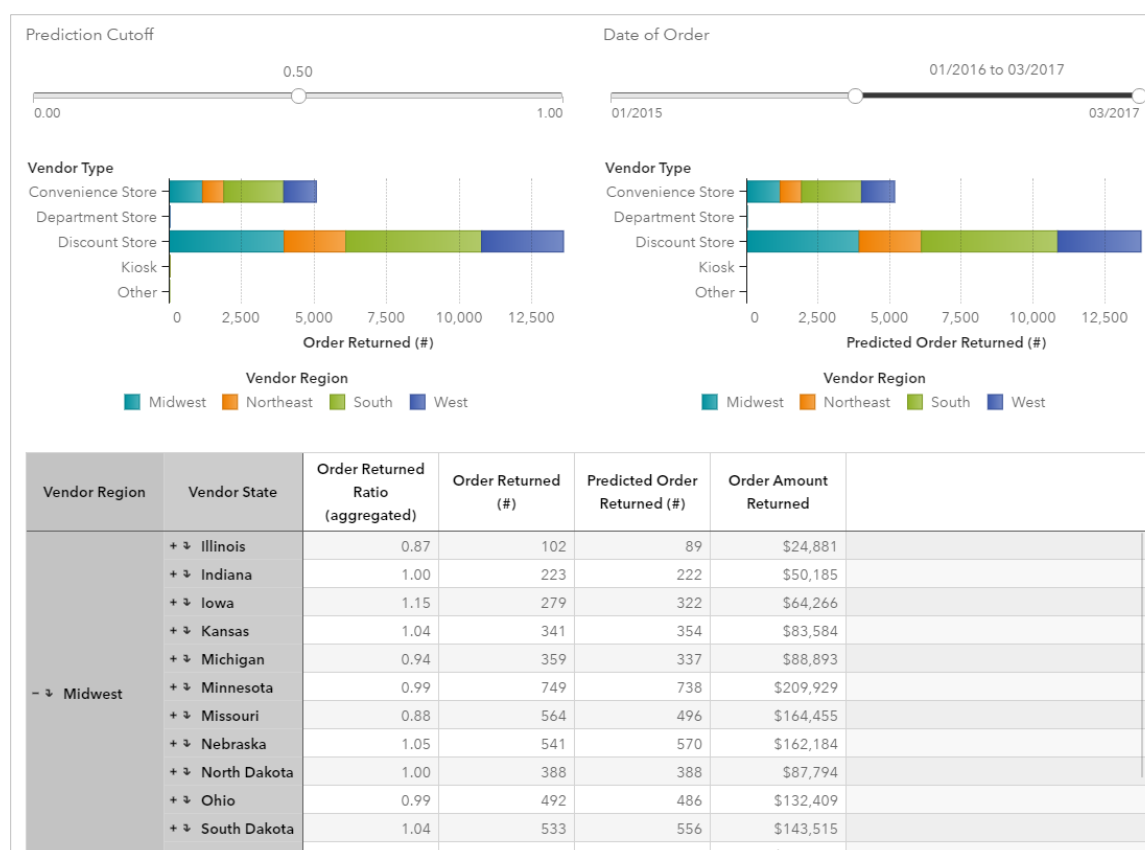


Figure 18. Model Review

You review the results of your logistic regression model and you feel good about its ability to help you predict which vendors are returning orders.

You have two analytics that are perfect for segmentation: k-means clustering and decision tree. You use clustering to segment based on Probability Order Returned=Y and Vendor Satisfaction. You are using the results of your logistic regression model, Probability Order Returned=Y, as a cluster input variable. The other input variable, Vendor Satisfaction, was consistently one of the most important predictors identified during modeling. Figure 19 shows the results of your segmentation.

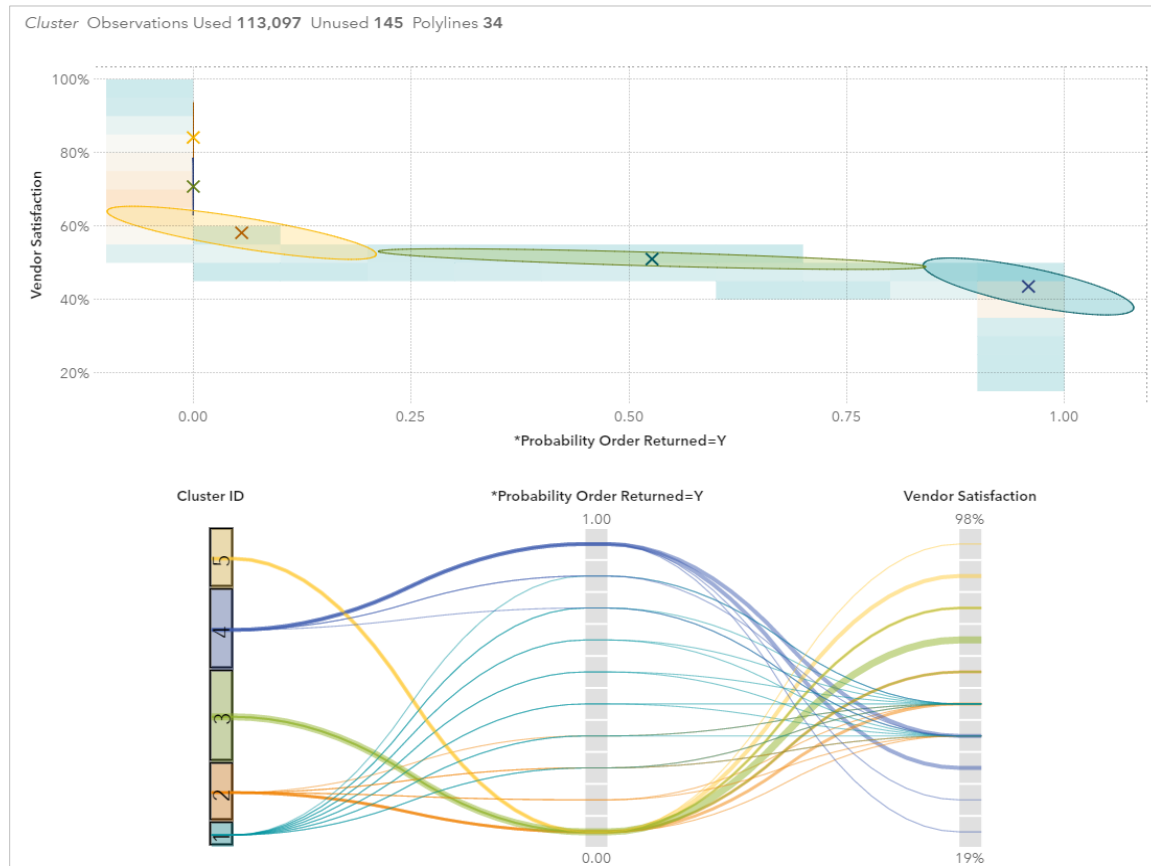


Figure 19. K-Means Clustering

You can interactively change the view in the parallel coordinates plot, as show in Figure 20. You observe from the parallel coordinates plot that the highest probability for an order to be returned is with cluster 4 and there is some contribution from cluster 1. You also observe that they have low vendor satisfaction, which aligns with everything you have seen so far.

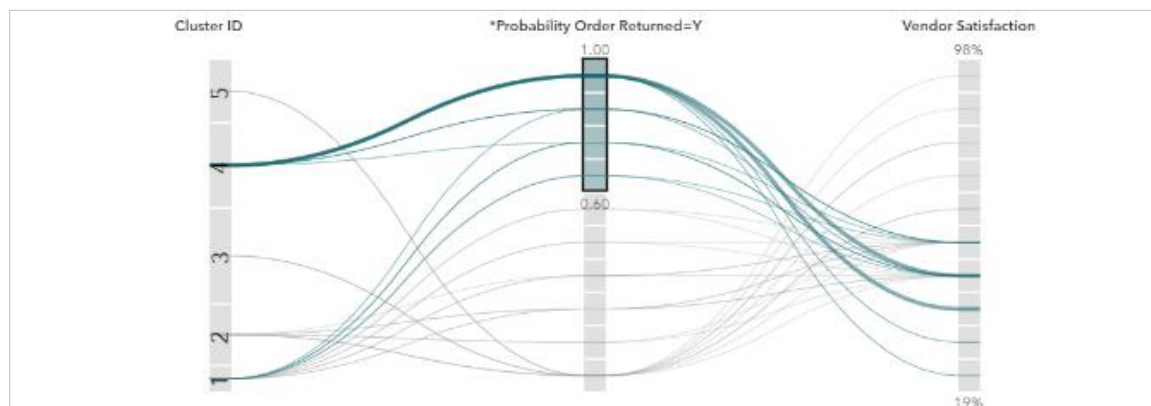


Figure 20. Clustering Parallel Coordinates Plot

You easily derive the clusters from the visualization and generate a new feature, Targeted Vendors – Cluster, to help you solve your business problem.

Next you want to visualize the results of your cluster-based segmentation by using Targeted Vendors – Cluster. Figure 21 shows a highly interactive set of visualizations that allow you to filter the entire page by Vendor Region and Order Date. It also allows you to only display a specified number of top vendors. On this page you applied a filter to show information about the Midwest starting in January of last year. At the bottom of the page is a stacked container that allows you to view details on the amount and number of orders returned.

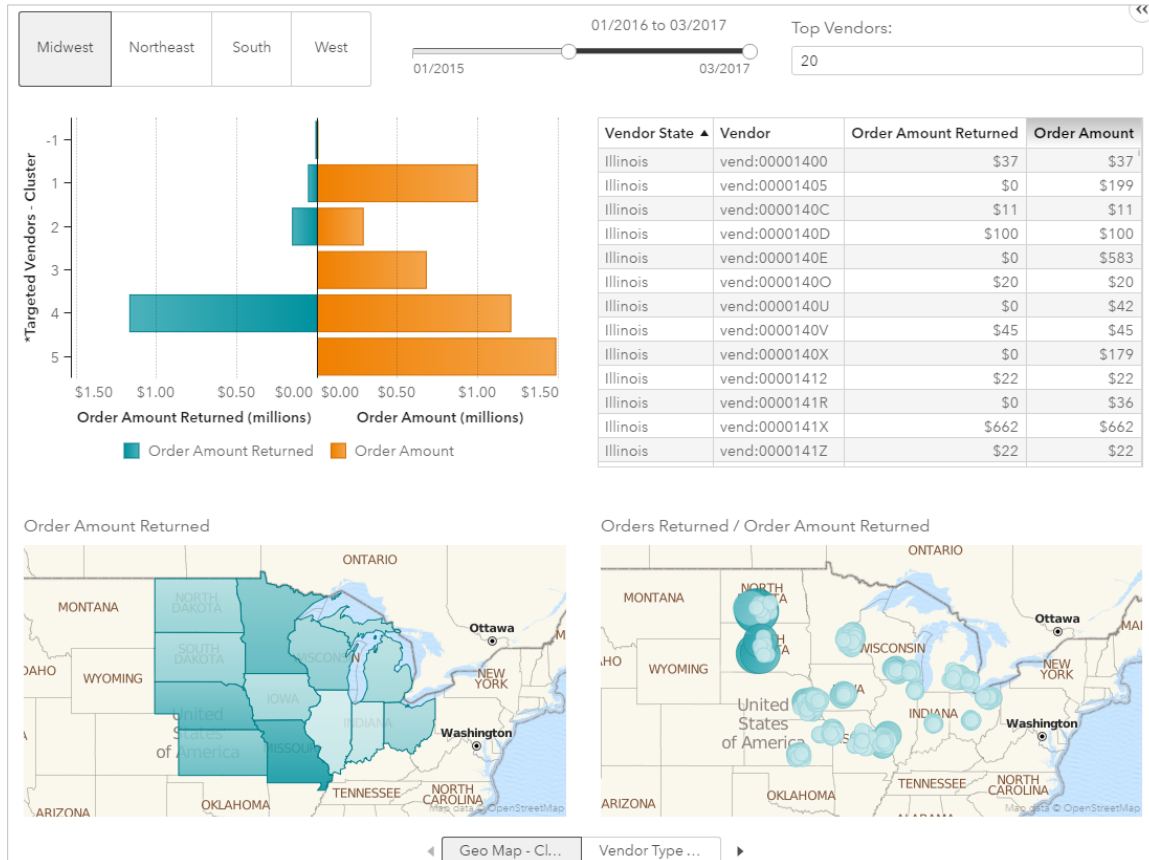


Figure 21. Cluster Targeted Vendors – Geographic

The stacked container allows you to navigate from the Geo Map view shown at the bottom of Figure 21 to the Vendor Type view shown in Figure 22.

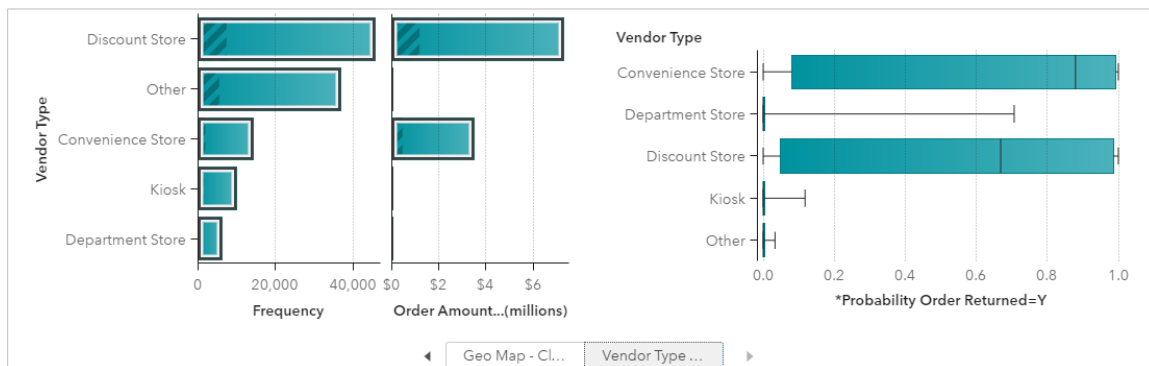


Figure 22. Cluster Targeted Vendors – Type

The Vendor Type view shows you what percentage of the data for each Vendor Type comes from what is displayed in the butterfly chart. It also shows you a box plot of the probability an order was returned by the type of vendor.

Next you use a decision tree to segment Predicted: Order Returned based on Product Line, Vendor Type, Vendor Satisfaction, and Vendor State. Similar to the clustering segmentation, you are using the results of your logistic regression model, Predicted: Order Returned, as your response variable in your segmentation. Figure 23 shows the results of your segmentation.

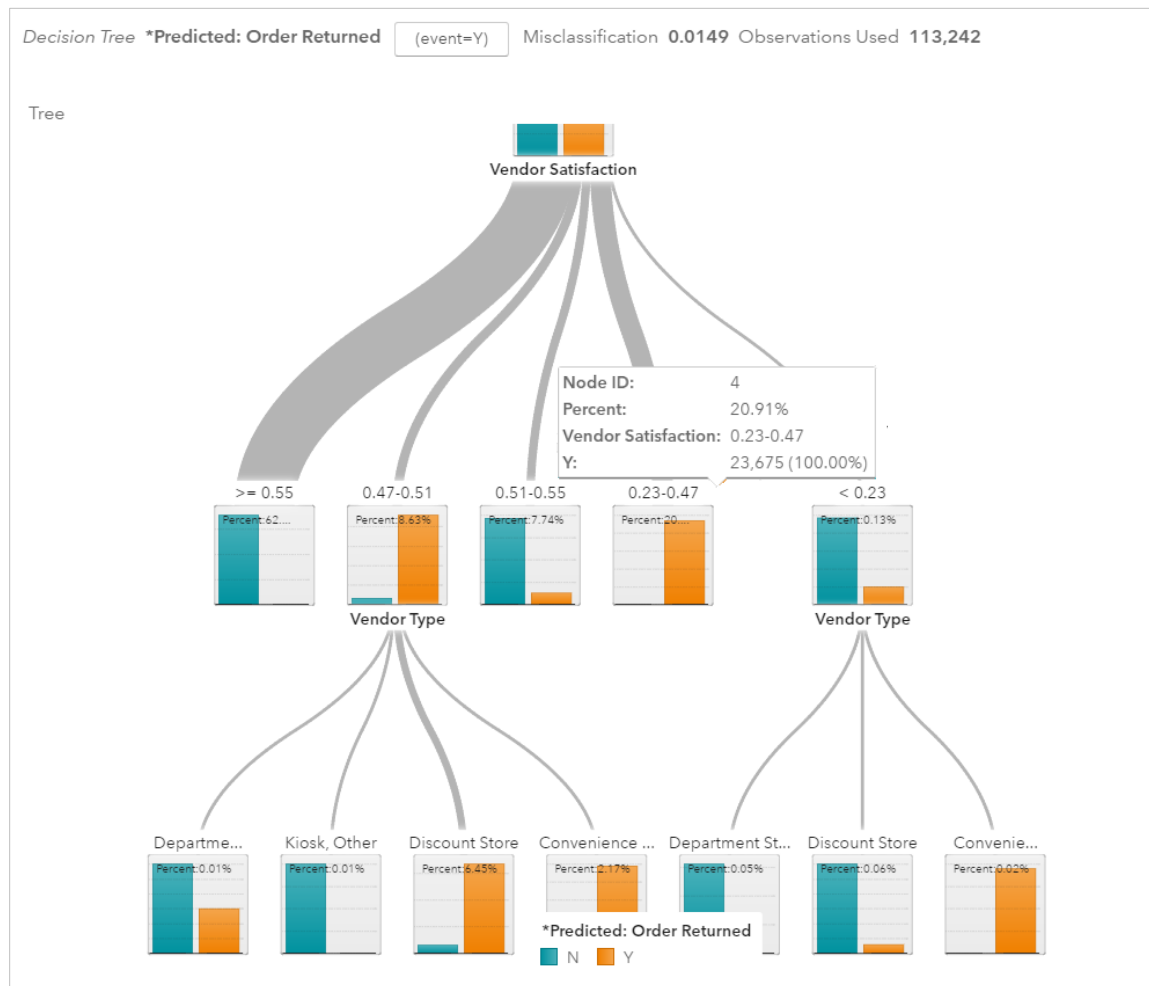


Figure 23. Decision Tree Segmentation

You observe from the decision tree that one node, the fourth one in from the left on the second level, has the highest percentage of Y(es) observations for Predicted: Order Returned. You easily derive the decision tree node IDs from the visualization and generate a new feature, Targeted Vendors – DTree, to provide a second segment to help you solve your business problem.

To compare the two features you have created, you visualize the results of your decision tree based segmentation by using Targeted Vendors – DTree in the same type of visualization you used for Targeted Vendors – Cluster. Figure 24 and Figure 25 show these visualizations.

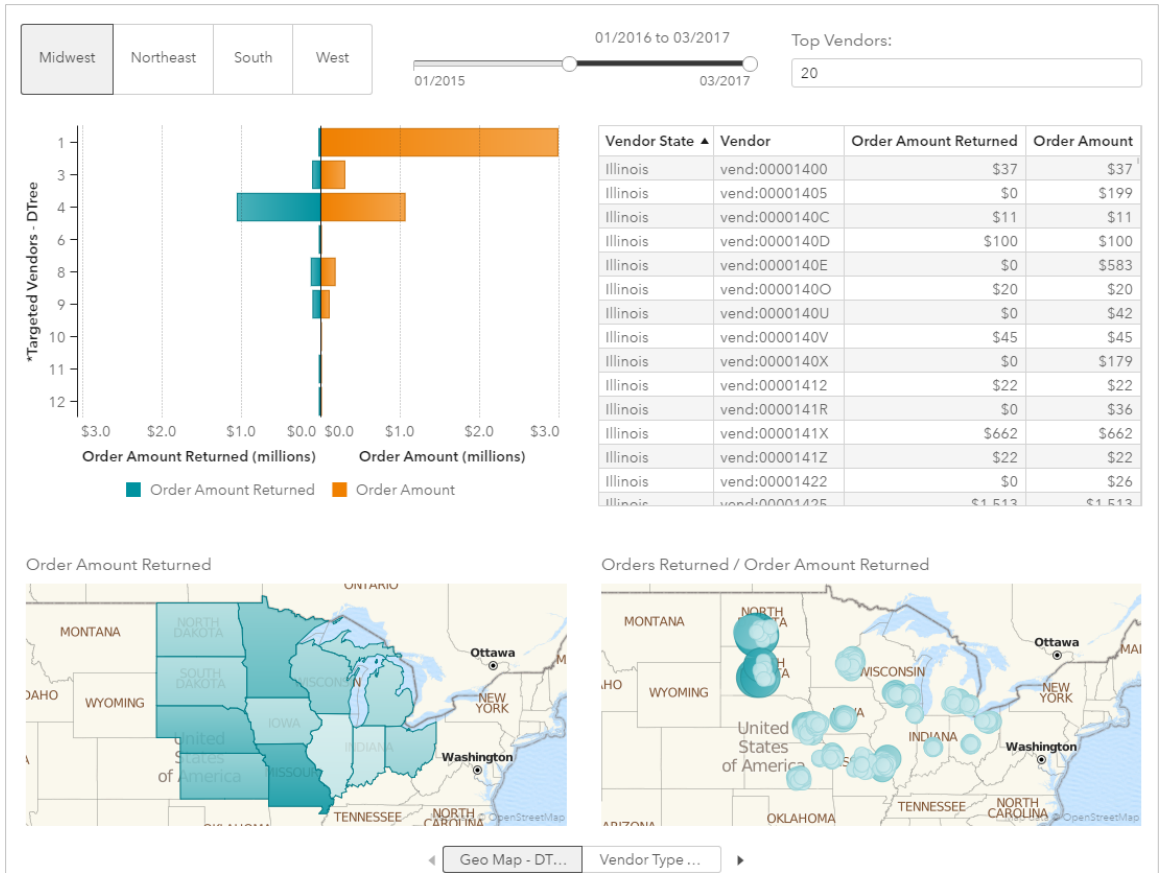


Figure 24. DTree Targeted Vendors – Geographic

From these results you confirm that node 4 has the largest contribution to Order Amount Returned.

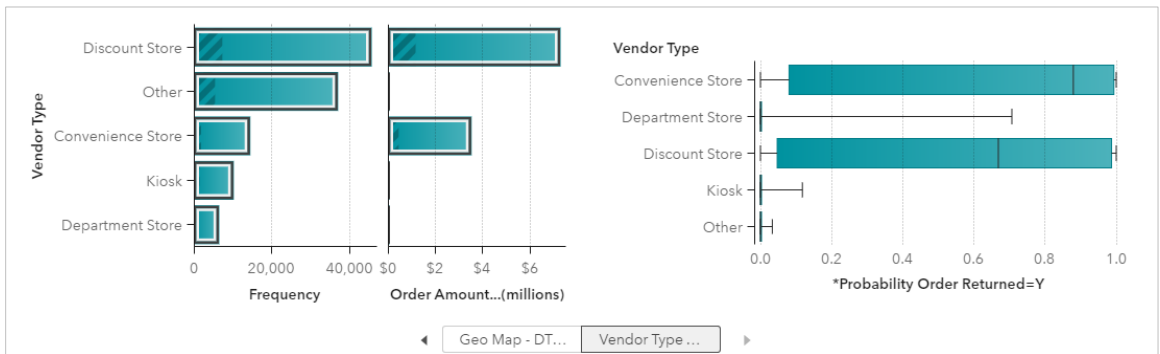


Figure 25. DTree Targeted Vendors - Type

GENERATE PROPOSALS TO THE BUSINESS PROBLEM

Now it's time to put your analytics on the line and make your manager look good. With some additional feature engineering you have constructed a page for the cluster proposal and a page for the decision tree proposal. Each page allows you to enter one or two segments and a targeted return rate for every sales office. Based on this information, a bar chart will dynamically show you how the actual profits and projected profits would compare to the targeted profits for each quarter over the last couple of years. A set of donut charts will dynamically update to show you how many vendors need to be contacted in each state.

Figure 26 shows you the cluster proposal.

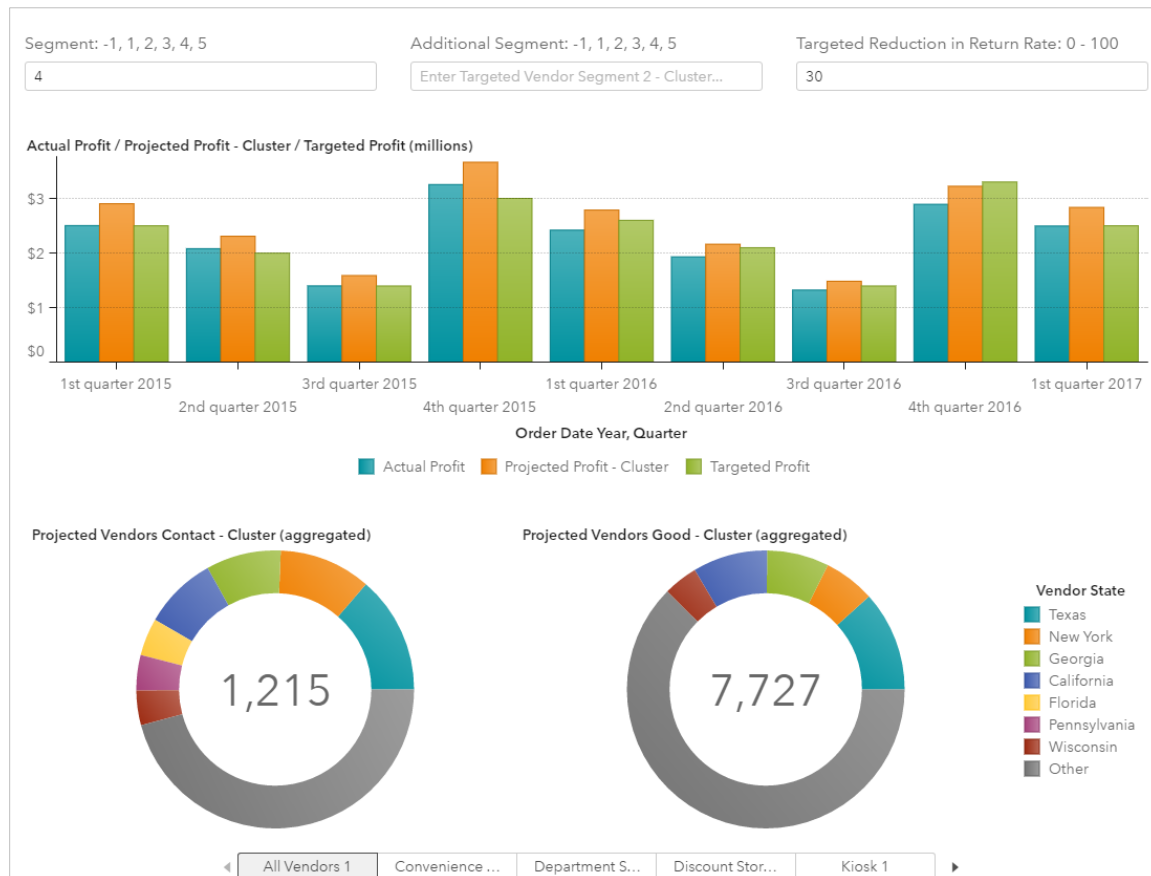


Figure 26. Cluster Proposal

You enter cluster 4, the segment containing the majority of the orders returned, and a value of 30 for the targeted reduction in return rate. The reduction in return rate specifies the percentage of returns that need to be eliminated. If the sales organization can reduce returns by 30%, then Insight Toy would have seen profits exceed the target for every quarter but one.

The donut chart at the bottom of the page shows you the vendors that should be targeted for this campaign. There are 1,215 vendors to contact and 7,727 that do not need to be contacted. The state that has the most vendors to contact is Texas, and it has 152. In addition to the overall visualization for projected vendors to contact, you can navigate to targeted donut chart based on Vendor Type.

Figure 27 shows you the decision tree proposal with the same visualizations as the cluster proposal.



Figure 27. DTree Proposal

You enter node 4, the segment containing the majority of the orders returned, and a value of 30 for the targeted reduction in return rate. The results for the decision tree proposal are similar to those for the cluster proposal. One difference between the proposals is that fewer vendors need to be contacted for the decision tree proposal.

The stacked container at the bottom of the page also contains a list table with details on the vendors that need to be contacted. In Figure 28 you sorted the list by Order Amount Returned. Your sales organization can access this report to determine which vendors they need to contact. They can also export the list to a spreadsheet.

Vendor State	Vendor City	Vendor	Vendor Type	Vendor Satisfaction	*Probability Order Returned=Y	Order Amount Returned ▼
New Hampshire	Manchester	vend:000044FG	Discount Store	42%	0.98	\$151,887
New Hampshire	Manchester	vend:000044FK	Discount Store	44%	0.95	\$139,035
North Dakota	Bismarck	vend:00000NET	Discount Store	44%	0.98	\$104,827
South Dakota	Pierre	vend:00000LKK	Discount Store	43%	0.99	\$86,322
New Hampshire	Manchester	vend:000045OD	Convenience Store	51%	0.63	\$81,246
Massachusetts	Worcester	vend:00003UUJ	Discount Store	45%	0.95	\$80,484
South Dakota	Pierre	vend:00000KMJ	Discount Store	46%	0.95	\$76,829
New Hampshire	Manchester	vend:000044FH	Discount Store	49%	0.65	\$75,467
Nevada	Reno	vend:00000O69	Convenience Store	46%	0.97	\$75,186
South Dakota	Pierre	vend:00000IU5	Discount Store	47%	0.89	\$74,142
Connecticut	Hartford	vend:00003YI4	Discount Store	49%	0.67	\$71,280
Washington	Olympia	vend:00000LNE	Discount Store	44%	0.90	\$70,504

◀ Department S... Discount Stor... Kiosk 2 Other 2 Contact List 2 ▶

Figure 28. Contact List

You have a solid proposal with two options to present to your manager. The proposal is to get the sales offices to increase sales, but more importantly work with their vendors to decrease the number of orders returned. One aspect of the sales/vendor relationship that contributes to orders being returned is dissatisfied vendors. The sales team needs to engage the vendors to increase satisfaction.

Your work will help turn Insight Toy into a profitable company again.

CONCLUSION

The SAS Visual Statistics and SAS Visual Data Mining and Machine Learning add-on offerings to SAS Visual Analytics contain a robust set of tools that allow data scientists to explore their data, engineer features, interactively generate models, and use the model's output all within the same report. The integration of advanced modeling techniques, approachable analytics, and reporting capabilities provide the data scientist with a single tool for solving complex business problems and presenting the results in a business-friendly format.

CONTACT INFORMATION

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

Don Chapman
SAS Institute Inc.
Don.Chapman@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.