SAS® Viya® Trial

Develop Models Guide

Data Scientist Tasks

Intro

Data and AI Life Cycle: Develop Models

A recent study by The Futurum Group showed that SAS Viya increases data and AI team productivity by 4.6x.

The analysts compared SAS Viya to alternatives in an end-to-end customer churn prediction analysis, a common use case relevant to many industries.

The second step in the data and AI life cycle is **Develop Models**. This was performed by a **Data Scientist** persona, who explored the data set prepared by the Data Engineer and utilized it to build a model to predict customer churn.

This guide will walk you through the steps a Data Scientist took to complete the Develop Models portion of the life cycle in SAS Viya.

Data Scientist

Explore and Transform Data

Develop, Optimize, Validate and Document Models

Tasks

1. Visual Exploration and Insights Discovery
2. Visual Exploration – Augmented Analytics
3. Outlier Detection
4. Quick Model Prototyping
5. Templates in Model Studio
6. Model Competitions
7. Explainability
8. Bias Detection
9. Model Reports
10. Pipeline Competitions
11. Model Registration
12. Project Insights Report – Documentation
13. Sharing Projects – Read/Write

Resources

Watch before start

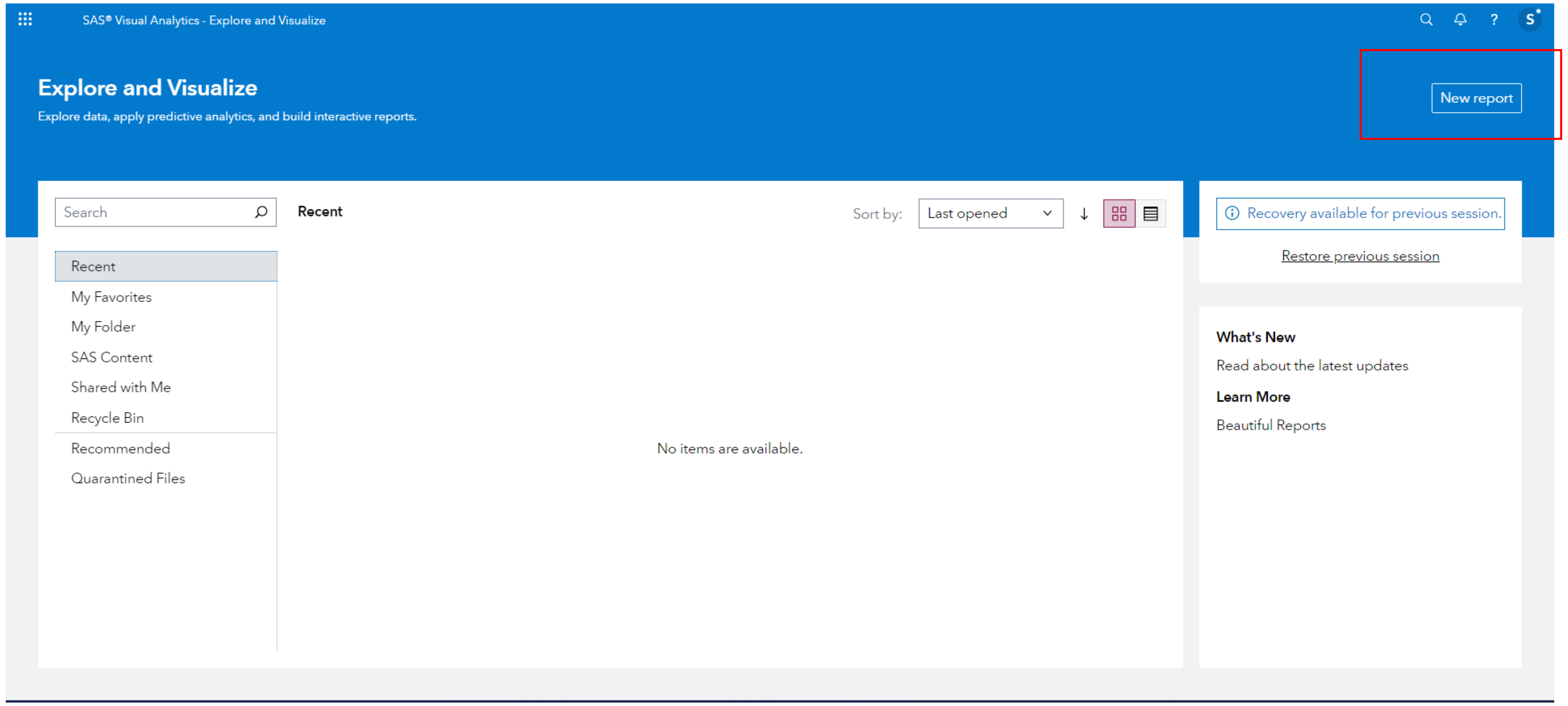
- Quick Start - Data & AI Life Cycle
- Quick Start - SAS Drive
- Quick Start - Manage Data
- Quick Start - Explore and Visualize Data
- Quick Start - Build Models
- Webinar - Getting Started With SAS Machine Learning

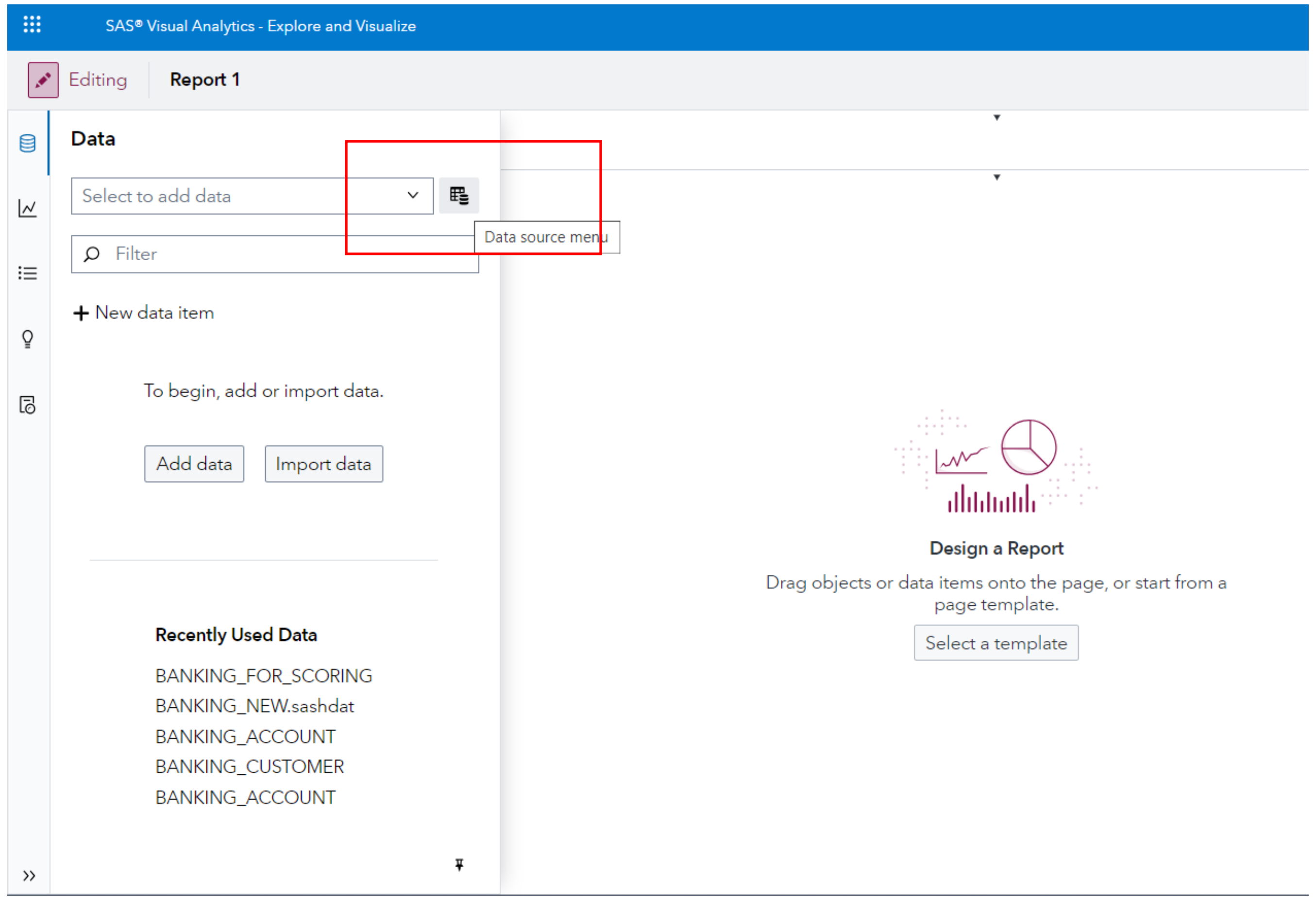
Visual Exploration and Insights Discovery

Visual Exploration and Insights Discovery

Automatic and GUI Created

To begin exploring the curated data set, go to the “Explore and Visualize” application from the applications tab on the left of the screen. Next, click “New Report.” Click the table icon next to the data tab and select “Add Data” to add the BANKING_NEW data to the report. Find the table in PUBLIC and select “Add.” All the variables should now be loaded in the Data tab and separated by categorical versus measure types.

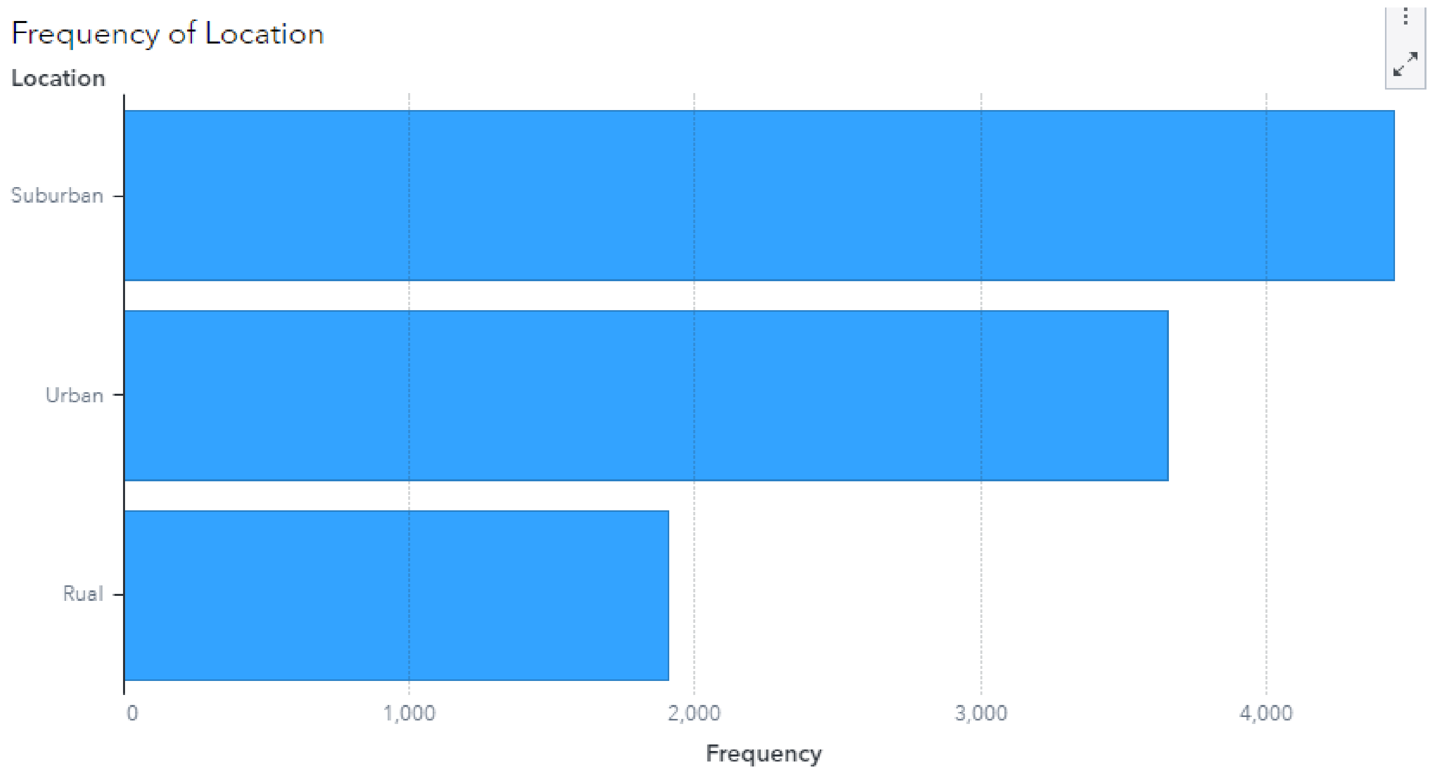


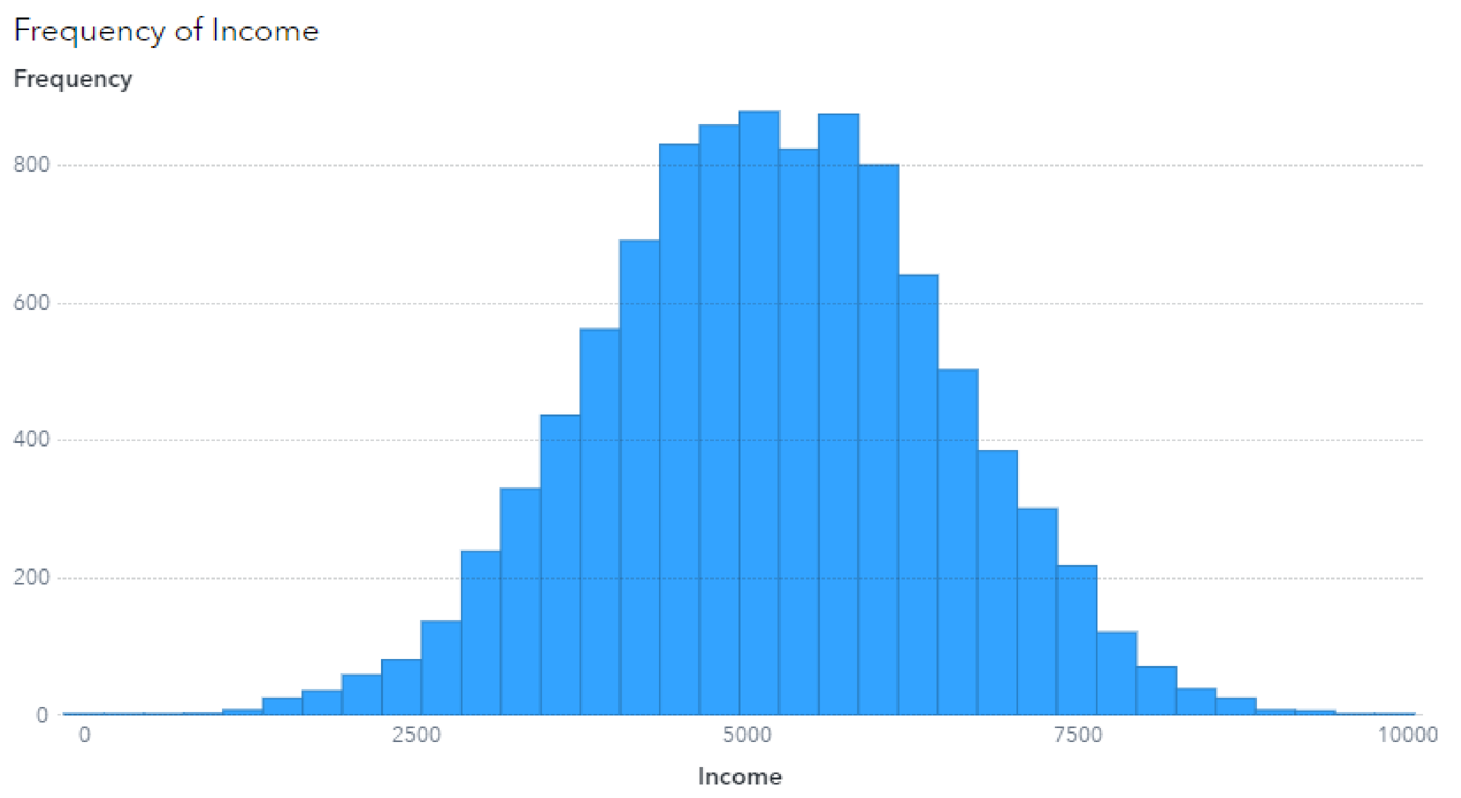


Visual Exploration and Insights Discovery

Automatic Distributions

- We'll start by exploring the data using basic data visualization. We'll explore a categorical and a numeric variable.
- When scrolling through the data tab (on the left of the screen), double-click the location variable. SAS will automatically create a bar chart for the viewer to see the three levels of the variable (suburban, urban and rural.) Scroll over the suburban bar to see how many observations are suburban in this data set (4,442 is the most for the three levels.).
- Open a new page by clicking the plus symbol "+" and then double-click the income variable to view the automatic histogram. You can modify the graphs if you want by using the options pane on the right side of the screen.

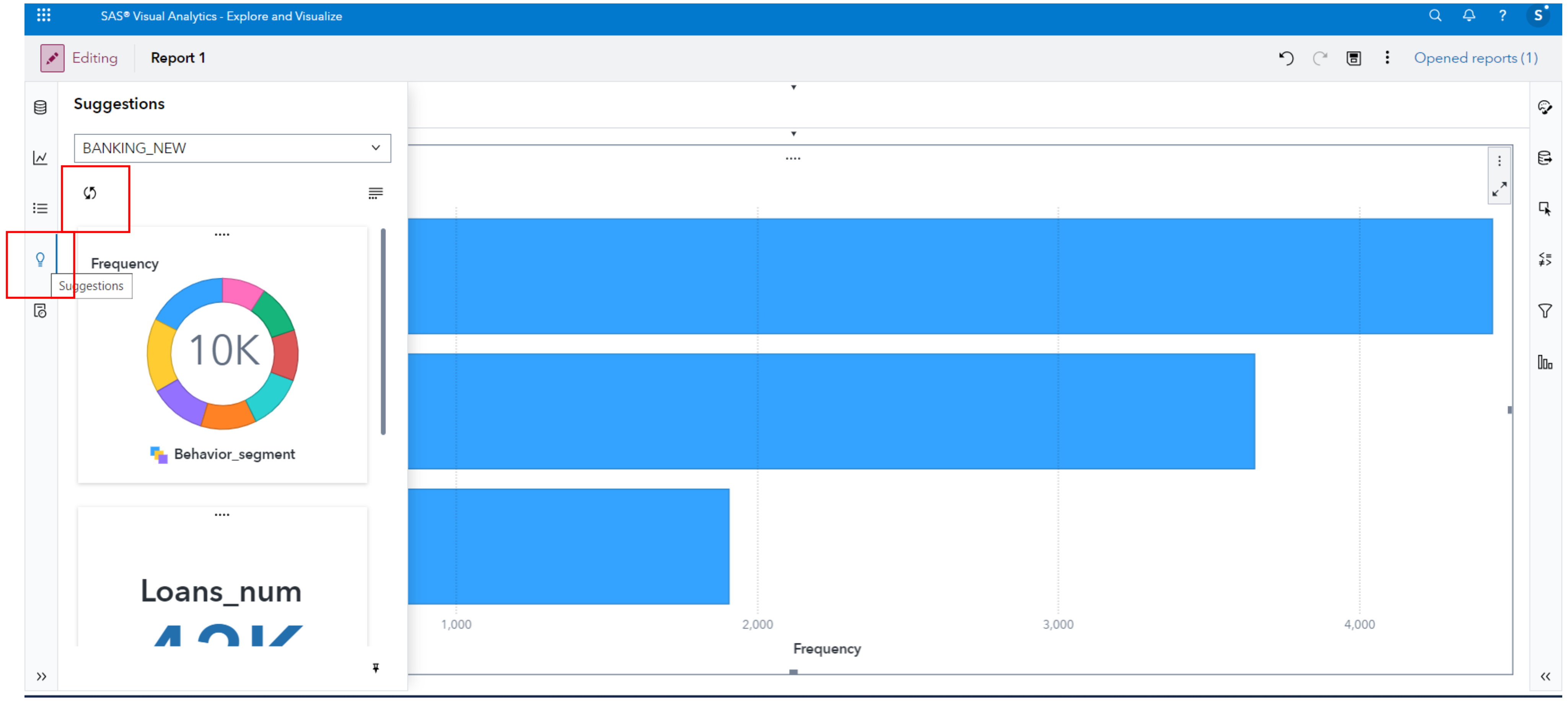


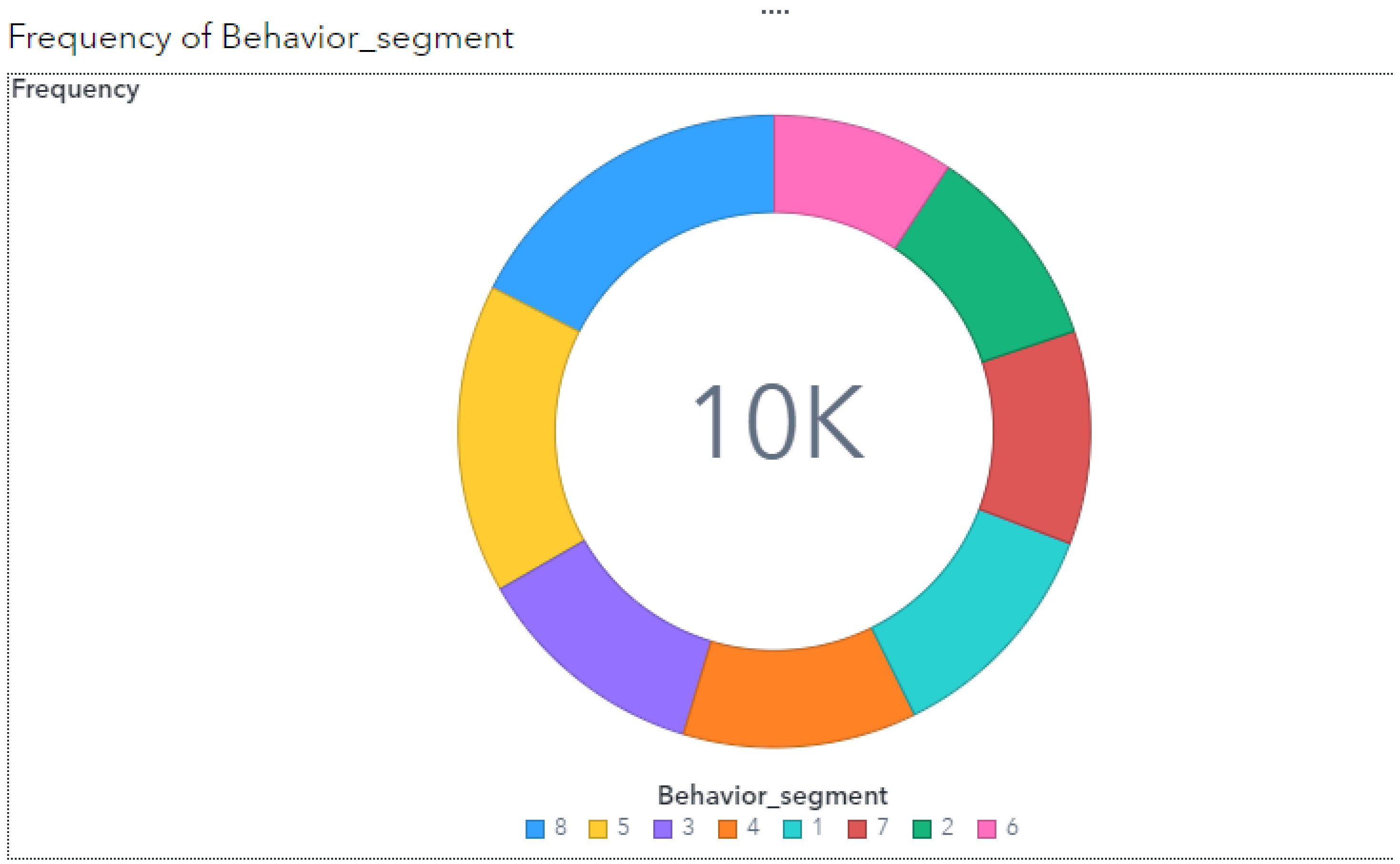


Visual Exploration and Insights Discovery

Automatic Graph Suggestions

- Now let's explore the automatic graph suggestions. These aim to provide AI-driven recommendations that could reveal hidden patterns and insights in your data in a quick and efficient way.
- On the left side of the screen, go to the "Insights tab" for automatic graphic suggestions. To view more automated recommendations, just hit the refresh icon, as seen in the first graph below.



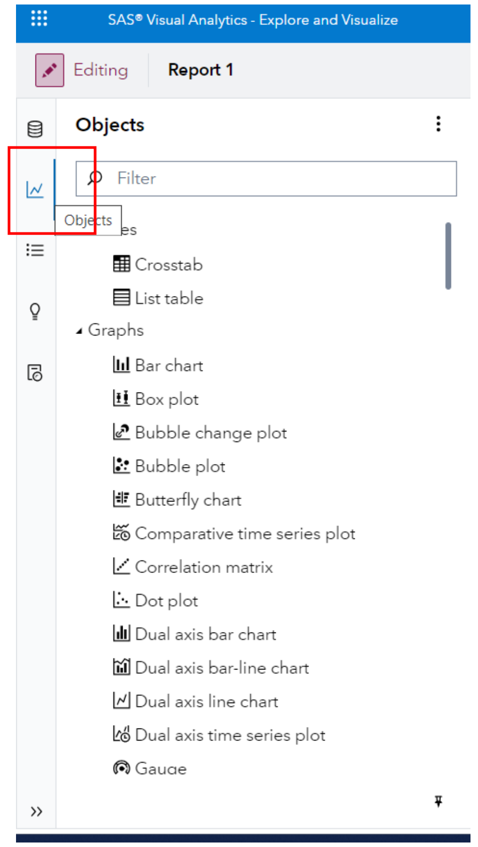


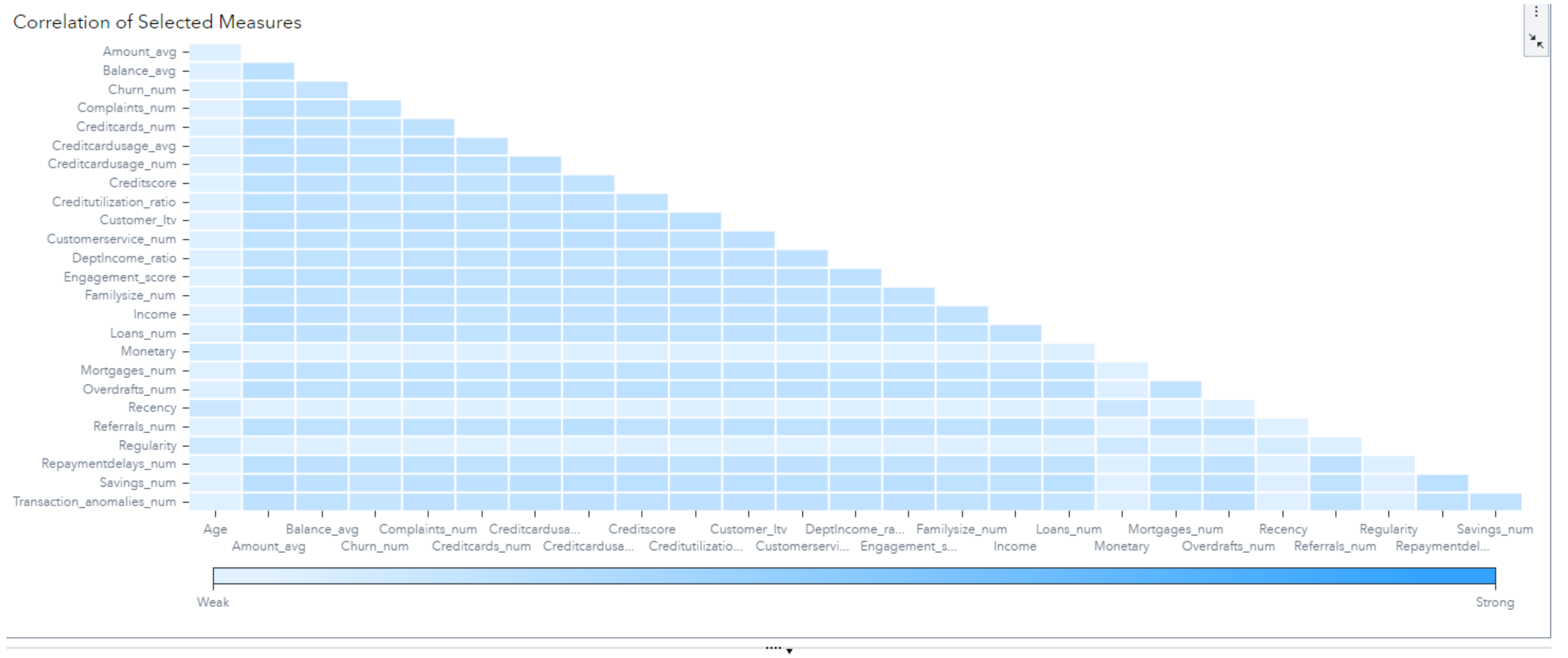
Visual Exploration – Augmented Analytics

Visual Exploration – Augmented Analytics

Correlation Matrix

- Now the Data Scientist needs to develop some common graphs to visually explore the data and understand the patterns and correlations of the independent variables in the data with the target (churn in our case). This can be done by creating the graphs manually using the drag-and-drop functionalities or completely automatically. Let's explore some manual graphs first.
- Open a new page and on the left side select the objects page. This lists all the possible explorations for the data (graphs and model prototyping methods). Select the correlation matrix graph. Assign data and add all the measure variables. Right-click on the graph and click "Maximize view." The correlation values appear at the bottom. Click the correlation term in the table to order the values from largest to smallest.



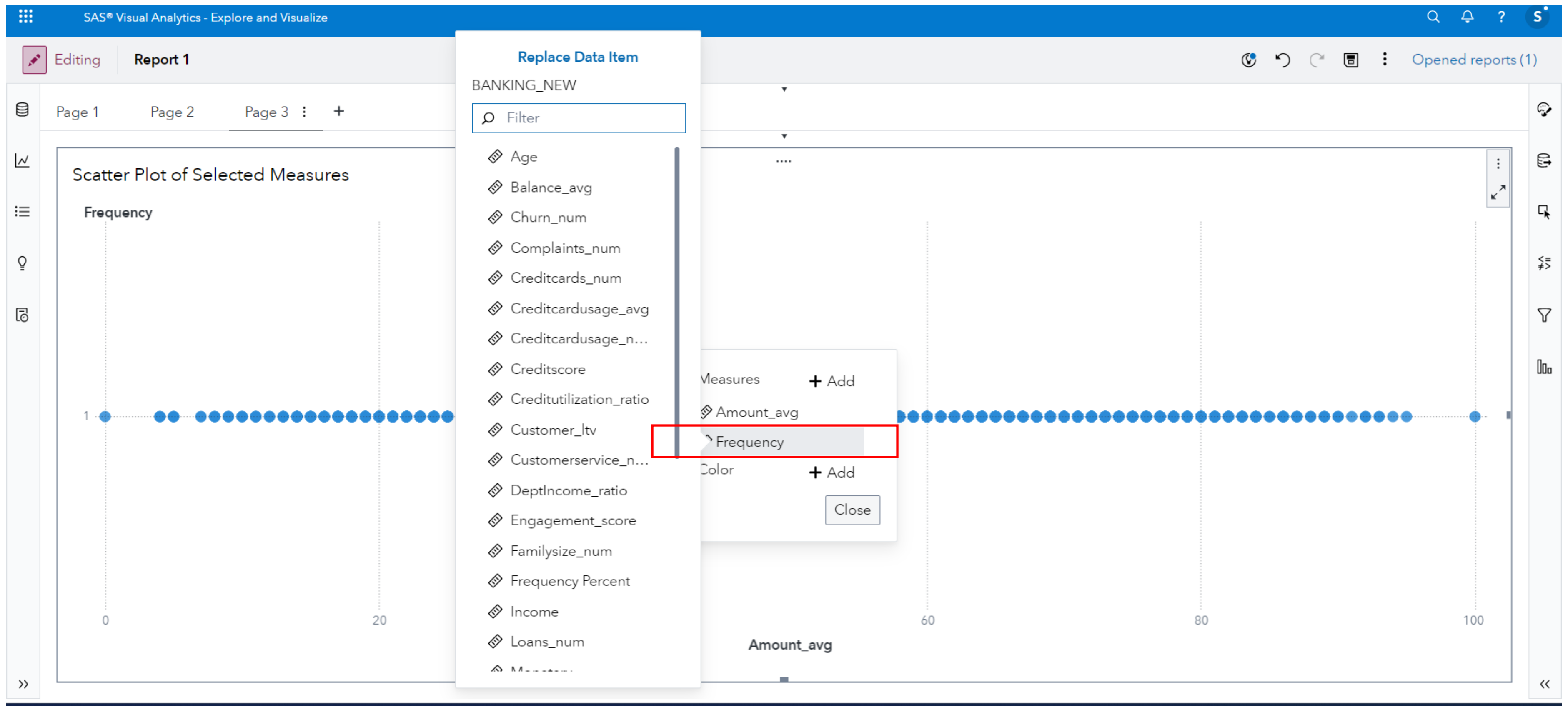


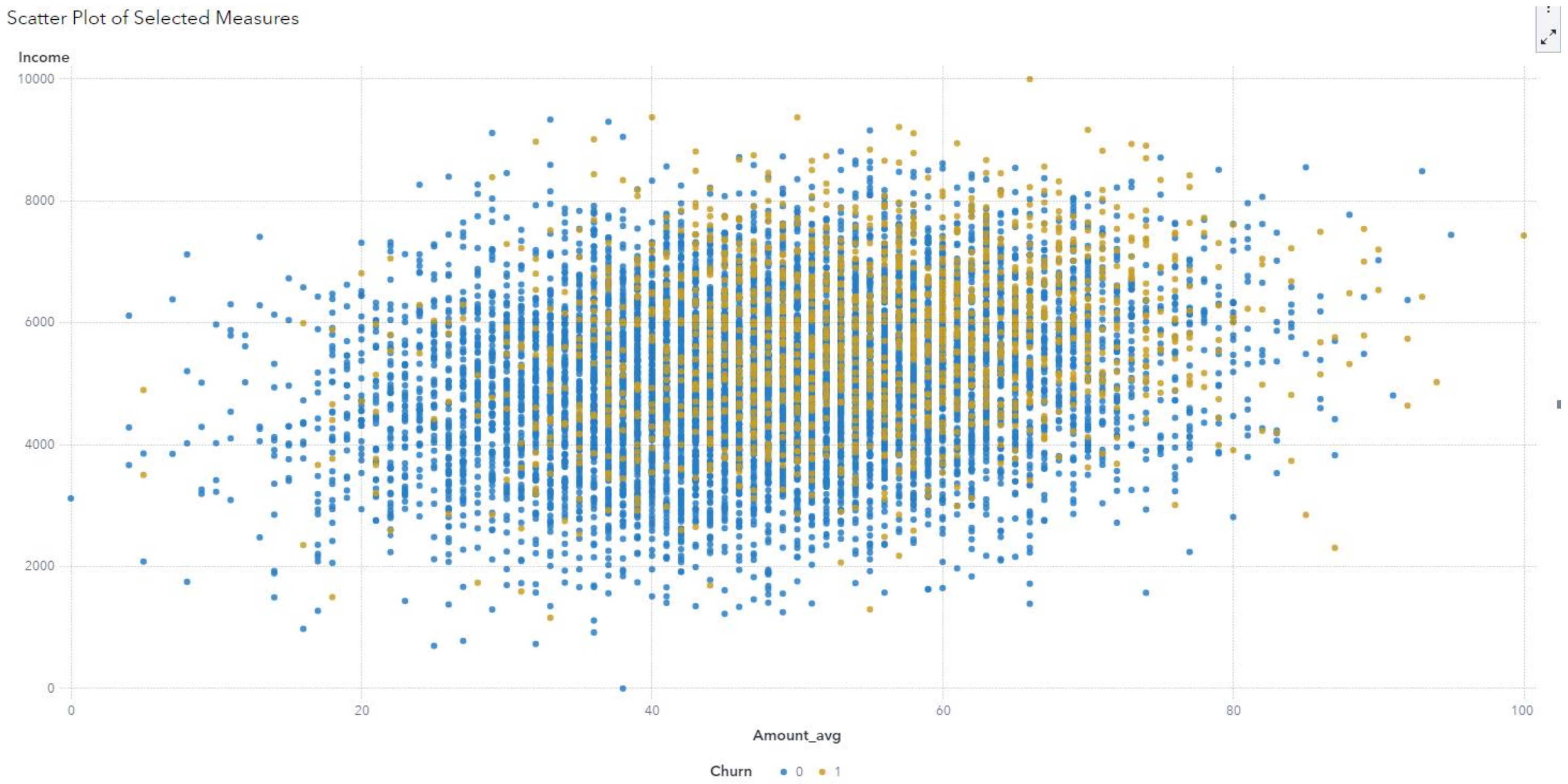
X Axis	Y Axis	Correlation ▼
Amount_avg	Income	0.2317
Amount_avg	Savings_num	0.2136
Complaints_num	Customer_ltv	0.2133
Creditcardusage_num	Referrals_num	0.2124
Engagement_score	Savings_num	0.2114
Amount_avg	Overdrafts_num	0.2113

Visual Exploration – Augmented Analytics

Scatter Plots

Based on your findings from the correlation matrix, we want to build a scatter plot. Open a new page. Select "scatter plot" from the object tab and add "Amount_avg" and "Income." Click on the "Frequency" measure that is automatically created and select the variable "Income" to replace it as measures and add "churn" (the target) as color.

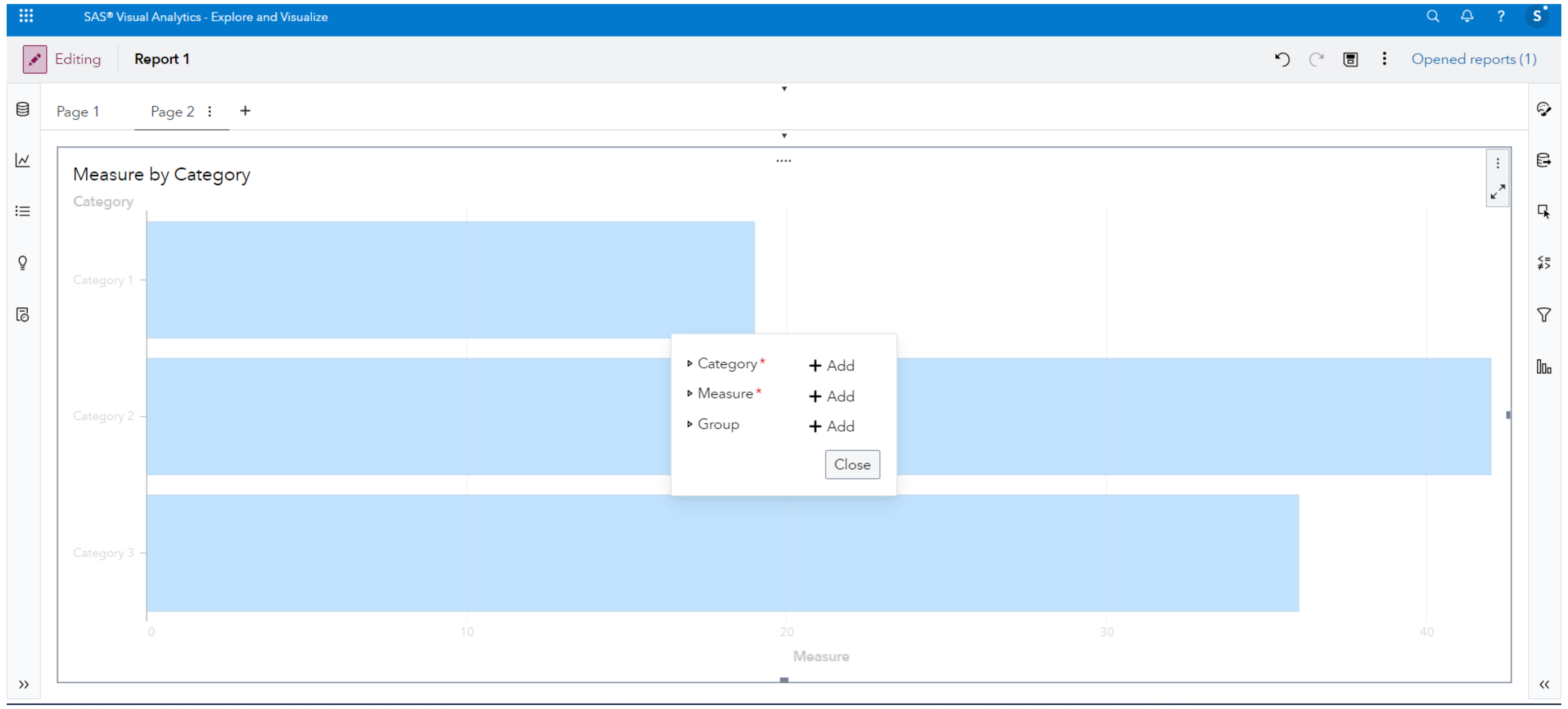


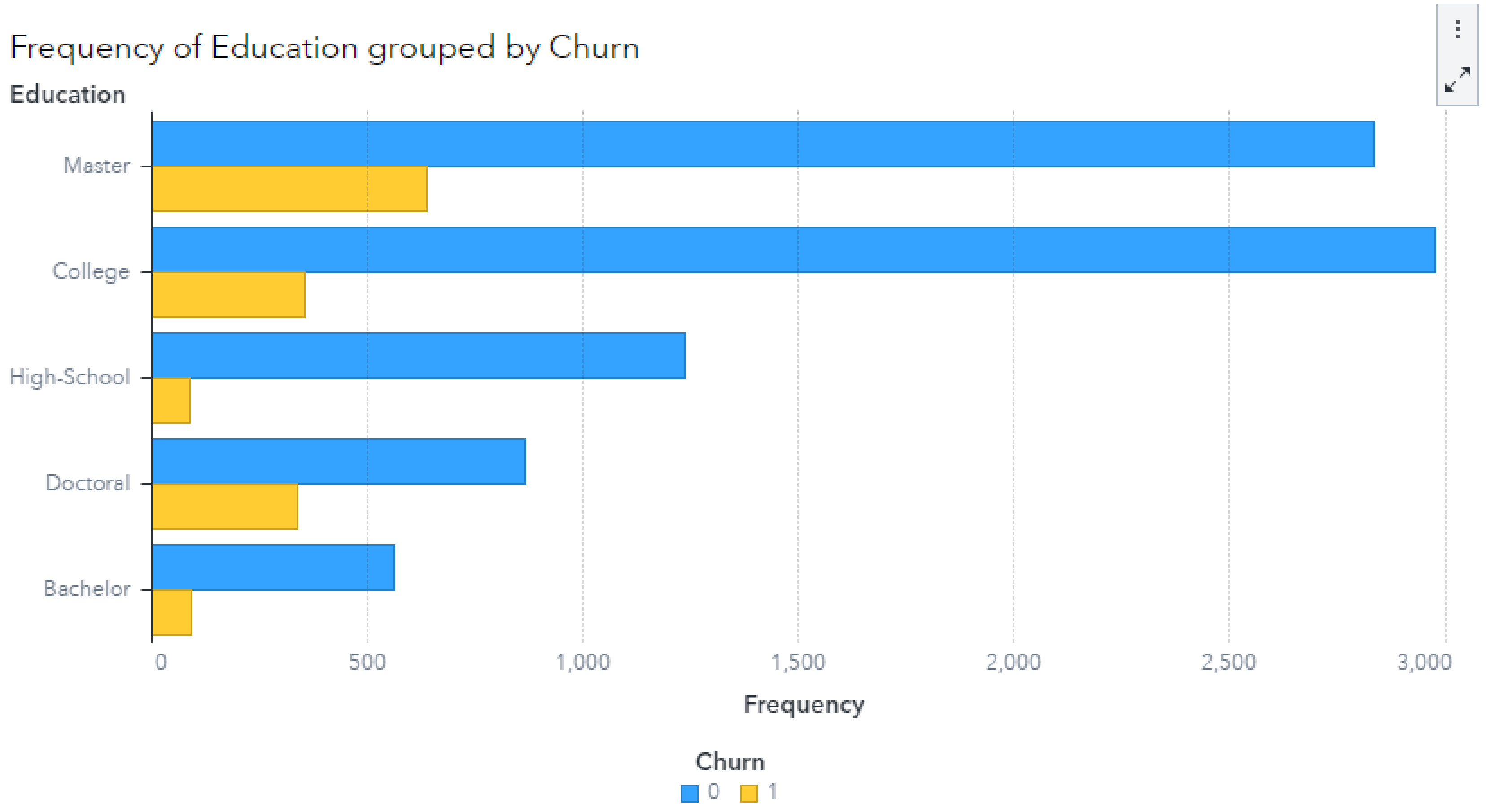


Visual Exploration – Augmented Analytics

Nested Graphs

- Sometimes we need to evaluate how the target variable (churn) behaves in our data for different groups. This way, we ensure that we don't discriminate against specific populations and that our data is unbiased. You can do this automatically when you move to the modeling step in SAS Viya but also manually, early on in the process, by using SAS exploration capabilities.
- In this step, we'll build a bar chart showing the Education level separated by the Churn level. Open a new page, select the bar chart from the objects page, and assign the education variable as the category and churn as the group.

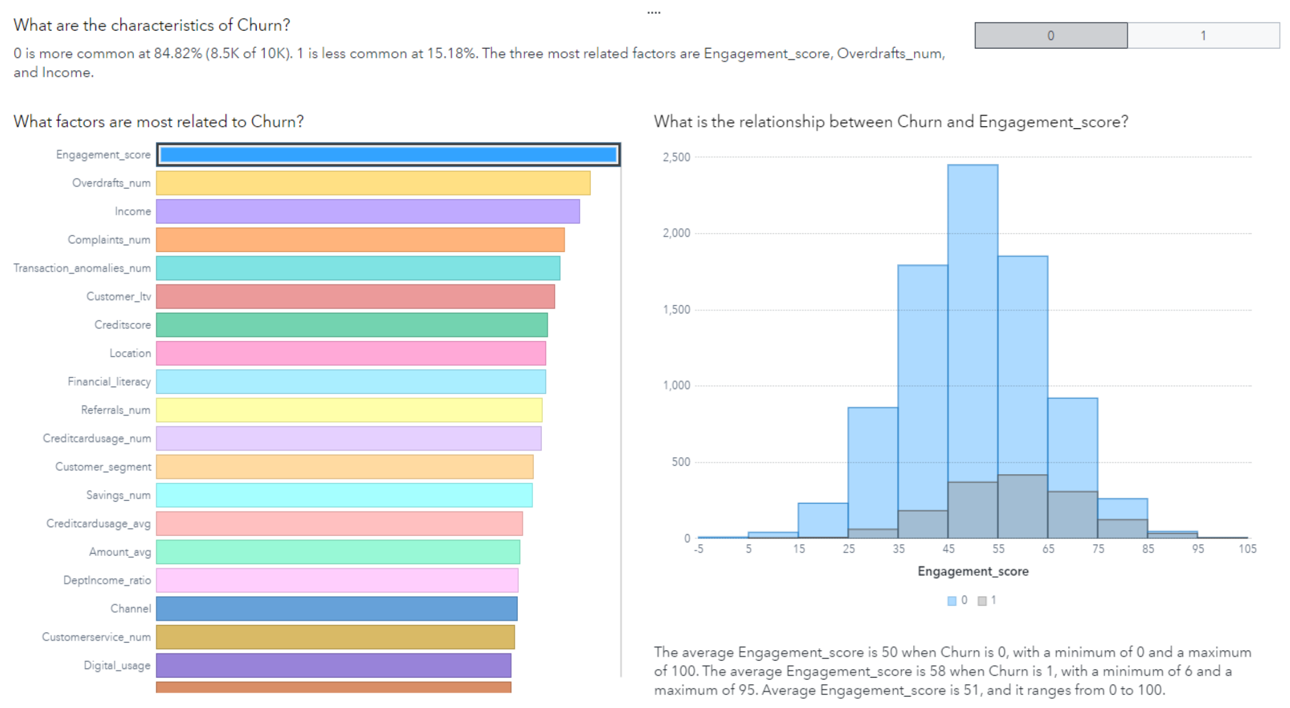


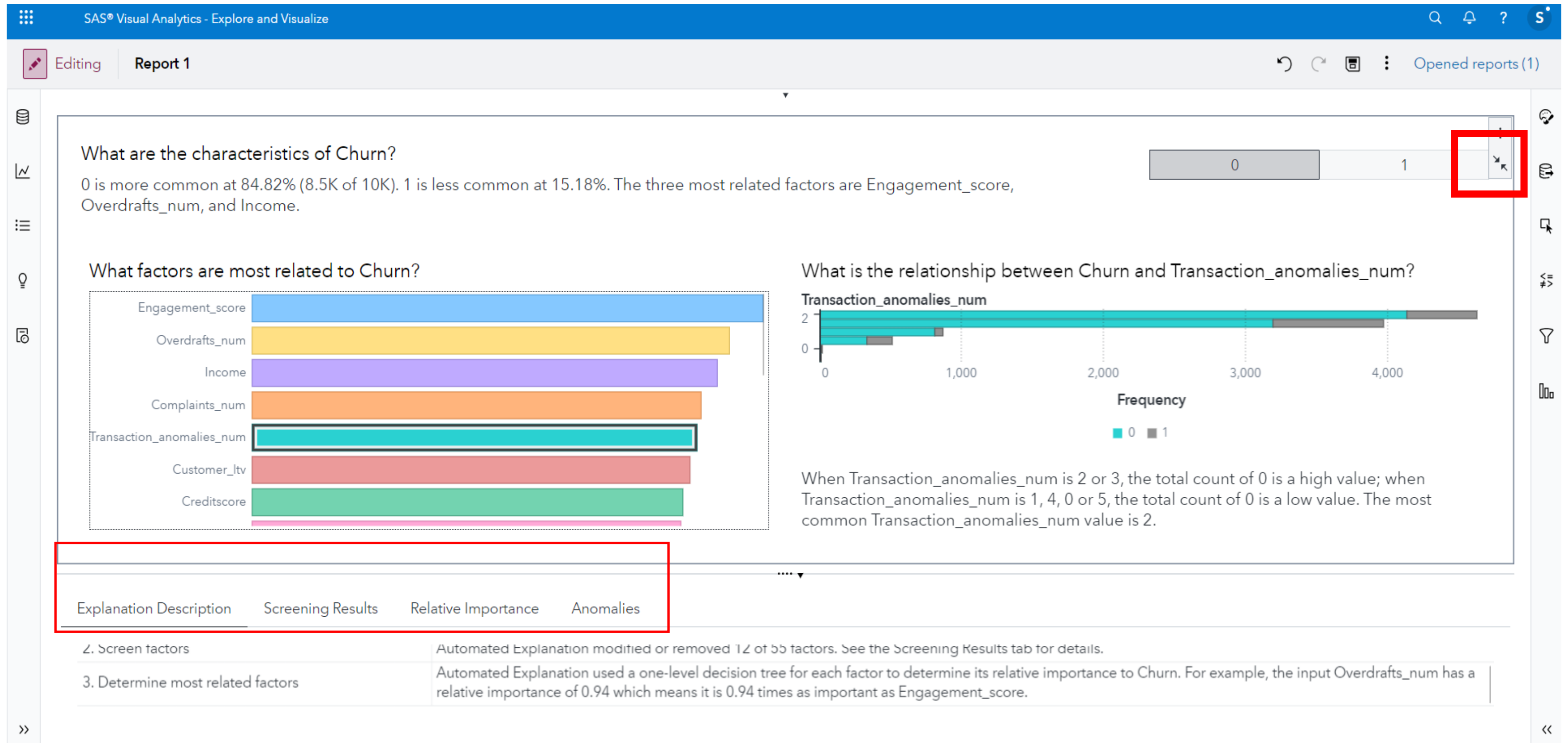


Visual Exploration – Augmented Analytics

Automated Exploration & Visualization

- Now let’s move to the automatic exploration. A Data Scientist usually has to present initial findings in the data to their stakeholders and discuss them before moving to the modeling phase. You can build a comprehensive and interactive report for your target variable (or the variable you want to explain) completely automatically. This would normally take many hours or even days to achieve.
- Open a new page, select “Automated explanation” from the object tab, and add “Churn” as the response. This graph displays the factors most related to the target based on variable importance. Here we see that the Engagement score is the most related variable. If you click any other variable from the bar chart, the relevant graph that shows the relationship with your target variable appears on the right. Finally, by maximizing the graph using the button on the top right, you can view further information about how this report is created, automated screening, relative importance values and any anomalies that are detected.
- After exploring the report, click the same button (arrows at the top right to maximize/minimize view) to keep editing the report and explore your data further.

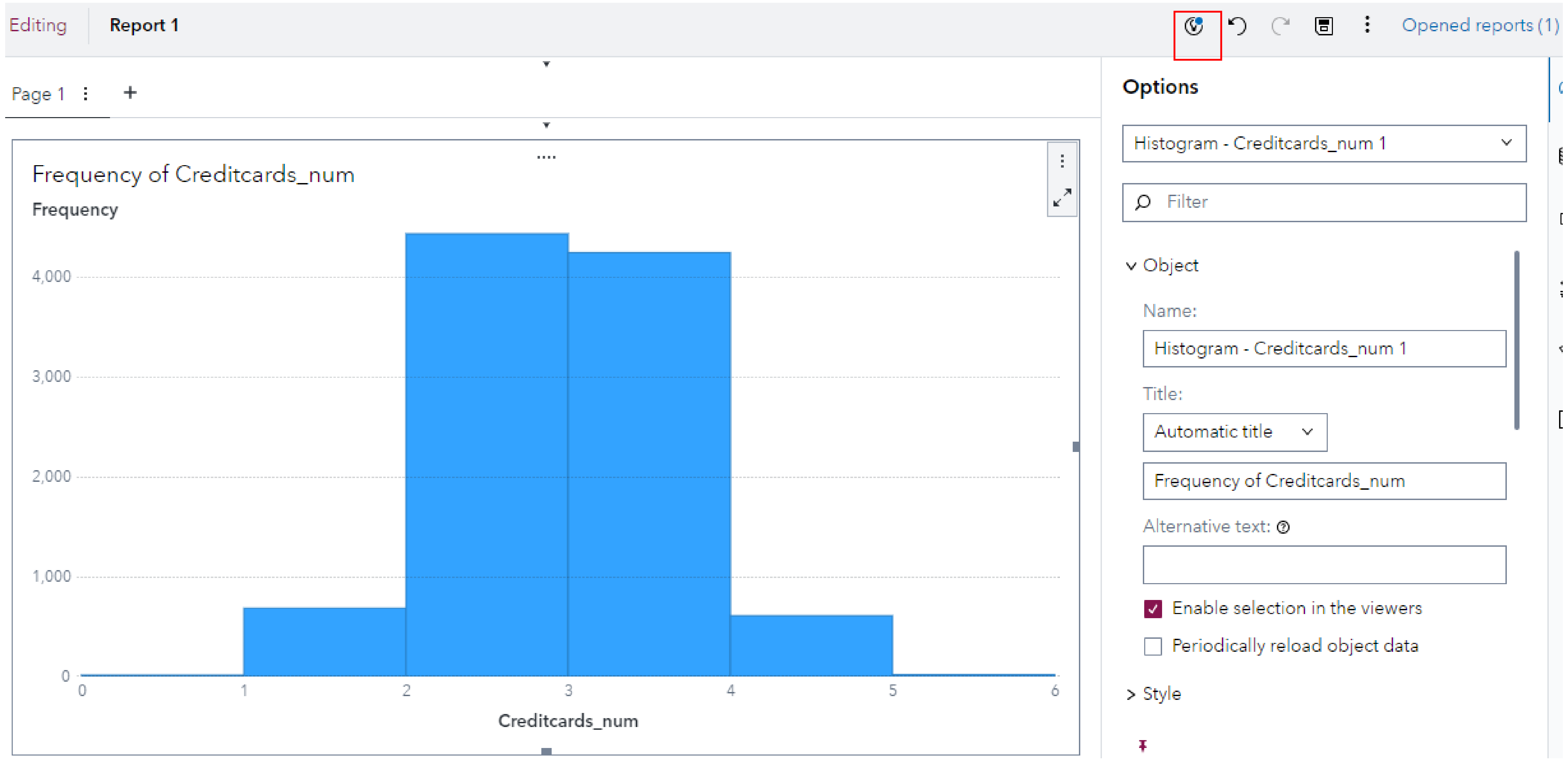




Outlier Detection

Outlier Detection

- We also need to check and take appropriate action if the data contains any extreme outliers that will negatively impact our models.
- In the "Explore and Visualize" application, open a new Page, then go to the data tab (left of the screen) and double-click the measure variable Creditcards_num. Notice this automatically generates a histogram of the variable. At the top right, the light bulb icon is automatically generated because SAS found outliers that need to be reviewed further.



Outlier Detection

Click the Insights button (light bulb icon) and then click “Analyze object” for impact. It automatically analyzes the variable for outliers but found that the outliers are not impactful in this case. Click the information icon next to the variable name and then click "View outlier analysis” to view the outlier report. In the report, you can see those outliers and how they affect the sum, average and median of this variable.

Outliers

The following data items are used by objects in the report and might be influenced by outliers:

Analyze Objects for Impact

Creditcards_num

Outliers

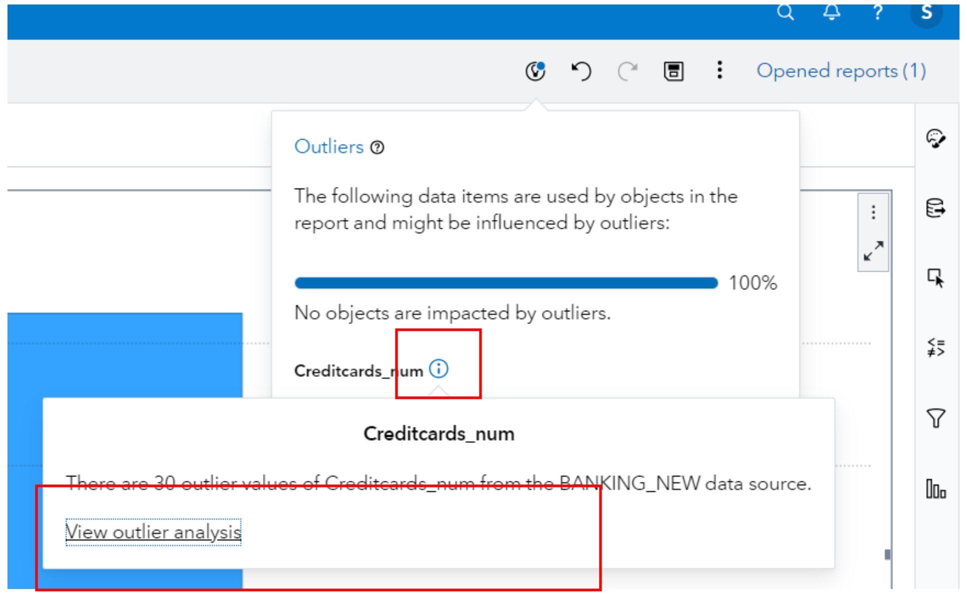
The following data items are used by objects in the report and might be influenced by outliers:

100%

No objects are impacted by outliers.

Creditcards_num

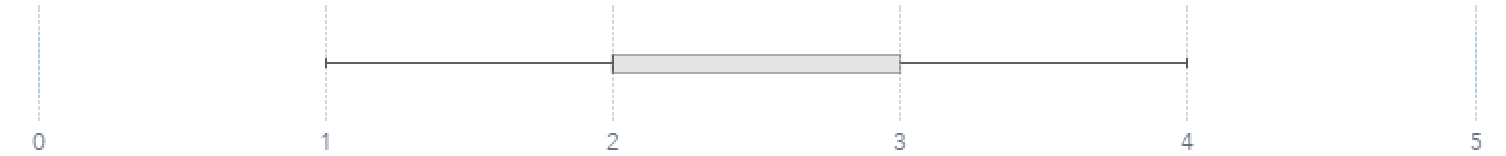
No objects are impacted by outliers.



Outliers of Creditcards_num

Are There Outliers Values of Creditcards_num?

There are 30 outlier values of Creditcards_num. These outliers do not change the overall sum, average, or median by more than 5%.



What are the Details of these Outliers?

	Creditcards_num	Savings_num	Creditutilization_ratio	Amount_avg	Balance_avg	Incon
	5	19	52.969984758	47	547.63975299	51
	0	17	11.546247511	36	471.9675259	50
	0	17	41.609299478	43	397.32784684	44
	5	22	52.700486193	43	275.97261209	56

What Is the Effect of Outliers on Creditcards_num?

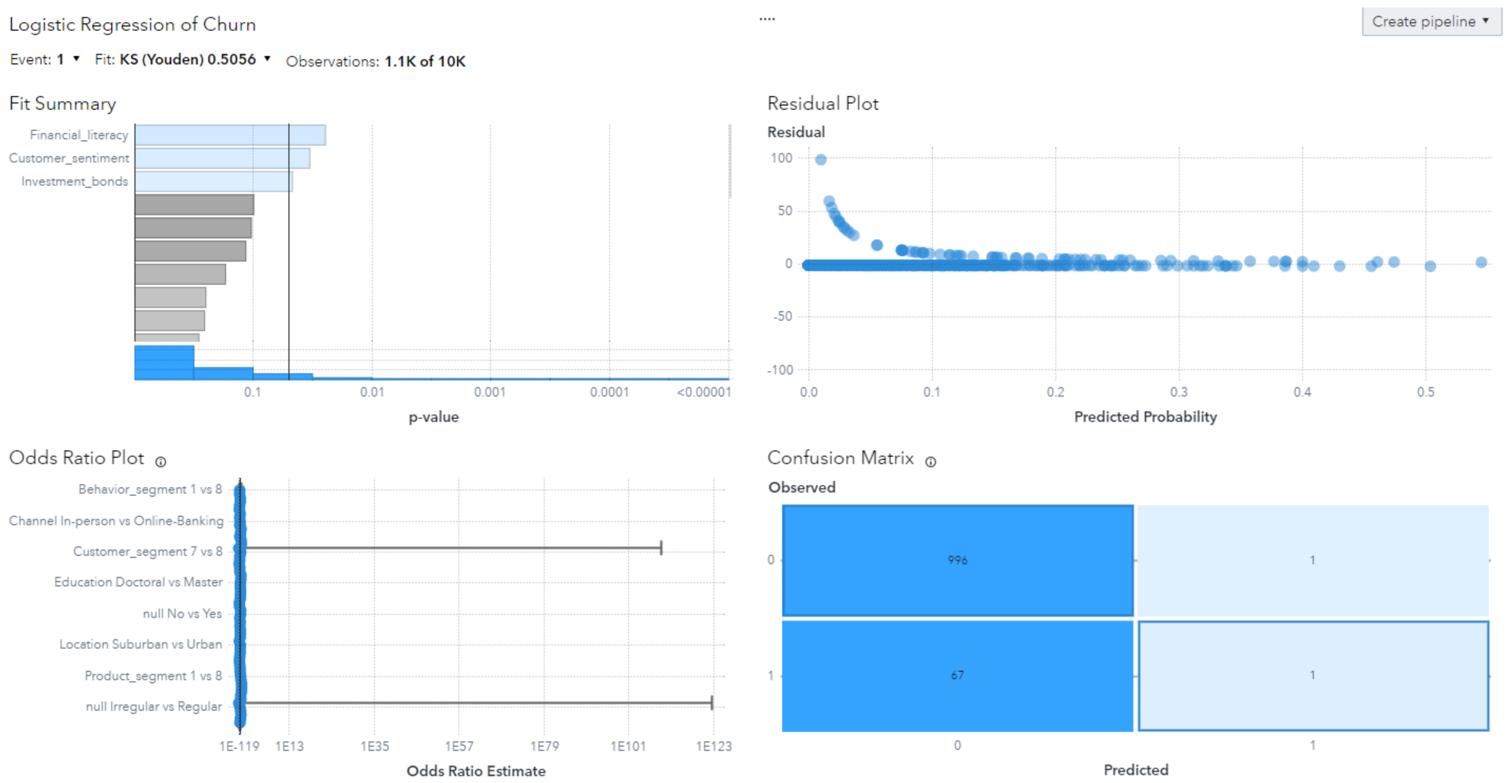
Metric	Including Outliers	Excluding Outliers	Outlier Impact	Difference
Sum	24801	24716	0.34%	85
Average	2.4801	2.4790371113	0.04%	0.0010628887
Median	2	2	0.00%	0

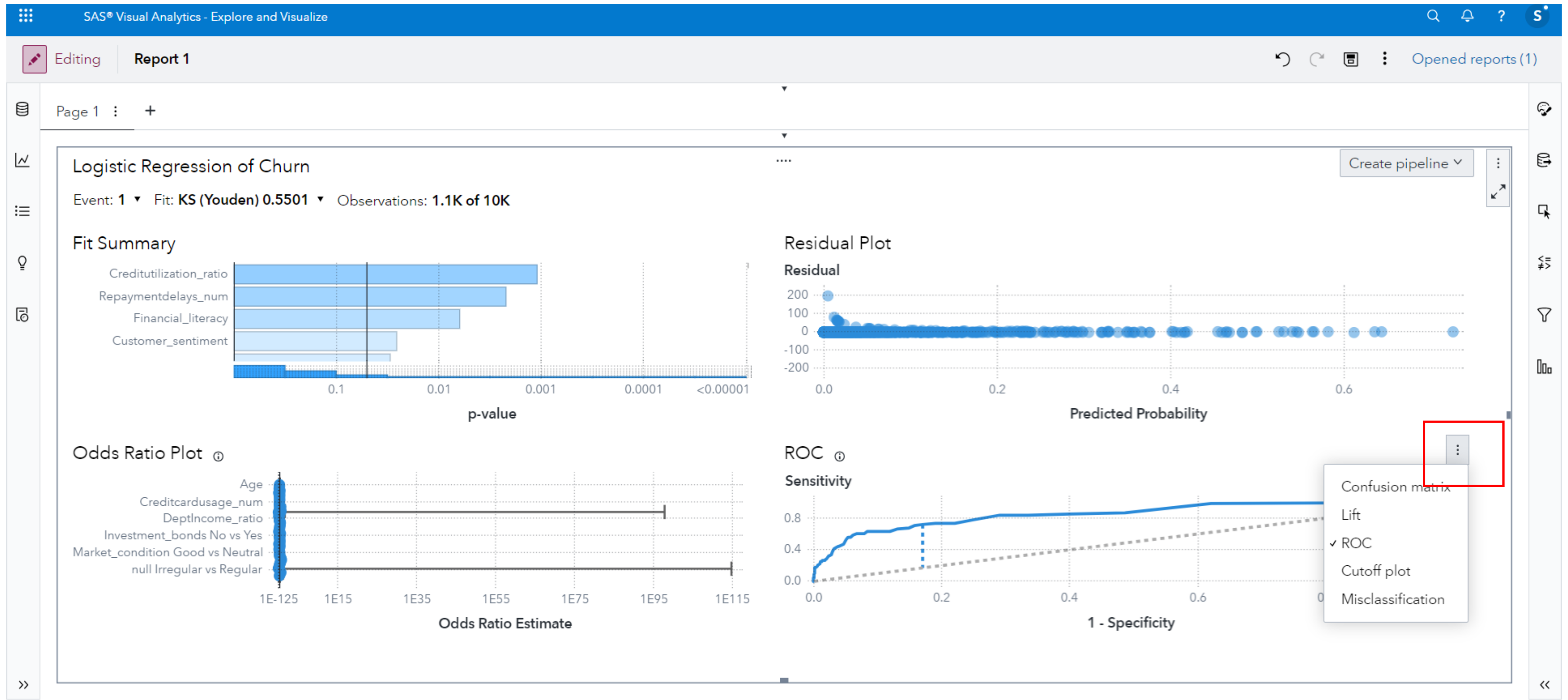
Quick Model Prototyping

Quick Model Prototyping

Logistic Regression

- Before building our models, Data Scientists may need to build a quick model prototype to understand if they could meet the accuracy expectations of their stakeholders and which variables they should consider as most predictive.
- Open a new Page using the “+" sign, then from the “Objects” tab, select “Logistic Regression.” Specify “churn” as the response and choose all the variables for continuous and classification effects except for name and surname. Under the options tab on the right, scroll to the bottom. Under the “Model Display” section, change the assessment plot from confusion matrix to ROC chart.



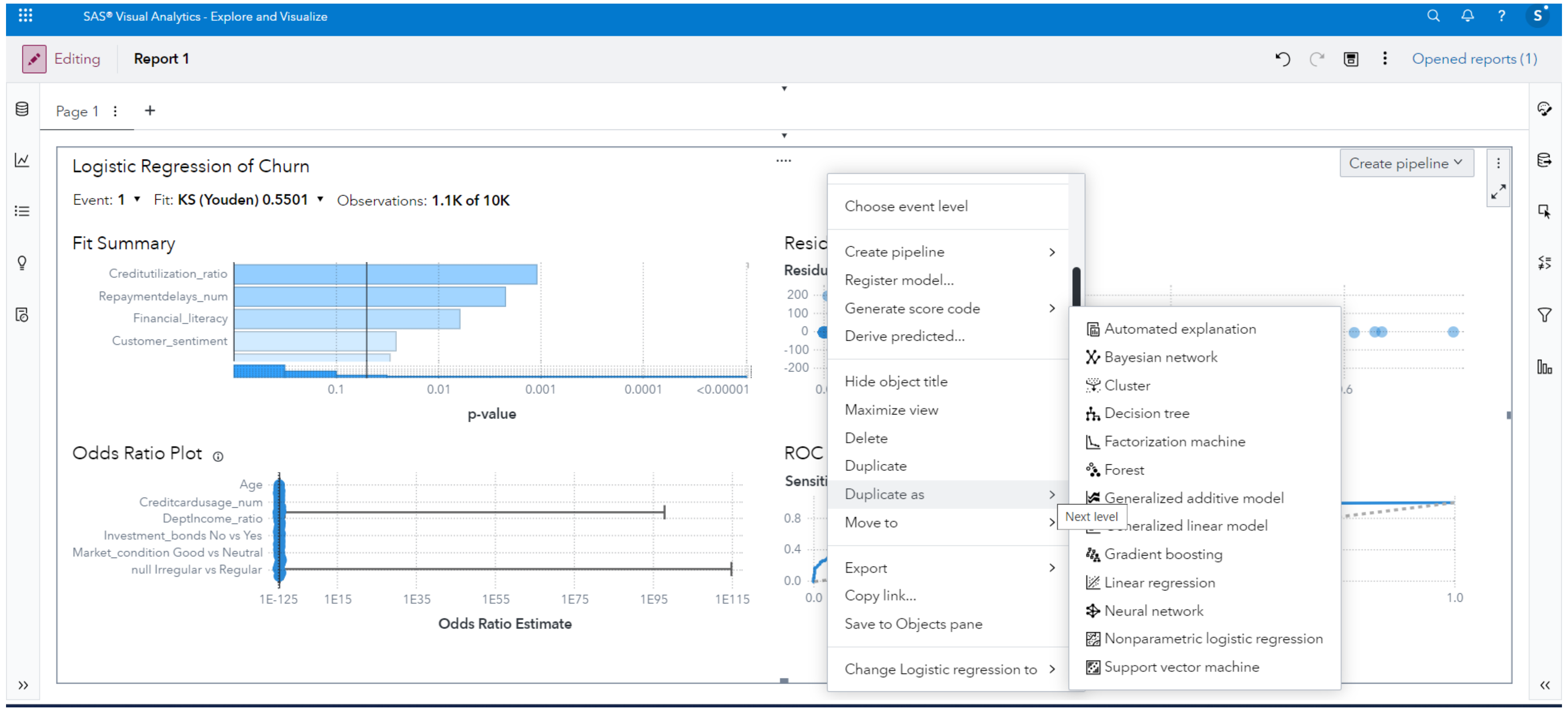


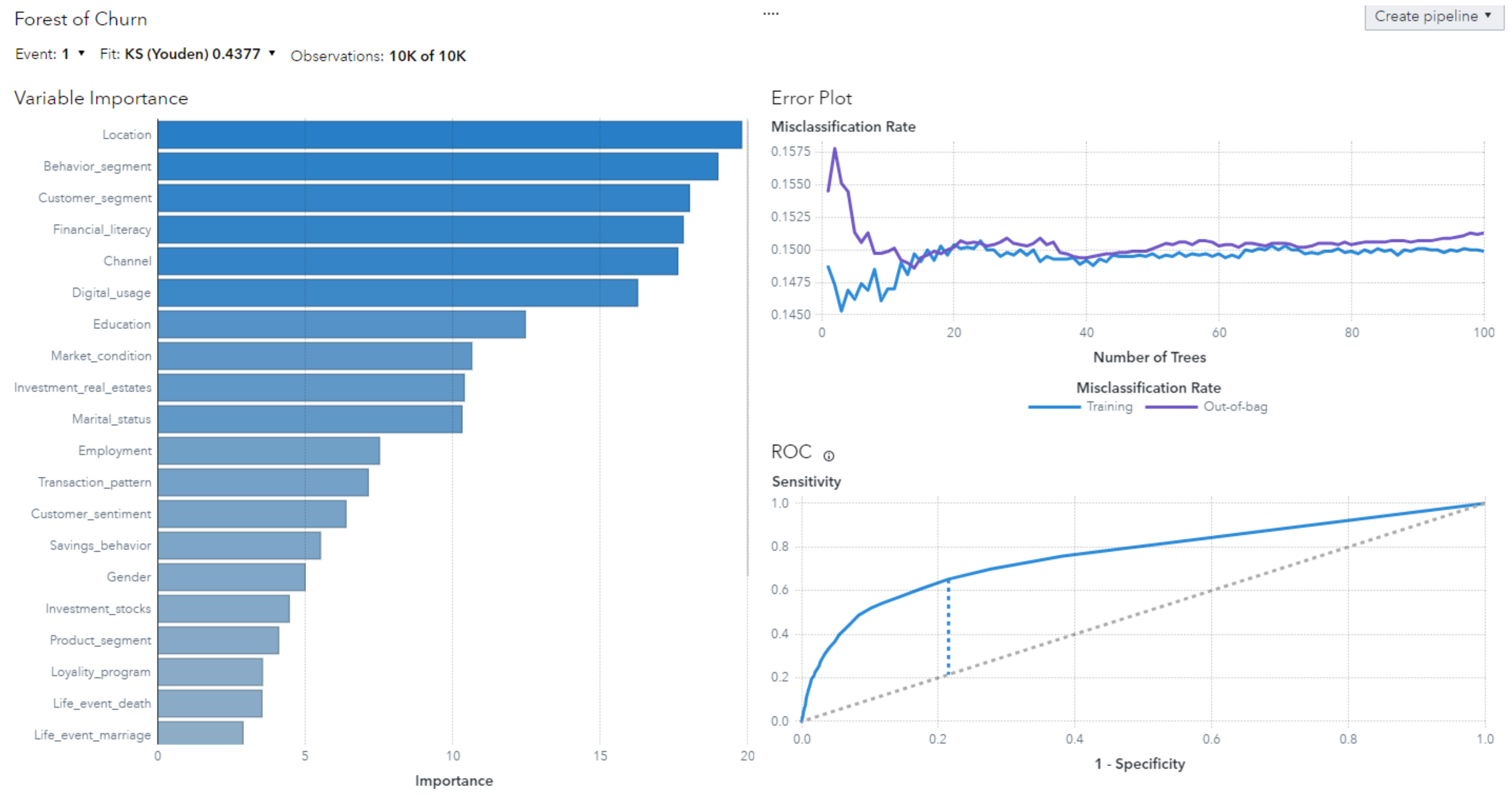
Quick Model Prototyping

Random Forest

You can prototype many different models quickly and automatically to explore how different models potentially would perform by duplicating the logistic regression page. This will keep the variables that you have selected the same for every model.

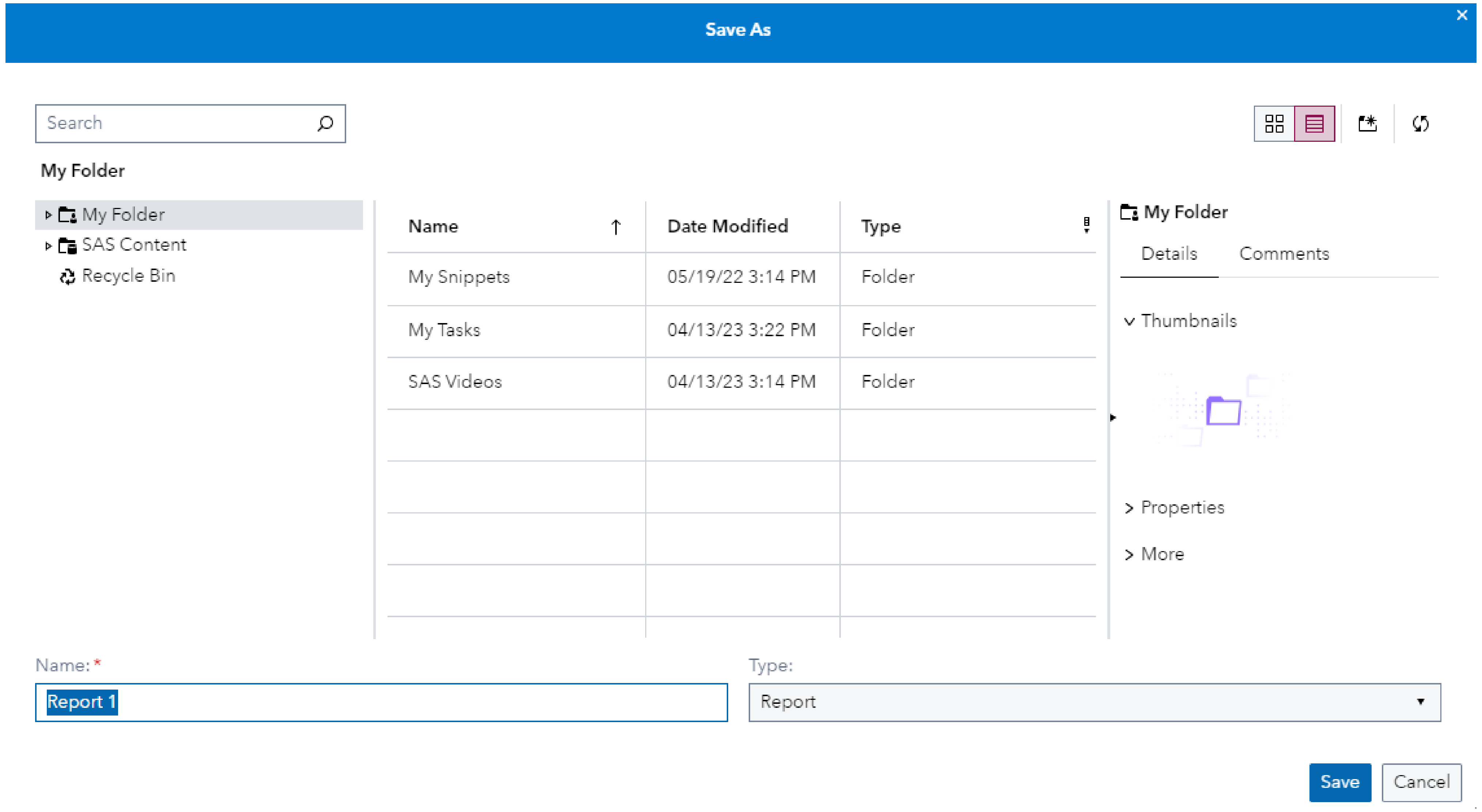
Right-click on the logistic regression template (anywhere around the middle of the screen) and select “Duplicate as” -> “Forest.” Then right-click on the forest template and select “Move to new page.” On the options page under assessment plots, select “ROC chart.”





Save Report

Now that you have gathered your insights, it is time to save your report and move to the modeling phase. Select the “Save” button on the top right to save the report and all the graphics created in the “Explore and Visualize” application. Give an appropriate name and click “Save.” The Data Scientist would then give read access to the project to their stakeholders from the “Share and Collaborate” tab in the applications menu or give “Read and Write” access to the project to their colleagues to perform further analysis.



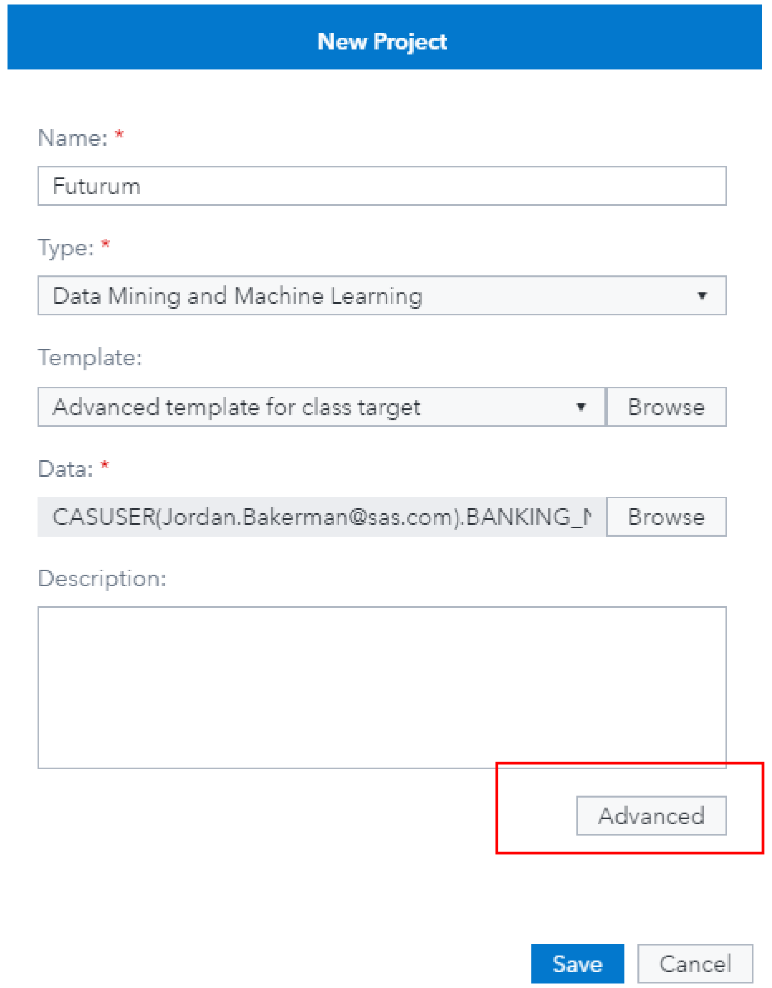
Templates in Model Studio

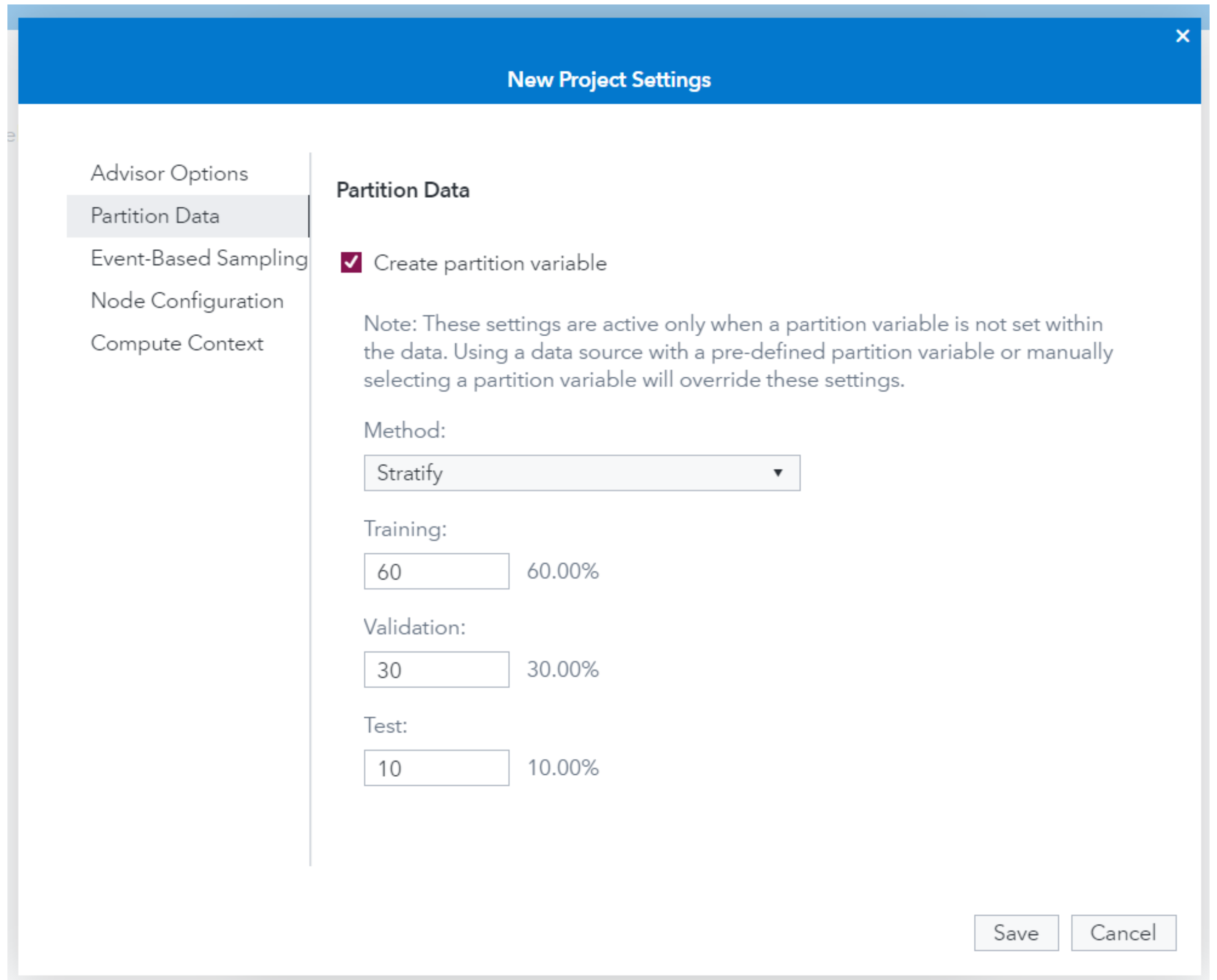
Build Models

Creating a Project

After we have properly explored the data, we move to the "Build Models" application from the applications menu on the left. At the top right, select "New Project." Specify a name for the project and choose the "Data Mining and Machine Learning Type." To begin, browse "Templates" and choose "Advanced Template for Class Target." These templates are provided by SAS, and they automatically create modeling pipelines (including a template for feature engineering), which are configurable and embed modeling and feature engineering best practices based on what the Data Scientist wants to achieve. Finally, choose the BANKING_NEW data set.

You could also click on the "Advanced" button at the bottom to modify the default project settings. These include partitioning options, events-based sampling, excluding variables where the missing values exceed a predefined threshold, and other configuration settings.

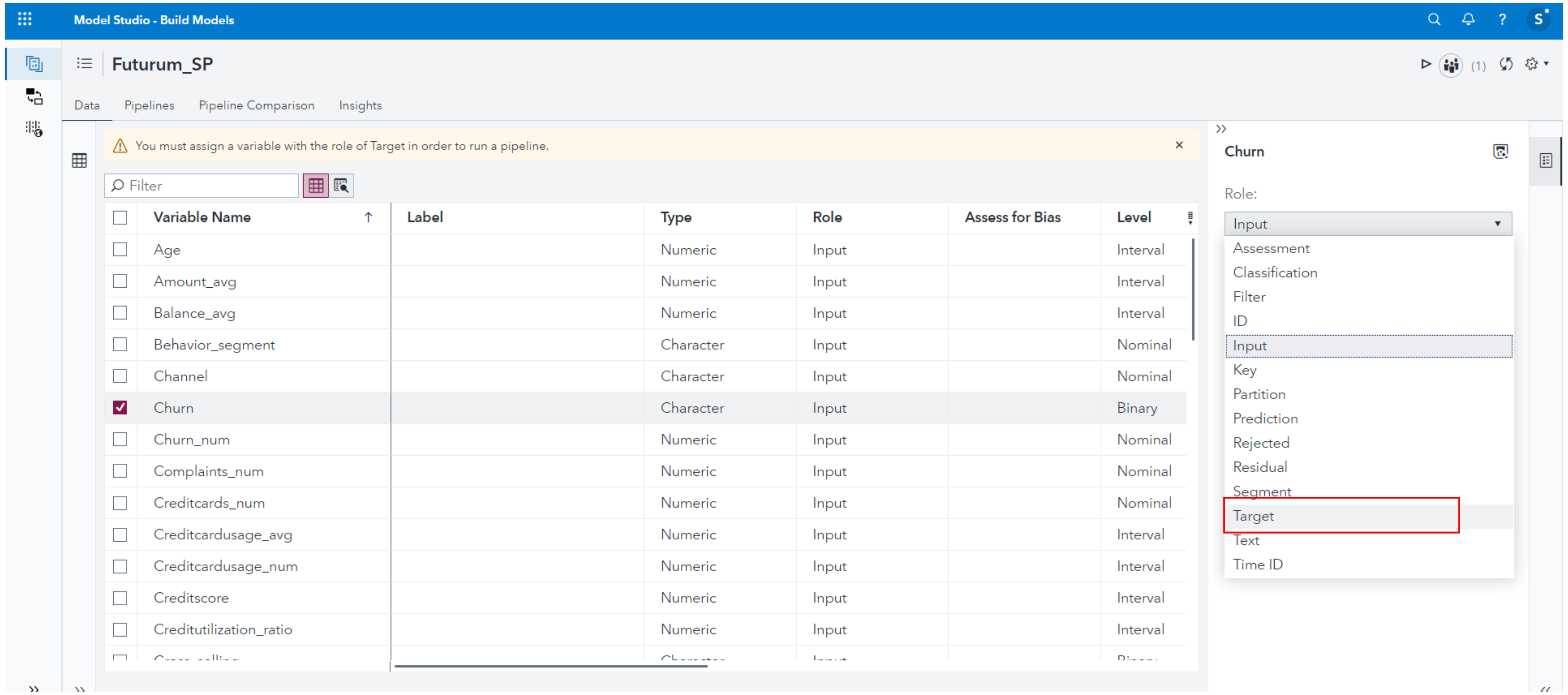




Build Models

Data Tab

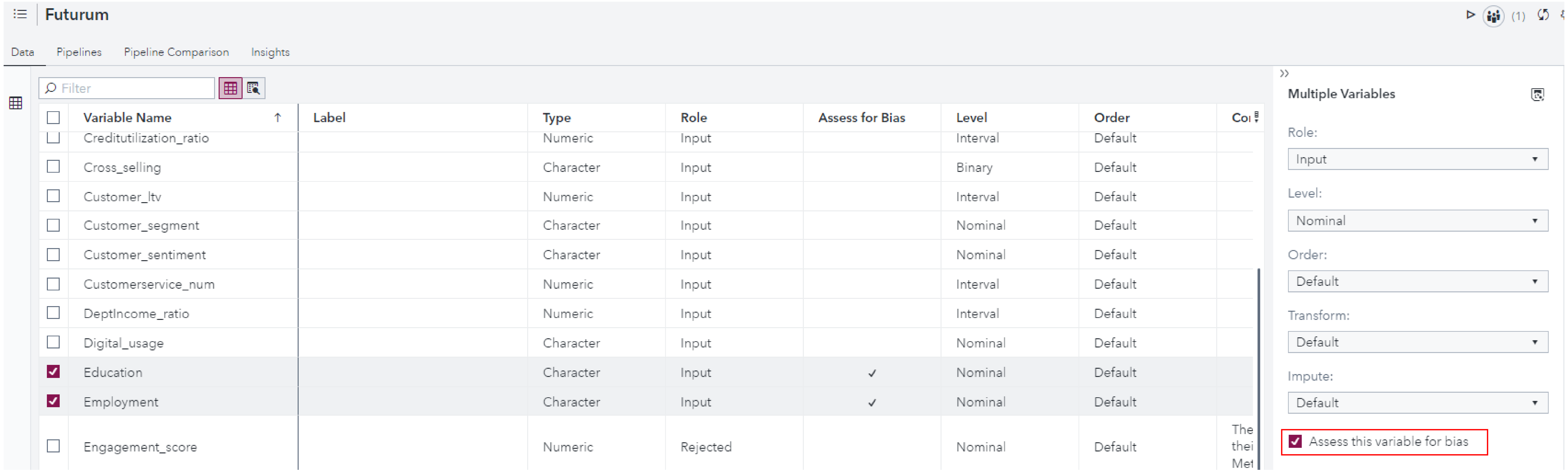
- When the project is created, SAS Viya guides us through the necessary modeling steps. We start with the “Data” tab to select the variables we are going to use for modeling.
- In the “Data” tab of Model Studio, change variable roles where necessary. Select and change “Churn” to “Target” (instead of ‘Input’) and then change “Age,” “Engagement_score,” “Gender,” “Loyalty_program,” “Name,” and “Surname” to “Rejected.” The rest should have a role of Input. The type and levels have the correct default values. Note that the default project partitions are 60% train, 30% validation, and 10% test, but you could change that in the project settings, as we showed in the previous slide.



Build Models

Assess Bias Variables

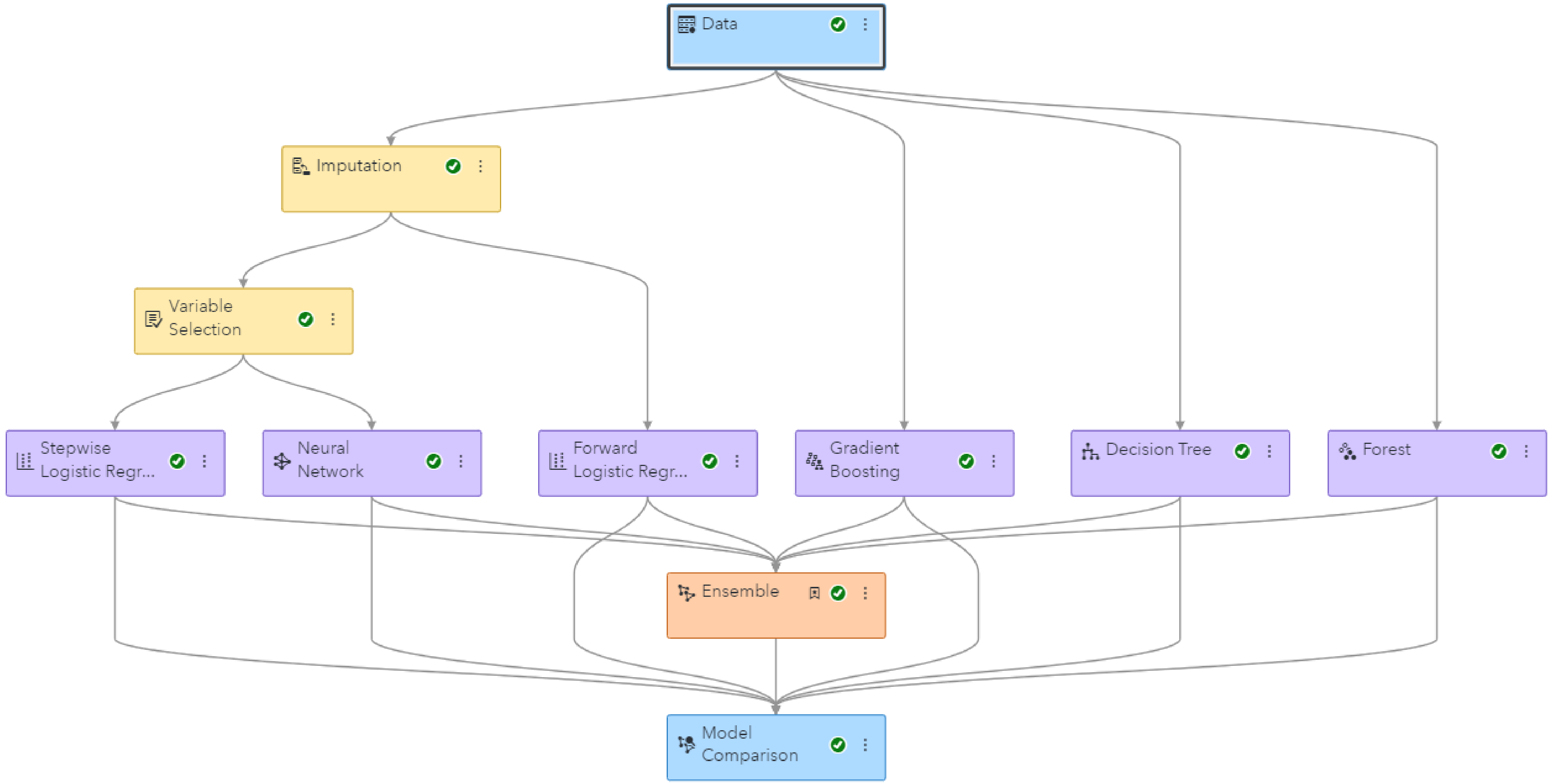
- Finally, choose variables to assess for bias in subsequent analysis. Select the Education, Employment, and Location variables and check the “Asses this variable for bias” box on the right. Each model will now include bias detection by default for the specified variables.
- Note that you don’t have to use variables as “Inputs” to assess potential bias based on those variables in your models. For example, we selected not to use “Gender” in our model and set the role to “Rejected.” However, we can still click the “Asses this variable for bias” button for “Gender’” and assess if our models include any “Gender” bias even though we don’t use this directly. In that way, we protect ourselves from proxy variables that could contain similar information and have that kind of impact on our models. Make sure that you select the “Gender” variable and click the box so bias is assessed for this variable as well.



Build Models

Pipelines

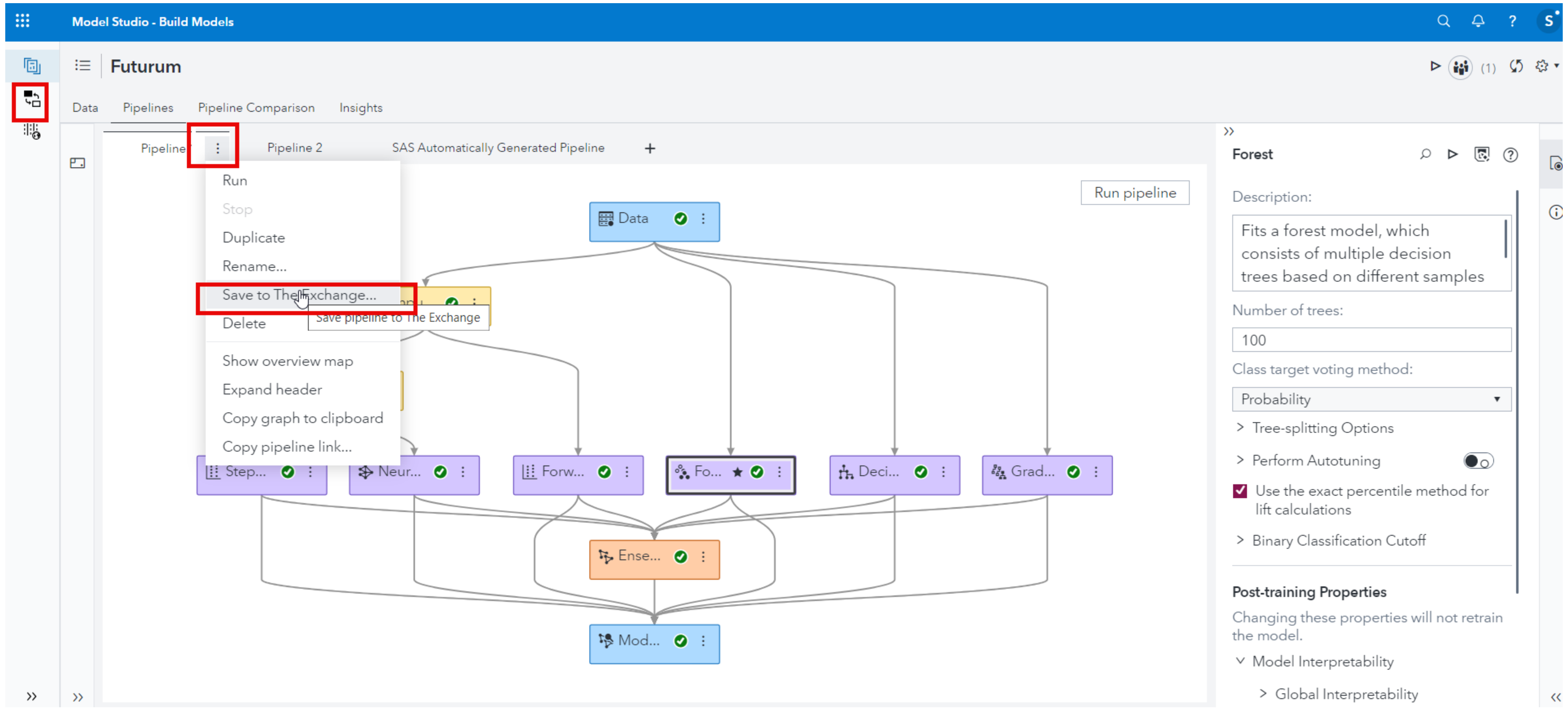
- Now, select the "Pipelines" tab. Notice the Advances template has been created. On the top right, select "Run pipeline" to build and compare all models in this pipeline. This pipeline creates two logistic regression models (one with automated preprocessing, imputation and automatic variable selection, and one without), a neural network, a decision tree, a random forest, a gradient boosting model, and an ensemble of all said models.
- This competition of models includes best practices for preprocessing of Neural Networks and Logistic Regression and also an Ensemble where you can select to use the "Average" or the "Maximum" value for the probabilities for our target. Before configuring this further, let's have a look at the results.

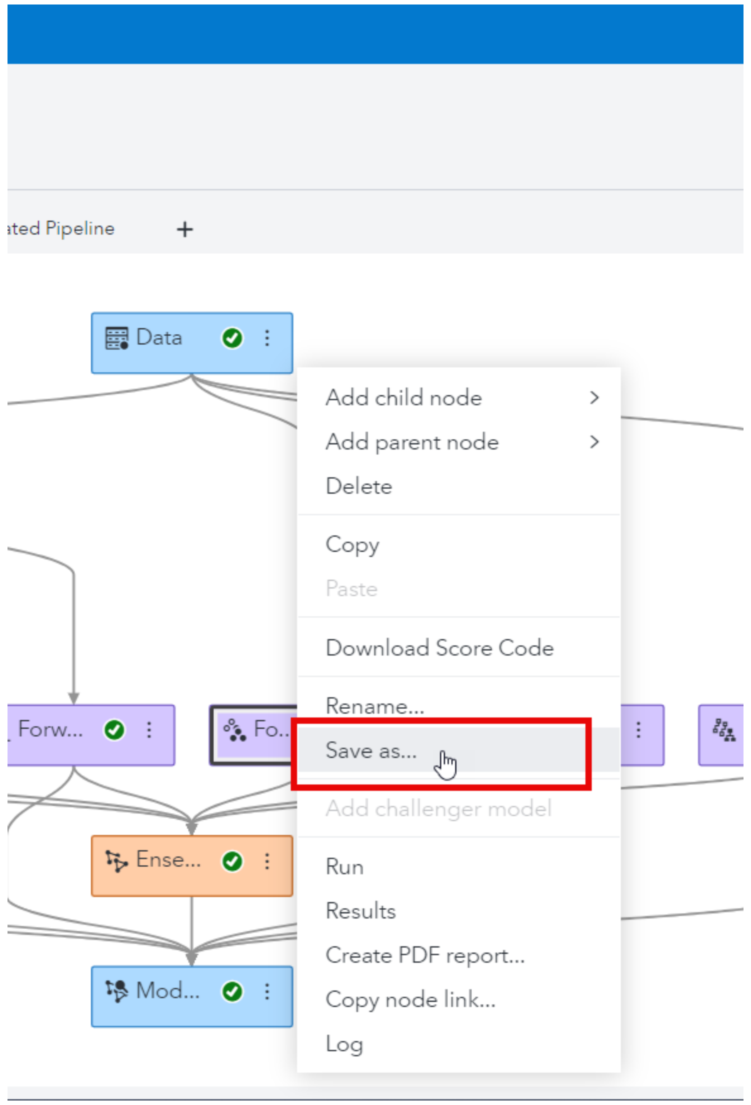


Build Models

Share Reusable Assets

- There may be cases where a Data Scientist would like to create custom pipelines by adding/removing nodes and setting certain hyperparameters/options for each model, using the nodes' options on the right of the screen. The Data Scientist will want this pipeline to be used by other business functions as well. In this case, there is a simple way to save his pipelines in a collaborative space called "The Exchange," so this custom template is made available to his team and other business functions every time they create a modeling project.
- To do this, click on the three-dot icon next to the pipeline name and click "Save to the Exchange." "The Exchange" can be accessed by using the little icon at the left of the screen circled in red in the graph below. You can also create custom nodes or modify existing ones by using the options pane and then save them to "The Exchange" as well so they can be reused. To do this, right-click on a modeling node and select "Save as."





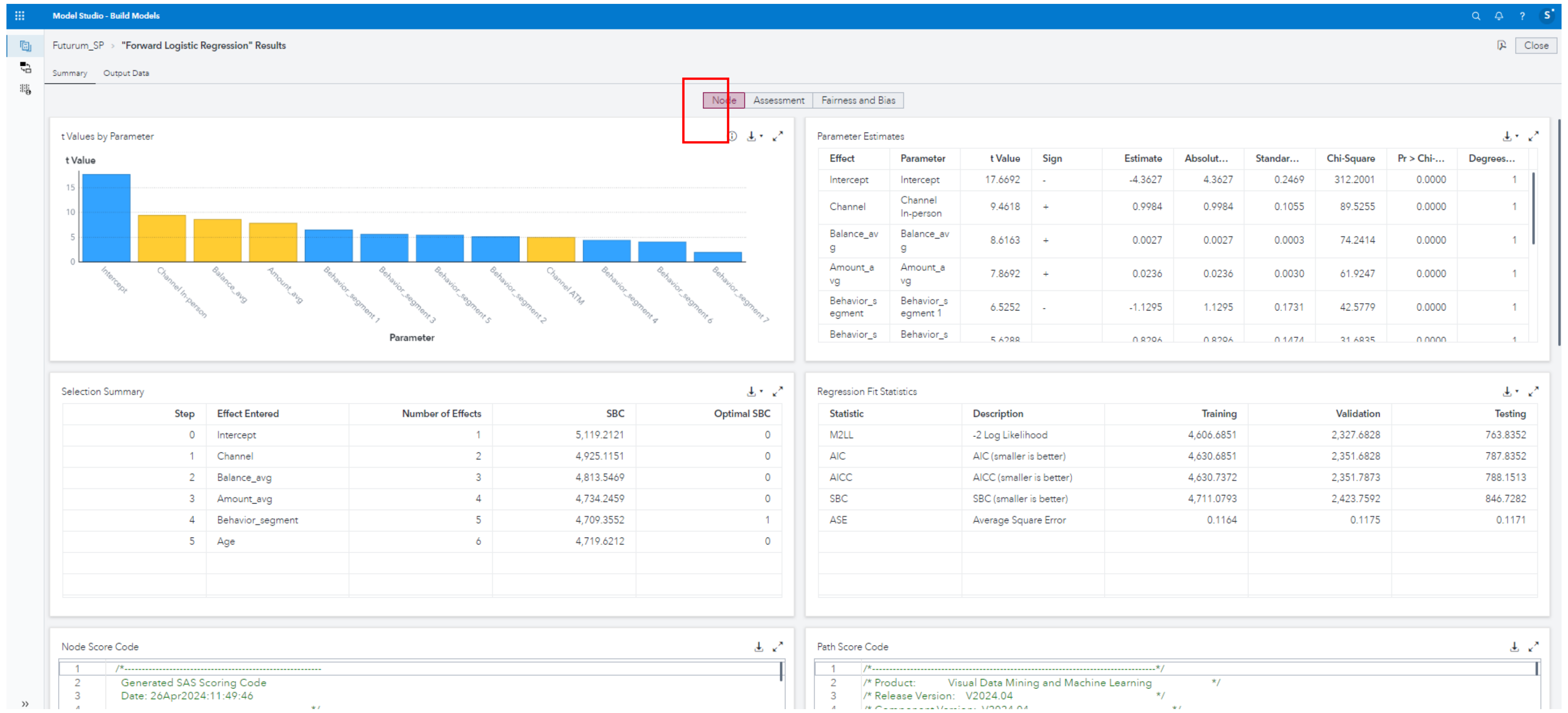
Model Competitions

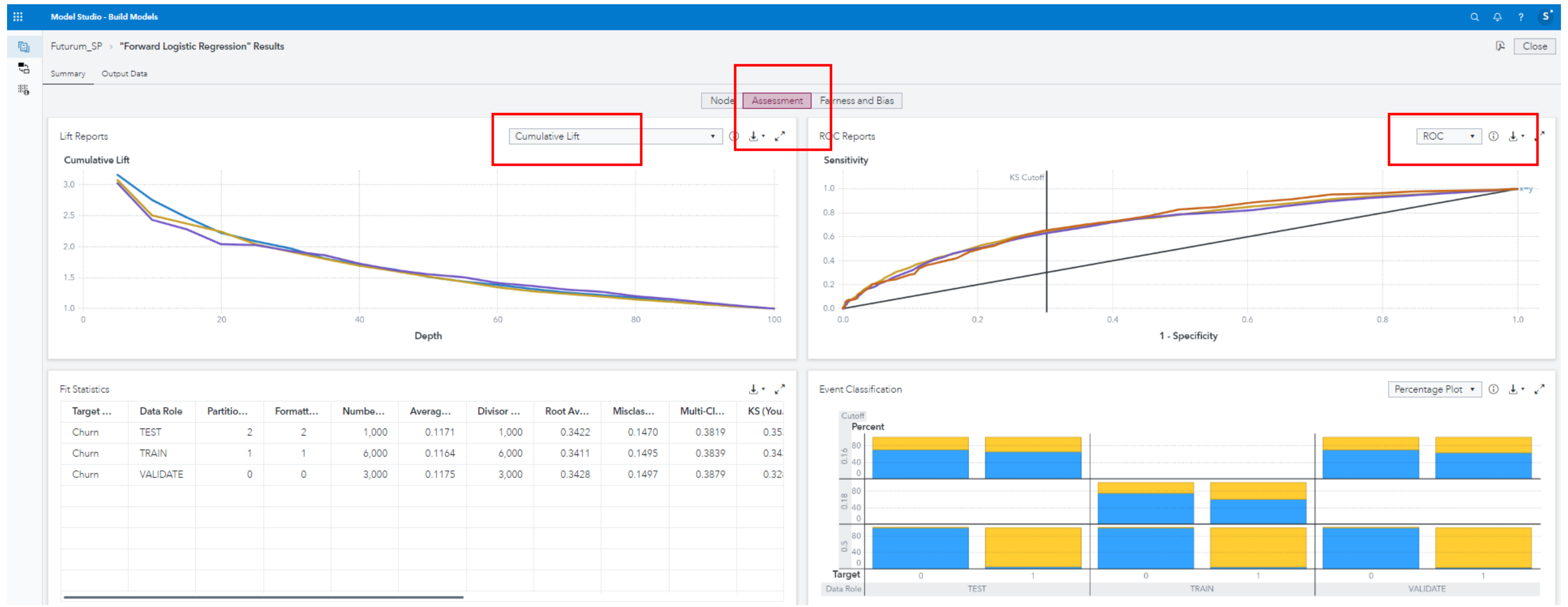
Model Competitions

Model Results

- Right-click on any model node in the completed pipeline to view results. For example, the forward logistic regression model shows t-values by parameter, parameter estimates, selection summary, fit statistics and score code. Maximize the graph to see the coefficients and an explanation of the graph in natural language.

- Move to the “Assessment” tab, which shows fit statistics and model graphics for the specified node. The first graphs you see are the “Cumulative Lift,” “ROC,” and “Event classification” reports. Click on the information “I” icon next to the name of the reports (top right in the graphs) or maximize it to view the full screen using the arrow icon, which expands the view. An automatic explanation of the graph appears so users can understand exactly what they see and how to interpret the graph. You can also switch the “Cumulative Lift” graph to view “Lift,” “Gain,” “Captured Response Pct,” etc. You can also switch the ROC graph to view the “Accuracy” and “F1 Score.” The same goes for the “Events Classification” graph, where you can view this information in terms of percentages, counts or as a table. All these metrics are developed automatically for all the models you created.





Build Models

Pipeline Model Comparison

- Exit the results of the model node and go back to your pipeline. Click on the “Model Comparison” node. On the right of the screen, you can configure the options so the node selects the best model based on your criterion of choice and also the partition you want to use. Leave the default settings.
- Right-click on the model comparison node and select “Results” to view model comparison metrics for this pipeline. The node tab displays that the ensemble had the best KS Youden statistic, and the Assessment tab shows fit statistics and graphics for each data partition and each model.

Node | Assessment

Model Comparison

Champi...	Name	Algorith...	KS (You...	Accuracy	Averag...	Area Un...	Cumula...	Cumula...
★	Ensemble	Ensemble	0.5211	0.8540	0.1060	0.8051	3.5526	35.5263
	Gradient Boosting	Gradient Boosting	0.3397	0.8480	0.1262	0.6750	2.4894	24.8939
	Stepwise Logistic Regressio n	Logistic Regressio n	0.4483	0.8580	0.1074	0.7810	3.3553	33.5526
	Forward Logistic Regressio n	Logistic Regressio n	0.4849	0.8620	0.1039	0.7947	3.5526	35.5263

Node | Assessment

Fit Statistics

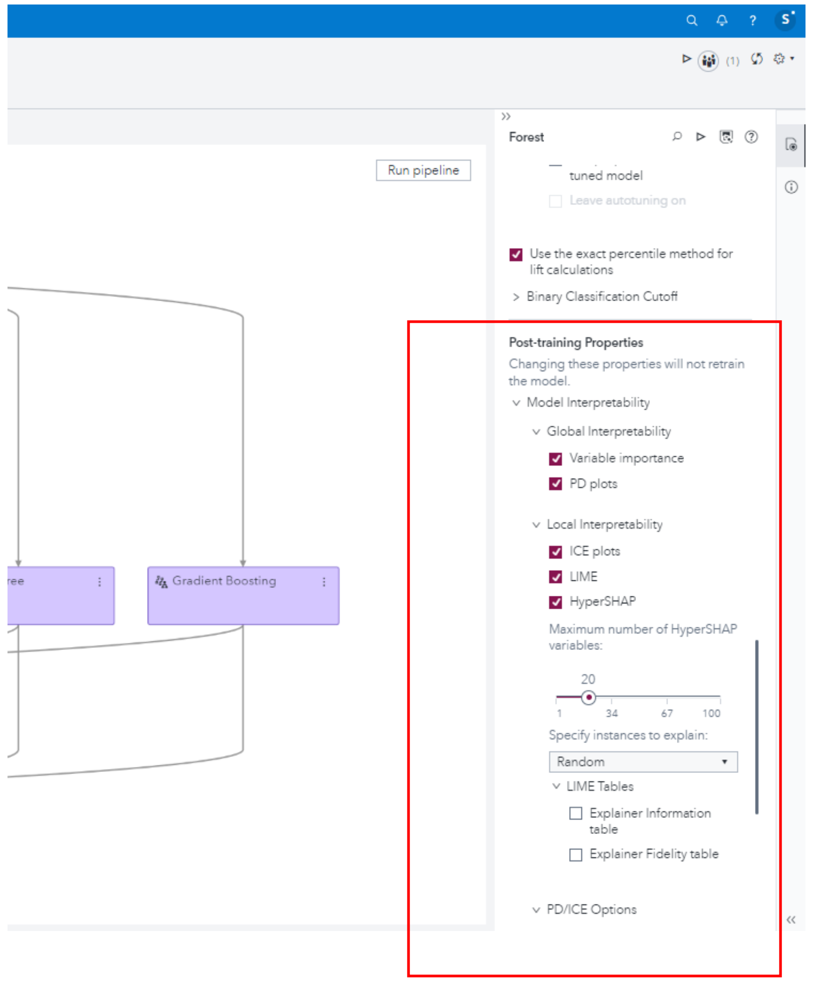
Statistic...	Train: G...	Validate...	Test: Gr...	Train: St...	Validate...	Test: St...	Train: E...
Area Under ROC	0.6944	0.6395	0.6750	0.7769	0.7490	0.7810	0.8383
Average Squared Error	0.1253	0.1264	0.1262	0.1059	0.1090	0.1074	0.0986
Divisor for ASE	6,000	3,000	1,000	6,000	3,000	1,000	6,000
Formatted Partition	1	0	2	1	0	2	1
Gamma	0.5778	0.4503	0.5147	0.5641	0.5082	0.5723	0.6928

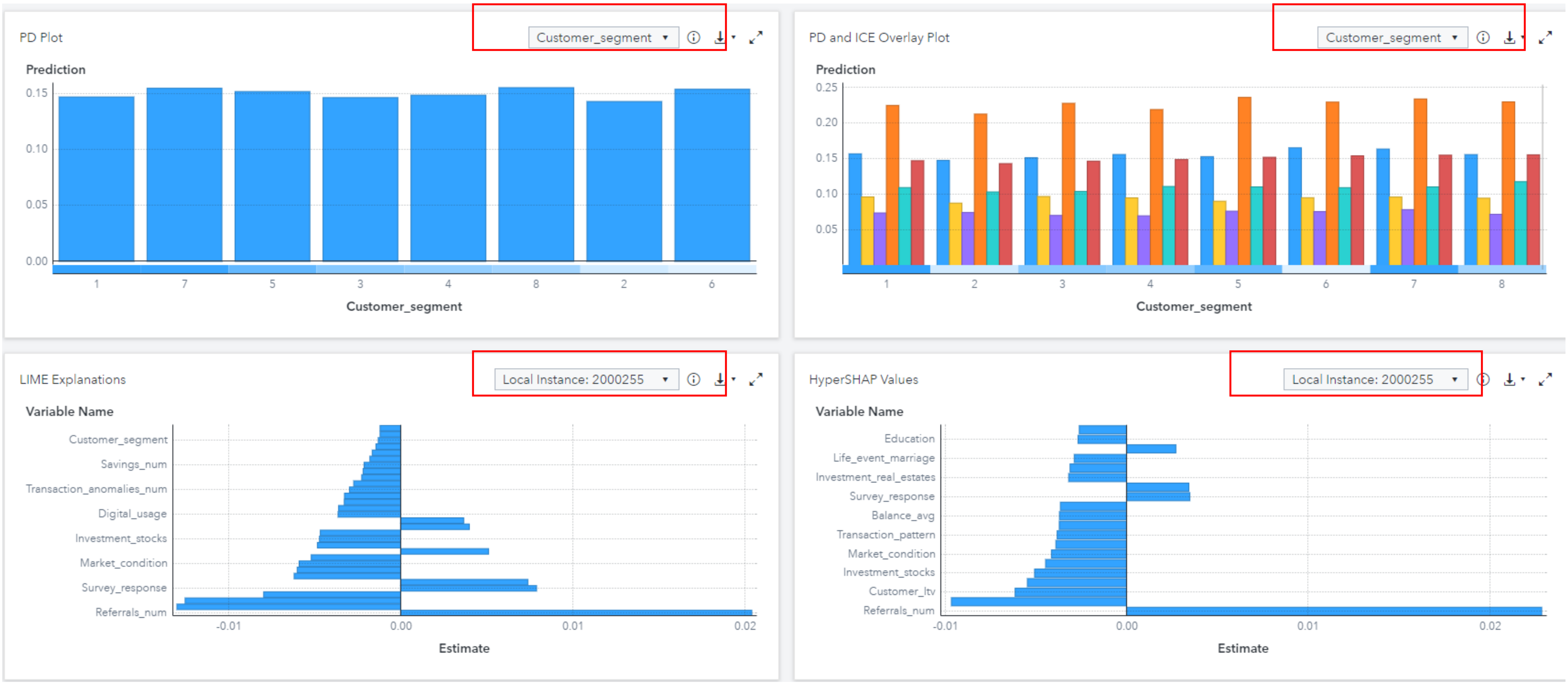
Explainability

Explainability of Models

Global & Local Interpretability Plots

- Now it's time to explain our predictions. Click the "Forest" node on the pipeline and on the options pane, scroll to the bottom to see post-training properties. Open Global Interpretability and check variable importance and PD plots. Under "Local Interpretability," select ICE plots, LIME and HyperSHAP. Run the pipeline.
- Right-click on the "Forest" node and select results and then "Model Interpretability." Notice each of the desired explainability plots ha s been created for each variable. To interpret the results, the Data Scientist can click on the information "I" icon on the top right of each graph. For global interpretability plots (PD, PD & ICE overlay), a Data Scientist can switch between the predictive variables to see how they behave in the model. For local interpretability, you can select the different instances to examine. (These are selected randomly, but you could have chosen specific observations to examine via the node's option pane).



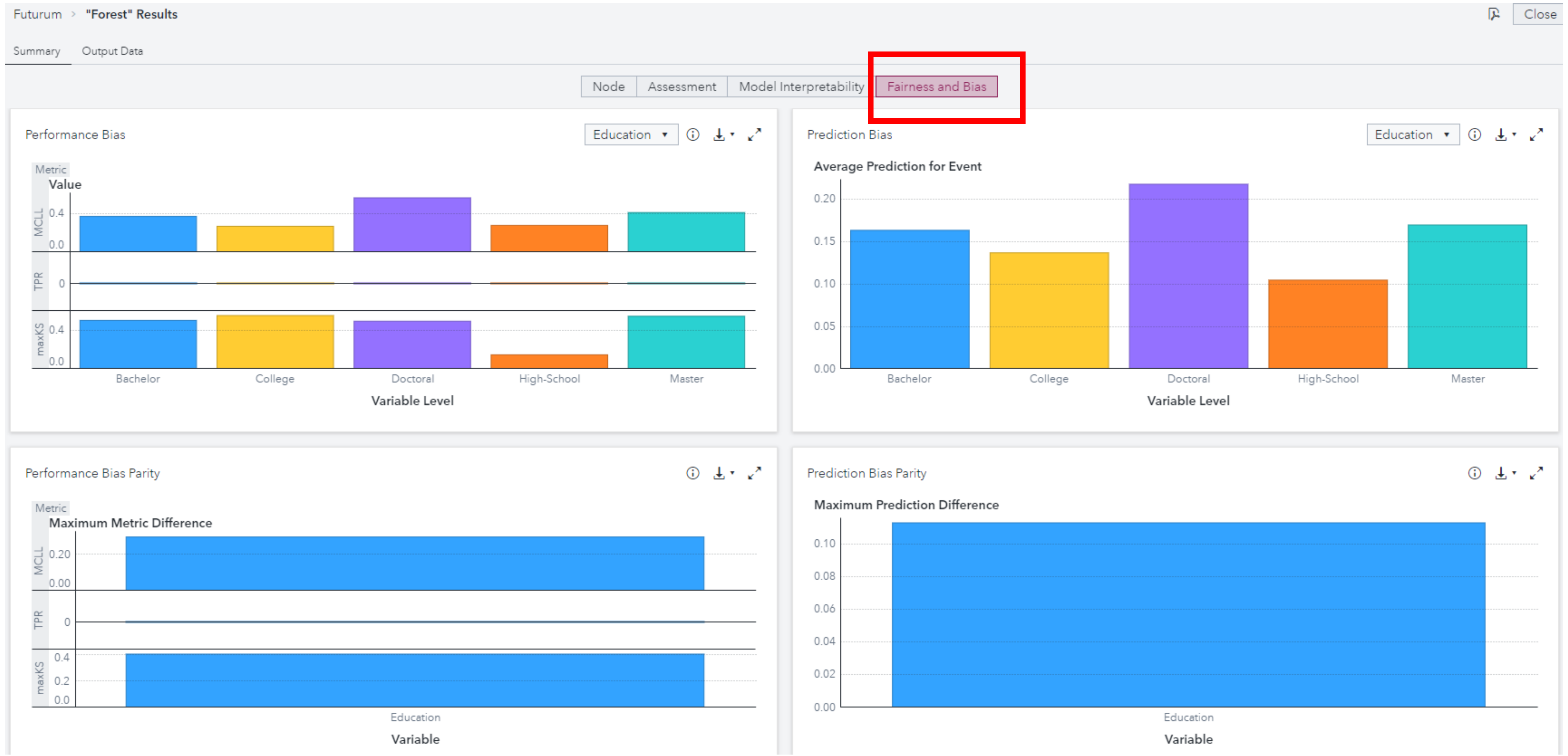


Bias Detection

Fairness & Bias Detection

Performance & Prediction Bias

- From the “Results” tab of your model), select “Fairness and Bias.” This tab was created because we selected assess bias for the chosen variables in the initial data tab. Remember that we can also see variables that we didn’t use in the models (such as we did with “Gender”).
- The results show different metrics for each level of the chosen variable, and the user can toggle between variables to see each bias detection result for the three variables. The bias detection is performed in two ways to ensure the fairness of the predictions. Both in terms of the performance of the developed models in each group (to see if the model fails to perform well for specific groups) but also in the average prediction bias for each group where we can see if we need to make any adjustments to the model. As in the previous graphs, users can press the information icon “I” on the top right of each graph to understand in natural language the meaning of the graphs.



Model Reports

Model Reports

Documenting Models

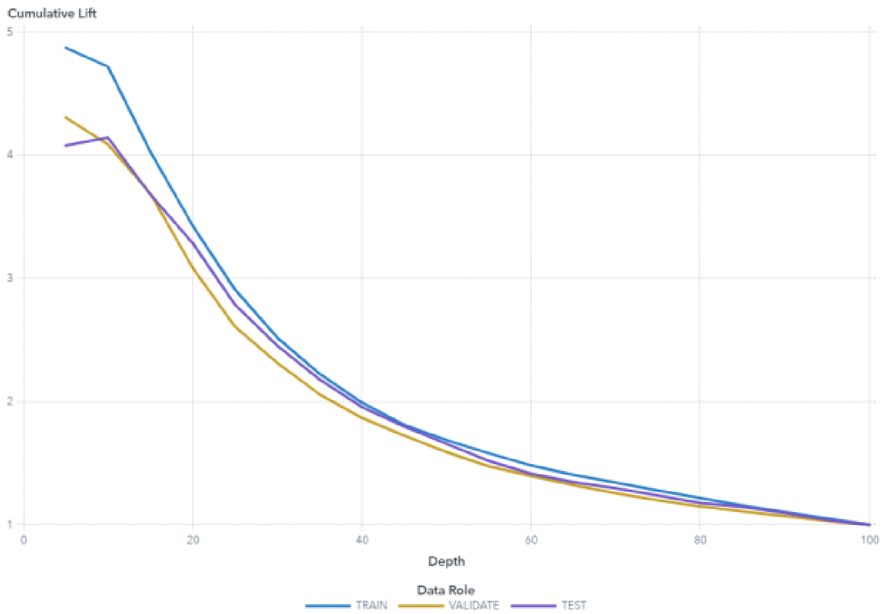
- All the modeling work we have done up until now must be documented. Sometimes, this is essential to take place not only at the project level but also at the model level for regulatory purposes. To save the full model report, select the desired model (use “Forest,” for example) and go to results. On the top right, select the PDF icon and choose Export. SAS creates a PDF of all the model information and graphics. The report provides detailed natural language-generated descriptions of the results.

Futurum

"Forest" Results

by: Jordan Bakerman

Cumulative Lift



The VALIDATE partition has a Cumulative Lift of 4.09 in the 10% quantile (depth of 10) meaning there are 4.09 times more events in the first two quantiles than expected by random (10% of the total number of events). Because this value is greater than 1, it is better to use your model to identify responders than no model, based on the selected partition.

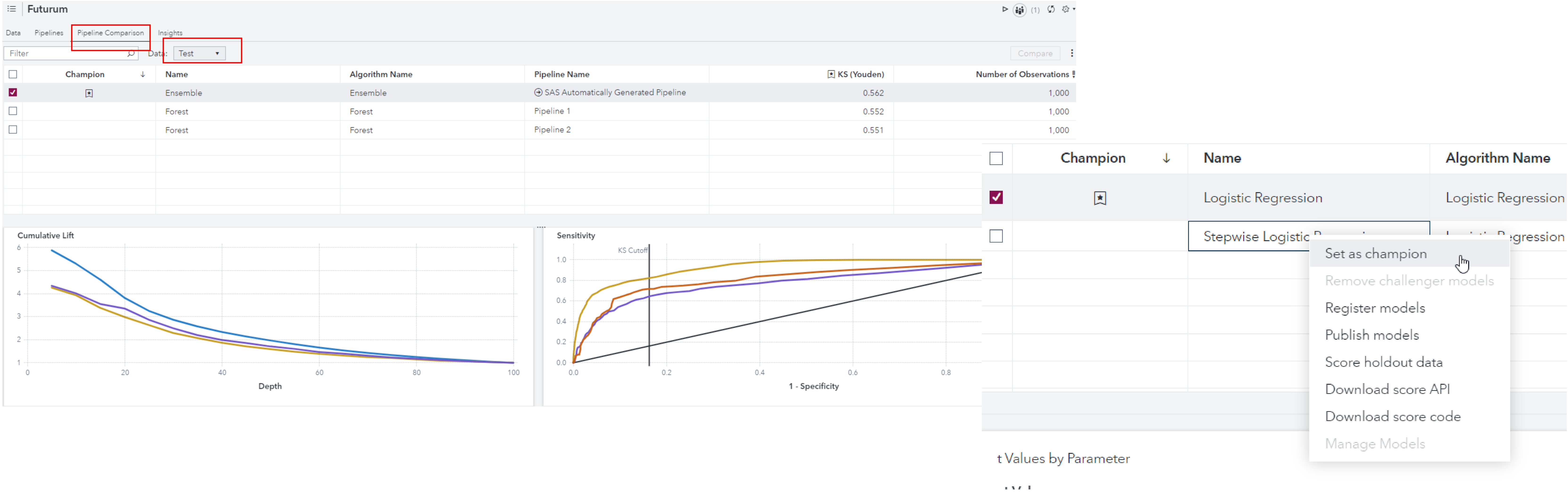
The TRAIN partition has a Cumulative Lift of 4.72 in the 10% quantile (depth of 10) meaning there are 4.72 times more events in the first two quantiles than expected by random (10% of the total number of events). Because this value is greater than 1, it is better to use your model to identify responders than no model, based on the selected partition.

The TEST partition has a Cumulative Lift of 4.14 in the 10% quantile (depth of 10) meaning there are 4.14 times more events in the first two quantiles than expected by random (10% of the total number of events). Because this value is greater than 1, it is better to use your model to identify responders than no model, based on the selected partition.

Pipeline Competitions

Pipeline Comparison

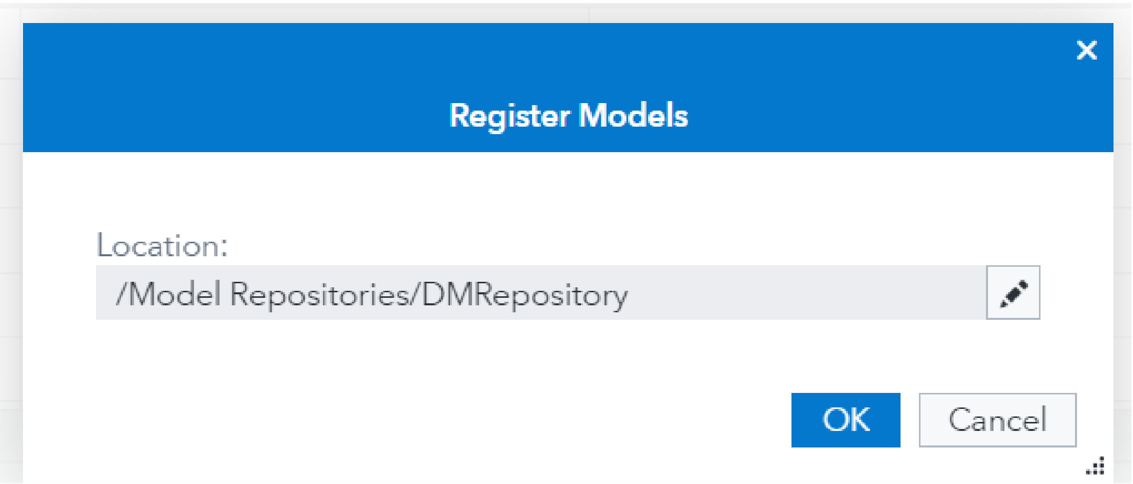
Now that we have run some pipelines and are happy with the model we developed, we can move to the next tab. The “Pipeline Comparison” tab will compare each pipeline using the test data and display the best model for each pipeline. Select the champion model in this view by clicking on it, which displays that model's summary information and graphics. You can also view this information for the “train” and “validate” data by using the button as shown in the graph below. For business reasons (explainability or bias, for example), you may want to put a different champion model into production than the one that is set by default based on its accuracy. To do that, just right-click on one other model and select “Set as champion.”

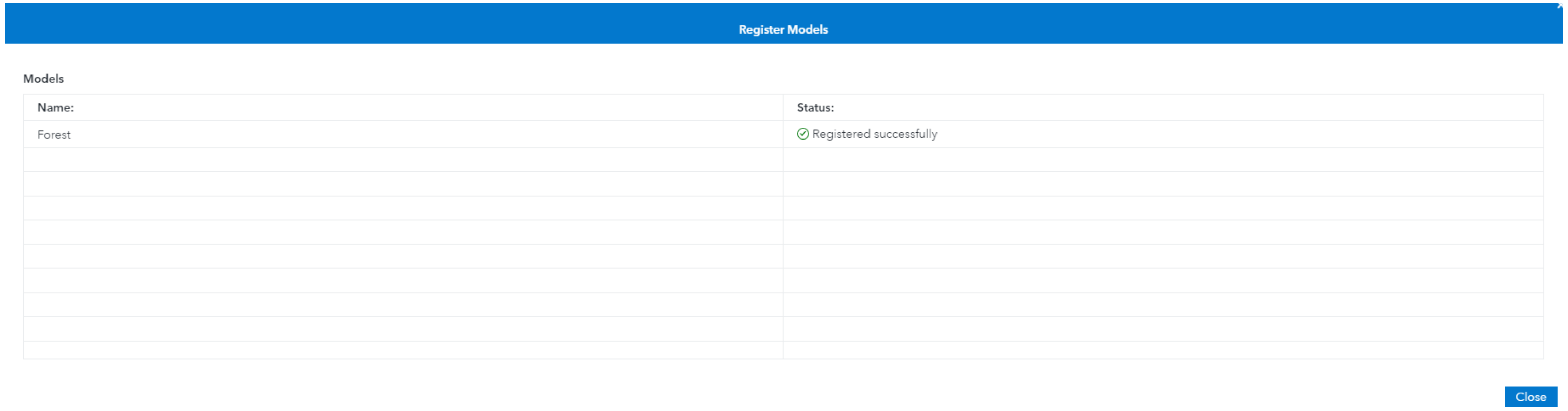


Model Registration

Model Registration

- At this point, the Data Scientist has created several models in multiple pipelines, and the Pipeline Comparison tab has determined the champion model. To register this model and all deployment artifacts to the central model repository, simply right-click on the champion model in the Pipeline Comparison tab as we did before and select “Register models.” Click on “OK” to save the default location.
- The work now moves to the ModelOps team. Note that more than one model can be registered per project to set champion/challenger competitions afterward and see which models perform best or are more stable over time. A Data Scientist can then choose to switch the models to production later.

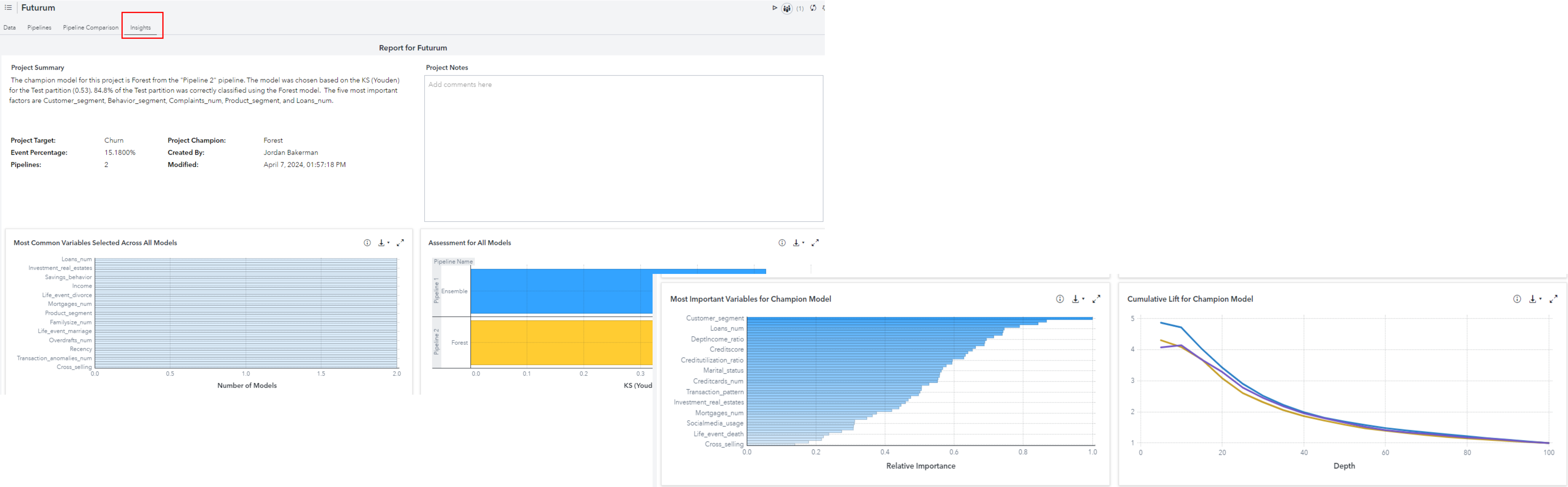




Project Insights Report – Documentation

Project Insights and Reports

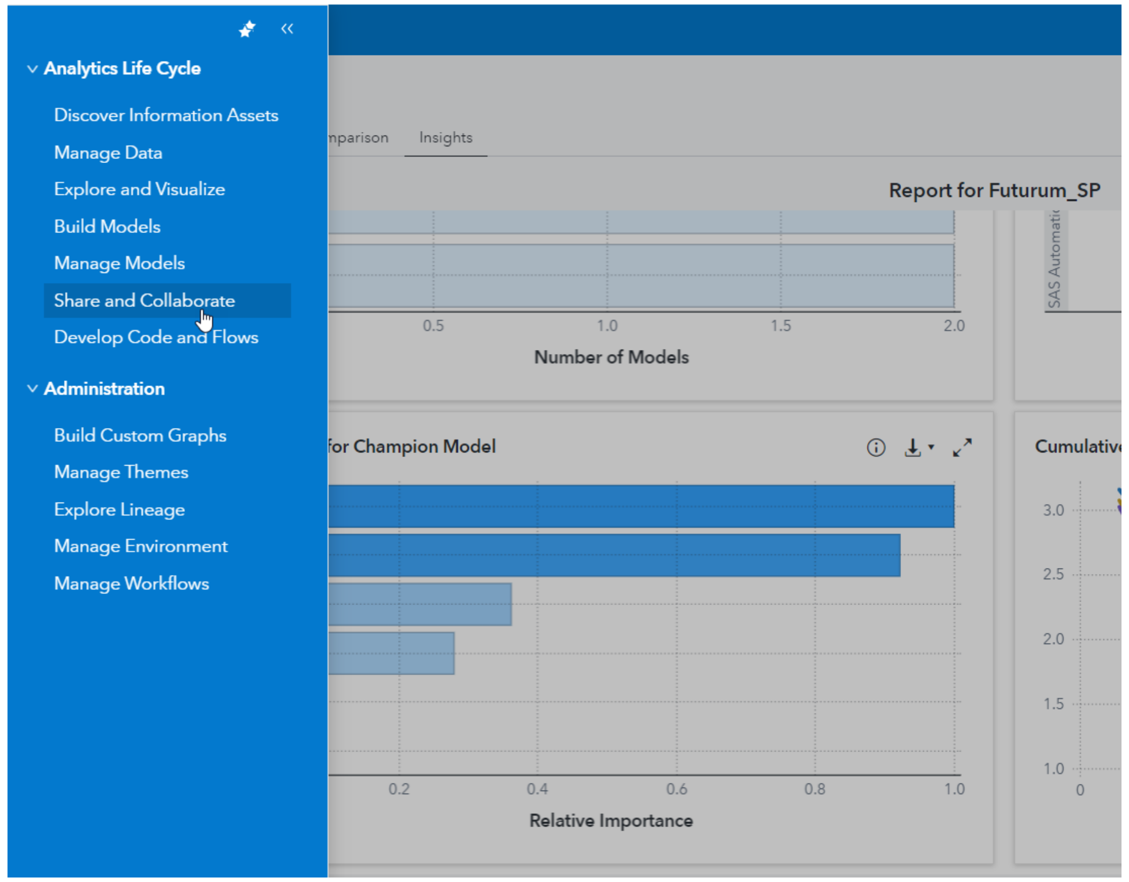
The last thing we want to do for this project is to check our insights and documentation, which are automatically generated. Move to the "Insights" tab, which provides a full description of the project using natural language generation. It provides a project summary and champion model metrics. A report can be created by selecting the PDF icon on the top right of the insights page. You can also write custom notes to document any relevant business information you need to remember about the project or information that needs to be shared with your stakeholders.

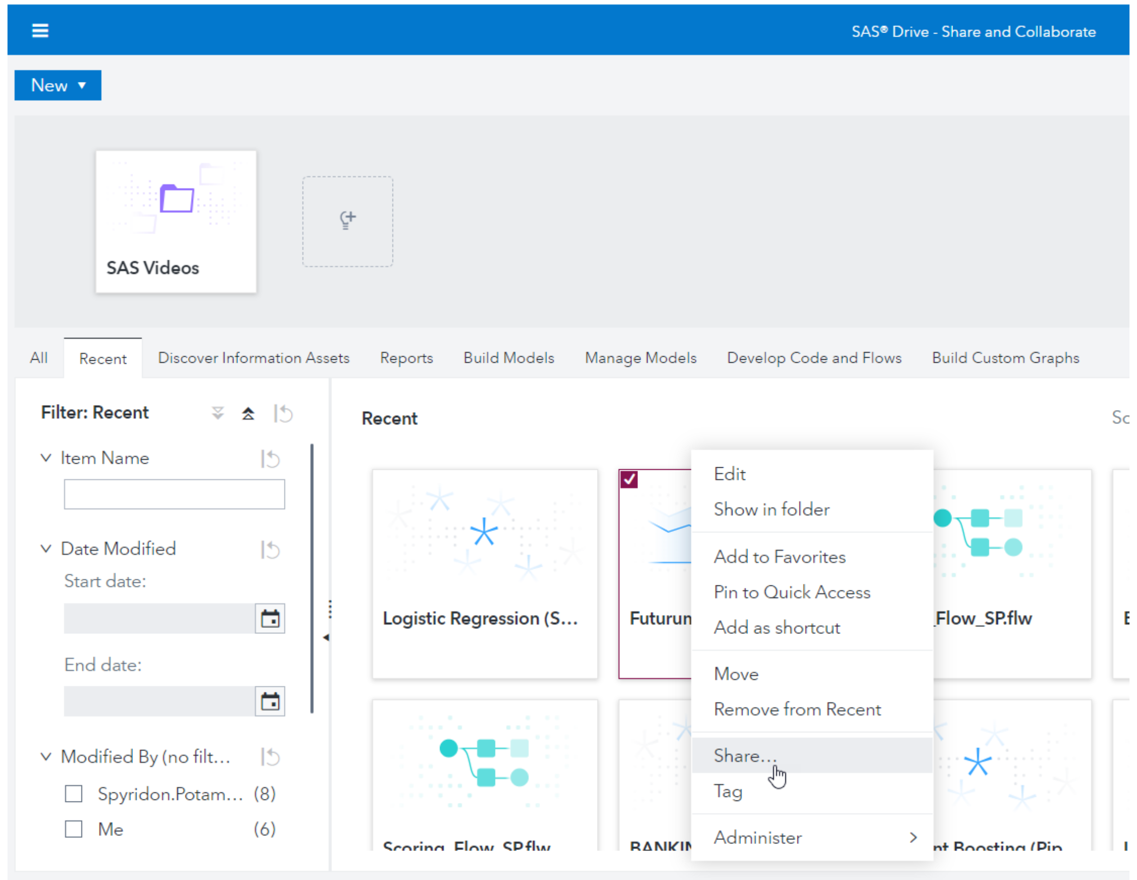


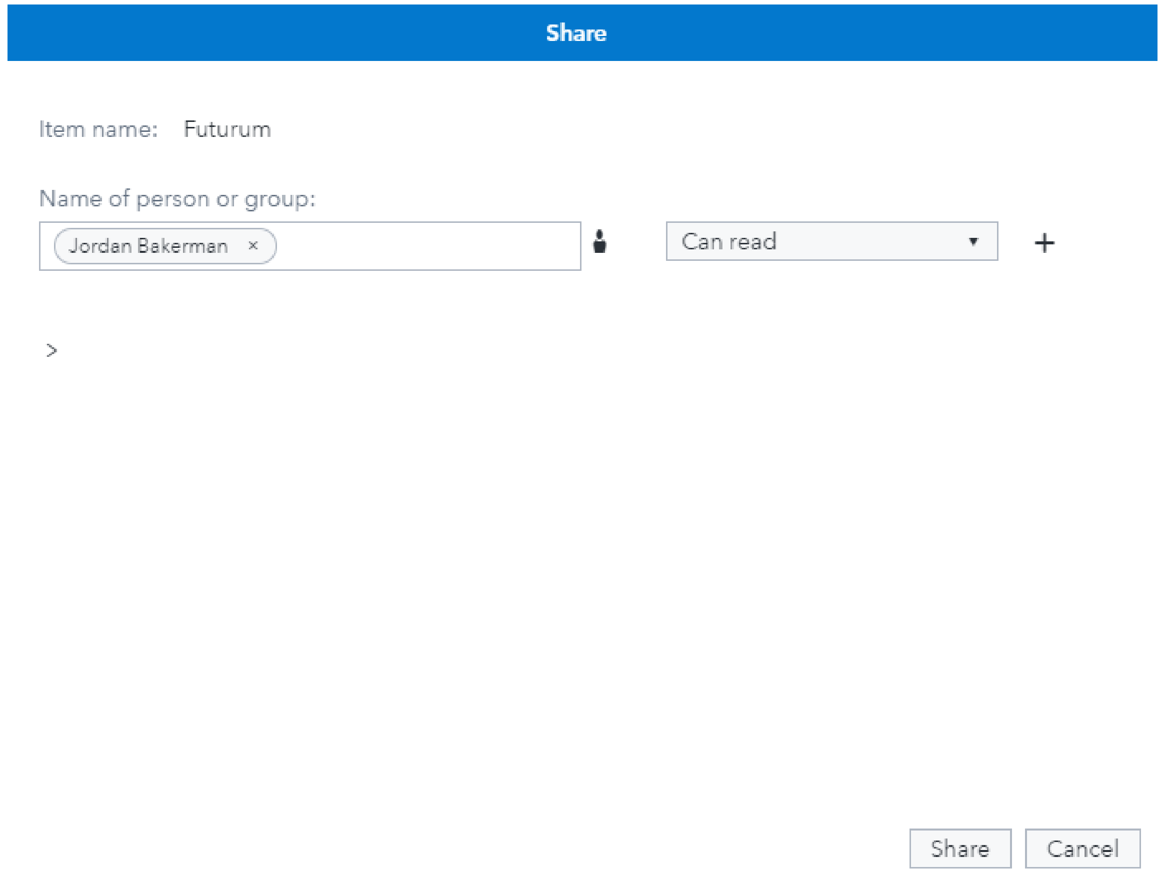
Sharing Projects – Read/Write

Share Projects

Your progress in the Model Studio (Develop Models) project is always saved automatically. To share the project, leave the “Develop Models” application and navigate to the “Share and Collaborate” application. Navigate to the “Recent” or “Build Models” folder. Right-click the name of the project we created (Futurum) in Model Studio and select “Share.” Select the persons or groups in your organization you would like to share the project with and select “Read access” or “Read and edit access,” and then click “Share.” In this way, you can easily collaborate with fellow Data Scientists on the project or share the project with business stakeholders or your management in “read” mode to view the key project information.







Thank you!