

IBM
Analytics



Data Science Workshop
SPSS Modeler Guide



© Copyright IBM Corporation 2018

IBM, the IBM logo, ibm.com, and IBM SPSS are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at “Copyright and trademark information” at www.ibm.com/legal/copytrade.shtml.

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

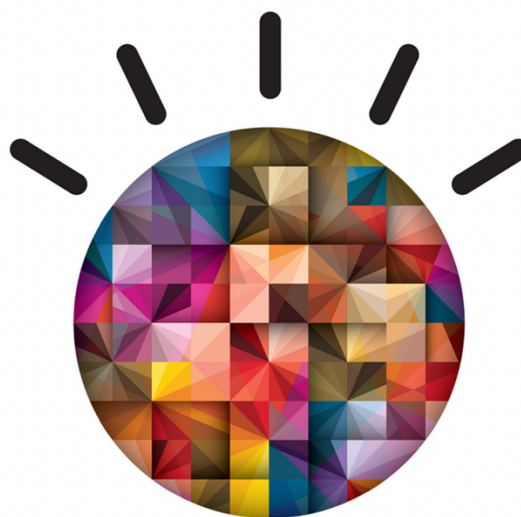
Contents

Exercise 1: Predict in 20 minutes	4
---	---

SPSS Modeler

IBM SPSS Modeler is a comprehensive data science platform, designed to bring **predictive intelligence** to everyday business problems, enabling front-line employees to make more **effective decisions** and **improve outcomes**.

This enables organizations to improve business processes and help people or systems **consistently make the right decisions** by delivering recommended actions at the point of impact. The result is a **rapid return on investment (ROI)** and the ability to **proactively** and **repeatedly reduce costs** and **increase productivity**.



This hands-on Data Science workshop is an instructor led session using IBM's data mining and predictive modeling software and is designed for those who are familiar with predictive analytics, as well as for beginners. Through this workshop you will experience firsthand how IBM SPSS Modeler works and how easy it is to implement predictive analytics.

Exercise 1: Predict in 20 minutes

Use Case

Goal: Approach:

Identify who has high probability to cancel the contract (Churn) to help marketing team to create a list of customer targets to receive specific marketing campaign

- Prepare data for modeling
- Define which fields to use
- Choose the modeling technique
- Automatically generate a model to identify who has cancel
- Review results

Why?

To save marketing cost and reduce the customer churn, identify those likely to churn and focus marketing efforts on those customers.

Predictive in 20 Minutes

1. Create a new model. Click on **+ New Modeler Flow**.
2. Type a name and description for our model and click **Create**.

IBM Watson Studio

Projects

Tools

Community

Services

Manage

Support

Docs

Modeler

New

From file

From example

Name*

Type name here.

50

Description

Type description here.

500

Select flow type

Modeler Flow

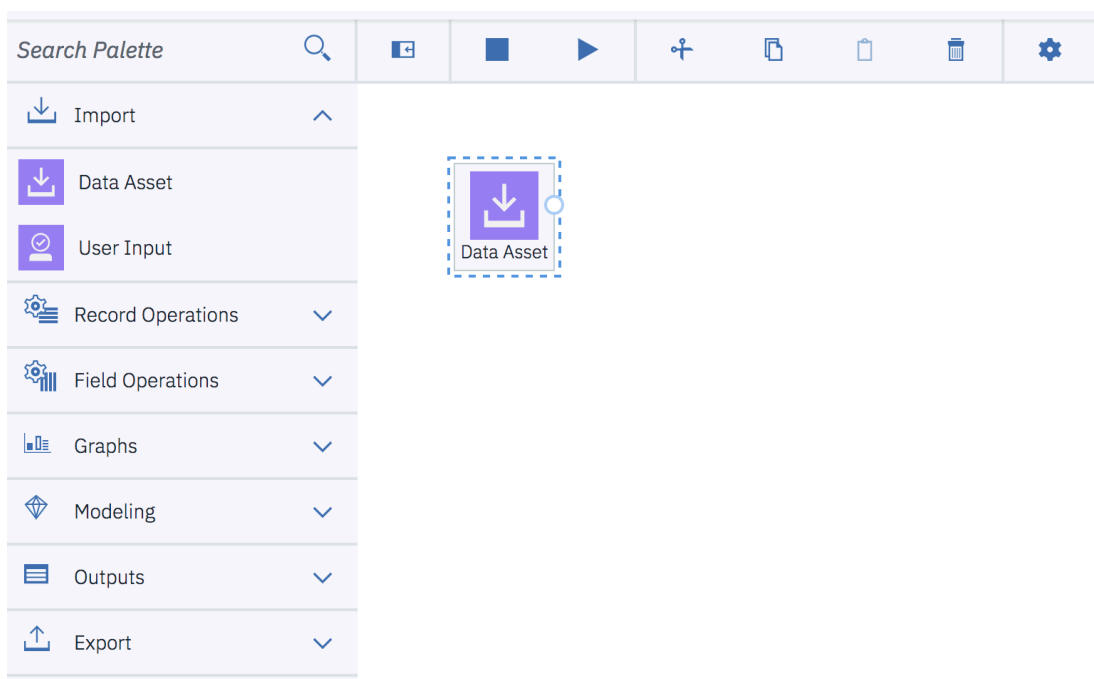
Neural Network Modeler BETA

Runtime

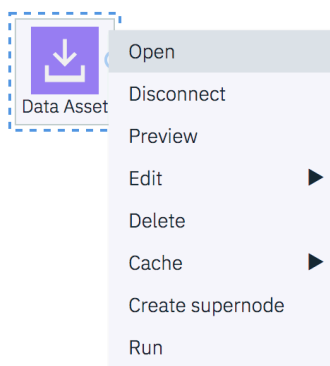
IBM SPSS Modeler

Scala Spark BETA

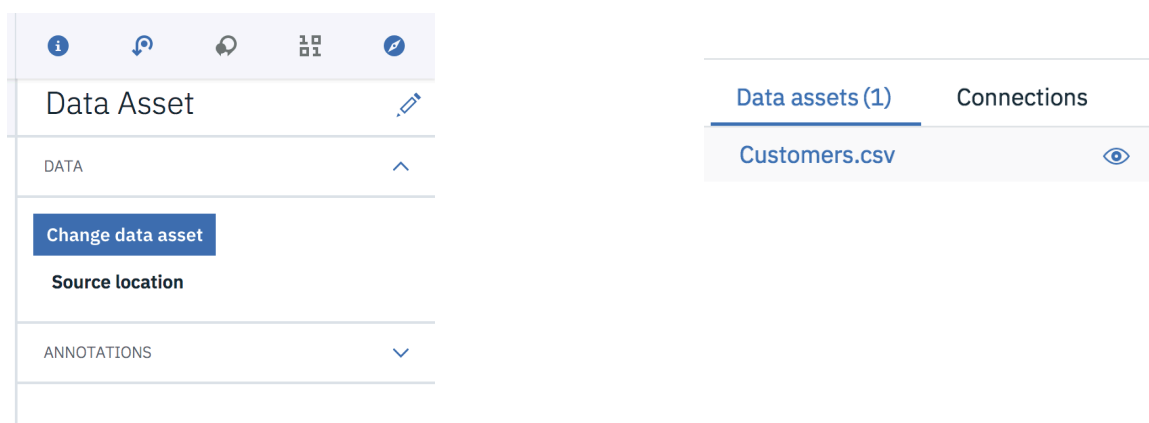
- From Import, add the **Data Asset** node to the workflow.



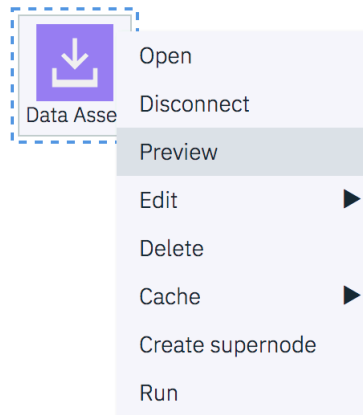
- Access the Data Asset properties. Right click on Data Asset and select **Open**



- Click on **Change data asset** to select the **customer_churn.csv** file already load in the project. Click **OK** and **SAVE**.



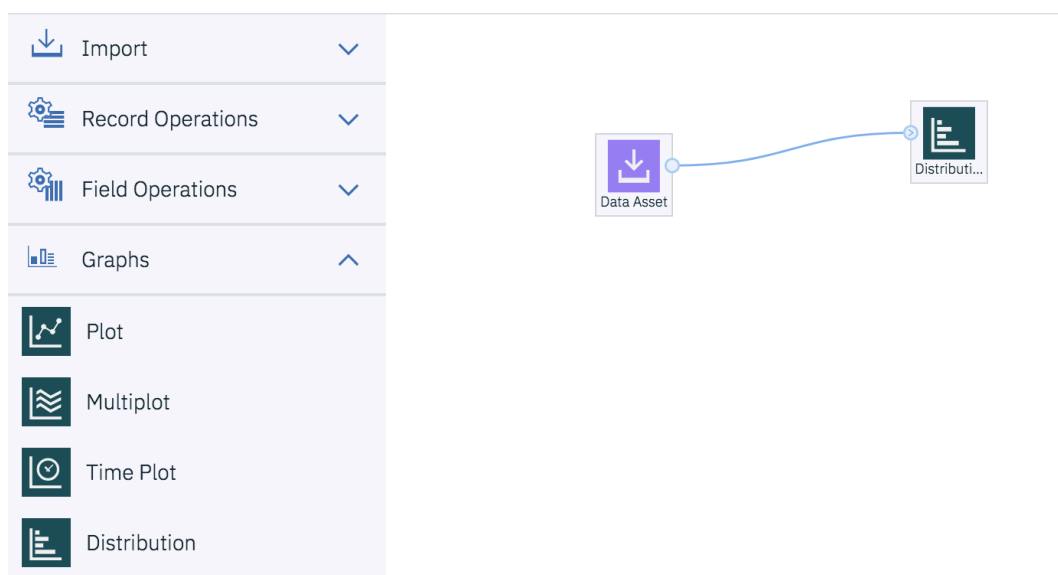
6. Right click on Data Asset and select **Preview**, to visualize the data available on file.



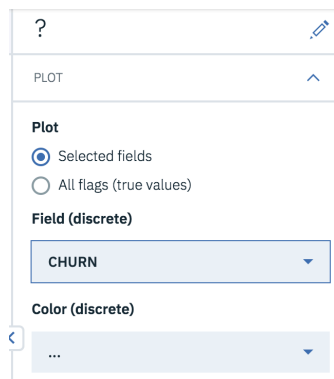
7. To go back to the workflow, click on the workflow name available on the top: My Projects / Project Name / **Workflow Name** / Preview.

	ID Decimal	Sex String	Status String	Children Decimal	Est_Income Decimal	Car_Owner String	Usage Decimal	Age Decimal
2	6	M	M	2	29616	N	75.29	49.426667
3	8	M	M	0	19732.8	N	47.25	50.673333
4	11	M	S	2	96.33	N	59.01	56.473333
5	14	F	M	2	52004.8	N	28.14	25.14
6	17	M	M	2	53010.8	N	58.87	18.84
7	18	M	M	1	75004.5	N	58.72	64.8

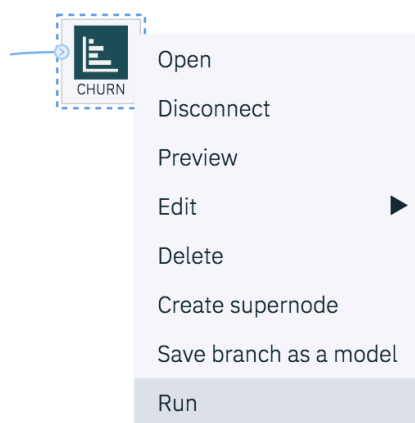
8. From Graphs options, add a Distribution node. Connect it to the data source. Click and hold the button on the first node, move the cursor to the second node and release when the cursor is on the mark of second node.



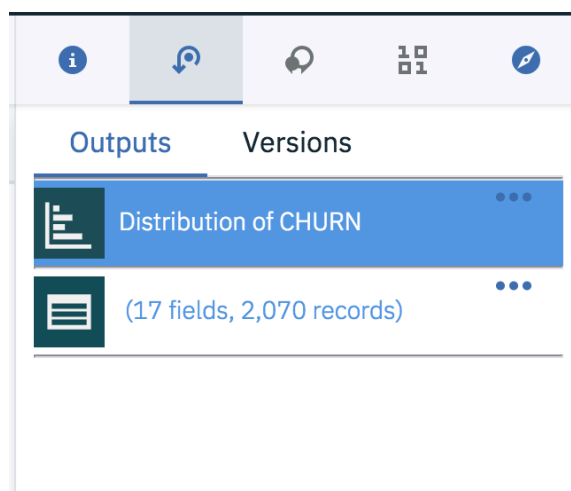
- Double-click on the Distribution node to access the properties. At **Field (discrete)**, select **CHURN** and click save.



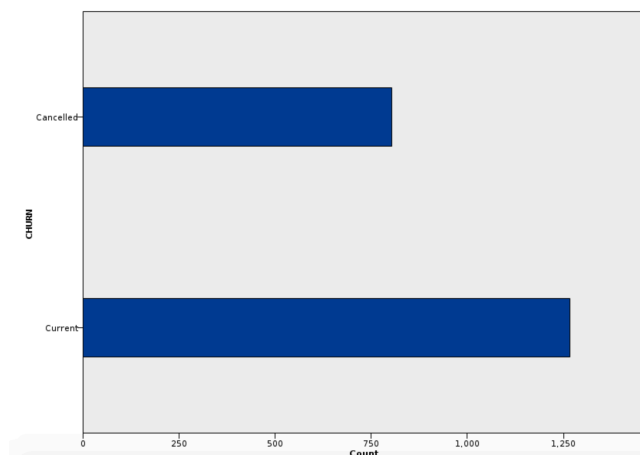
- Note that distributin node was rename to CHURN. Right click on it and select **Run**.



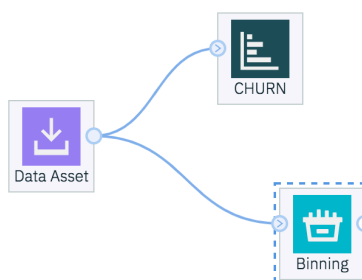
- Double-click on the graph Distribution of CHURN to open it.



12. The resulting graph shows that of the 2070 customers in this data set, around 40% of customers have cancelled their contract. The task is to build a model to understand the relationships within the data that led to the canceled their contracts.



13. From the Field Operations, add a Binning node to the workflow and connect it to the data source. The Binning node allows you to automatically generate bins (categories) using several techniques. In this case, we will be creating categories from the continuous variable Age.
14. Click and hold the button on the Data Asset node, move the cursor to the Binning node and release when the cursor is on the mark of second node.



15. Double-click on **Binning** to open properties.

Binning

SETTINGS ^


Bin fields











⊖ ⊕ Add Columns

16. Click on **Add Columns** and select **Ages**. Click OK.

Select Fields for Binning


Search in column Field name

Filter: 

<input type="checkbox"/> Field name ^	Data type ^
<input type="checkbox"/> ID	 integer
<input type="checkbox"/> Children	 integer
<input type="checkbox"/> RatePlan	 integer
<input type="checkbox"/> Dropped	 integer
<input type="checkbox"/> Est_Income	 double
<input type="checkbox"/> Usage	 double
<input checked="" type="checkbox"/> Age	 double
<input type="checkbox"/> LongDistance	 double
<input type="checkbox"/> International	 double
<input type="checkbox"/> Local	 double

17. Scroll down in the **Binning** properties, change the option **Bin width** to **5**. Click Save.

Bin width


5


☐ Use the same bins for all fields

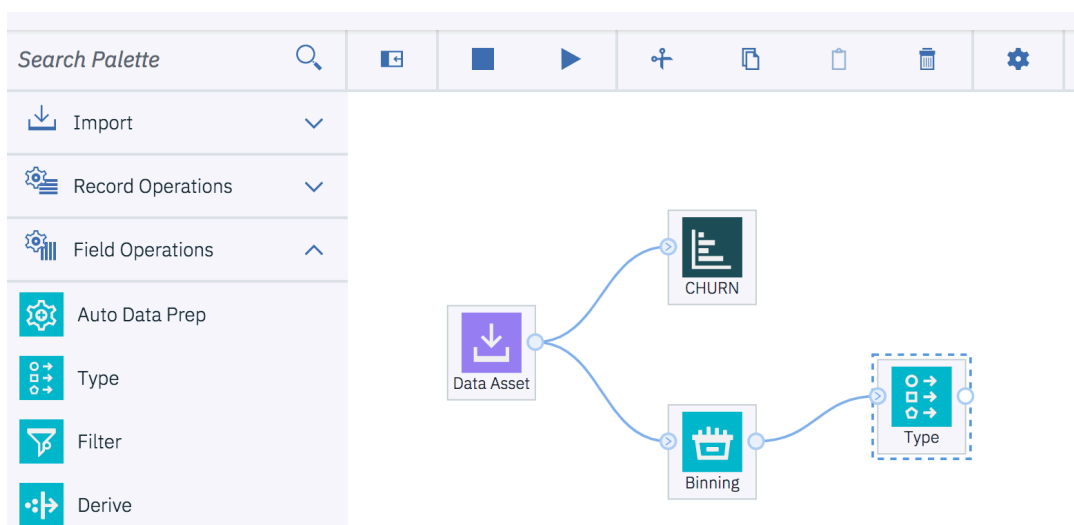
Recalculate bins

☒ Always
☐ If necessary

ANNOTATIONS



18. From the Field Operations, add a Type node to the workflow and connect it to the Binning node.



19. Open the Type node and click the Read Values button to scan the data as well as to display and update the range of values.

Using the drop-down box under Role, modify the following Fields:

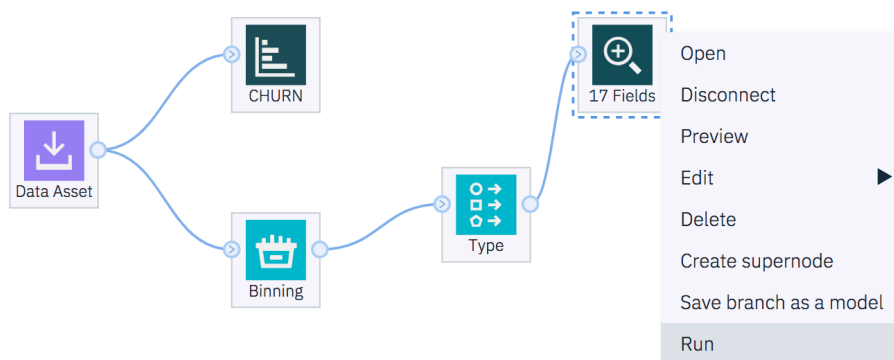
- ID = Record ID
- Age = None
- CHURN = Target

The Measurement of our Target should be set to Flag, which reflects two potential Values: Cancelled or Current. The remaining, including our new Age_Bin, will remain as Inputs in our analysis.

Save

Read Values		Clear All Values			
Field ^	Measure ^	Role ^	Value modeValues ^	Check	
Est_Inco...	Continuous	Input	Specify	96.33, 120000.0	None ⚙
Age_BIN	Nominal	Input	Specify	1, 2, 3, 4, 5, 6, 7, 8, ...	None ⚙
LongDist...	Continuous	Input	Specify	0.0, 59.0	None ⚙
Children	Continuous	Input	Specify	0, 2	None ⚙
ID	Continuous	Input	Specify	1, 3825	None ⚙
Age	Continuous	None	Specify	12.326667, 77.0	None ⚙
CHURN	Flag	Target	Specify	Cancelled, Current	None ⚙
Internati...	Continuous	Input	Specify	0.0, 9.7	None ⚙

20. From the Output, add the Data Audit node and connect it to the Type node. Right-click to select Run

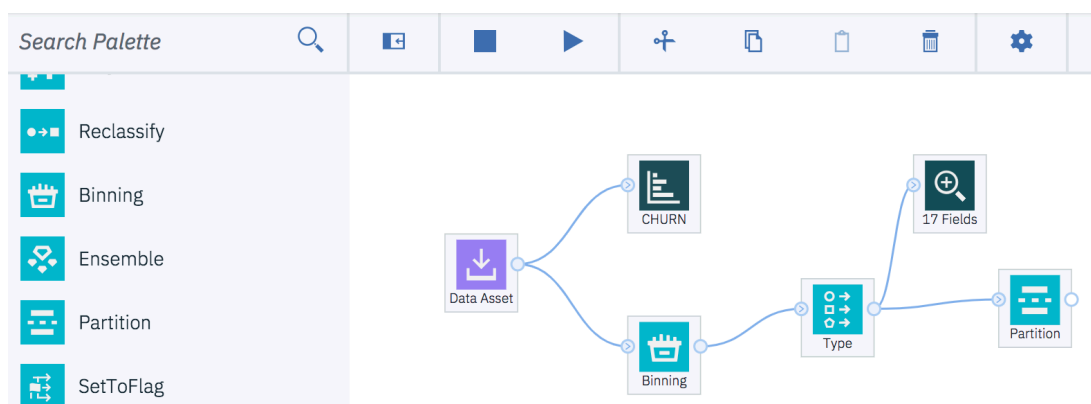


21. Double-click in the Data Audit (outputs) In the Data Audit results output, thumbnail graphs, storage icons, and summary statistics for all fields can be found. It also provides information about outliers, extremes and missing values.

Data Audit of [17 fields]

	Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev
1	ID		Continuous	1	3825	1901.152	1094.709
2	Sex		Flag	--	--	--	--
3	Status		Nominal	--	--	--	--
4	Children		Continuous	0	2	1.148	0.843

22. Now that we have explored our data, we can build a model to uncover the key drivers resulting in cancelled contract. Go back to the workflow and from the Field Operations, add a Partition node to the workflow and connect it to the Type node.



23. Double-click on the Partition node and change Training Partition to 70 and Testing Partition to 30. Click Save.

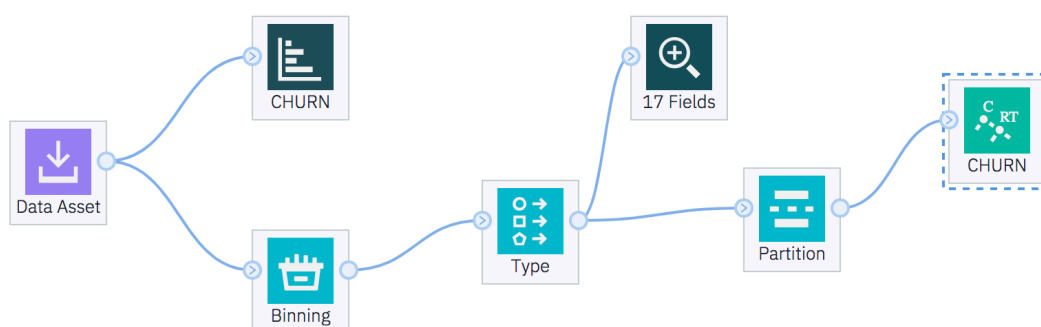
Training Partition

70

Testing Partition

30

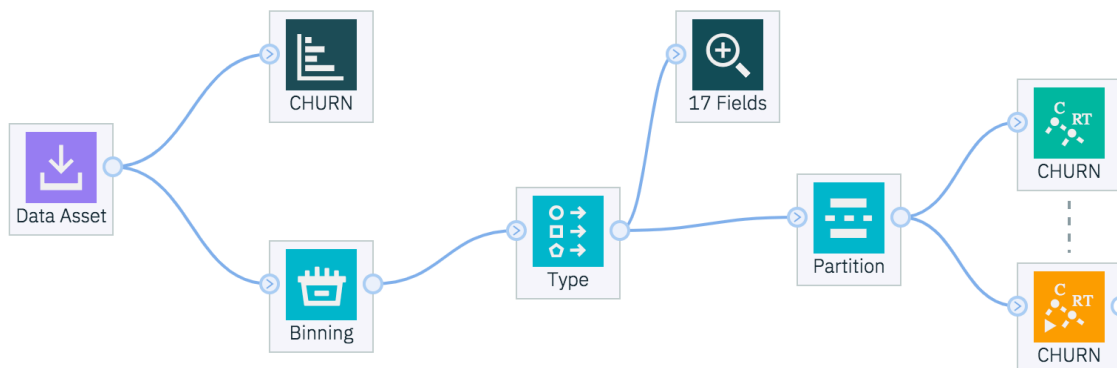
24. From the Modeling, add the C&R Tree node to the workflow and connect it to the Partition node. Note that the C&R Tree node name changes to CHURN when it is connected to the Partition node. This is because we defined CHURN as the Target Role in Step 20.



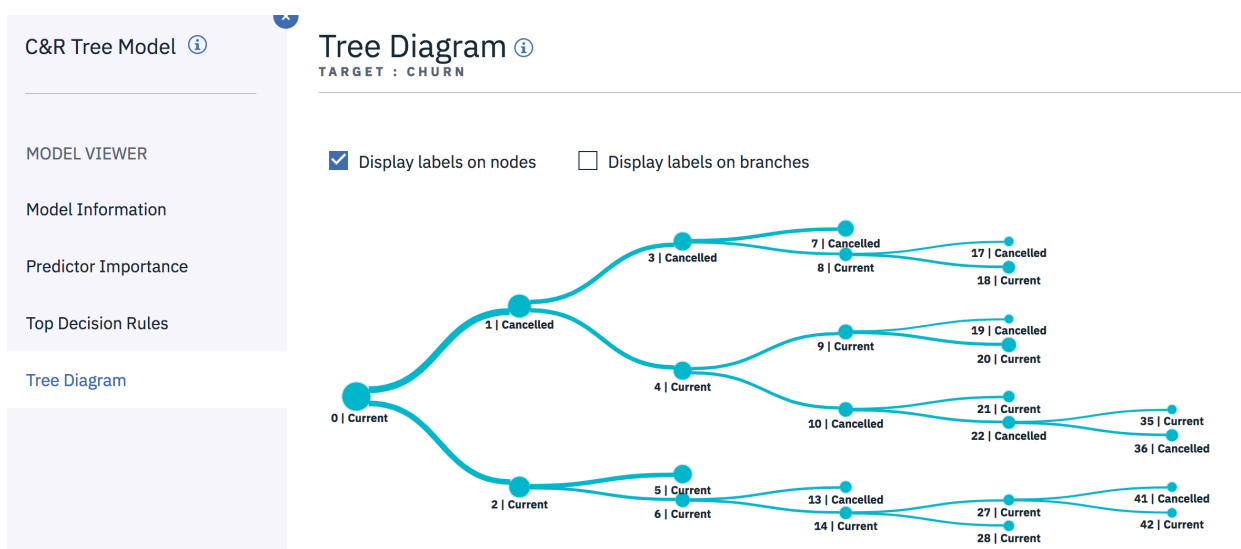
25. Open the C&R Tree node to view the settings before running the model. This example uses the defaults so closed properties and click Run.

CHURN	
FIELDS	✓
OBJECTIVE	✓
BASICS	✓
STOPPING RULES	✓
COSTS	✓
PRIORS	✓

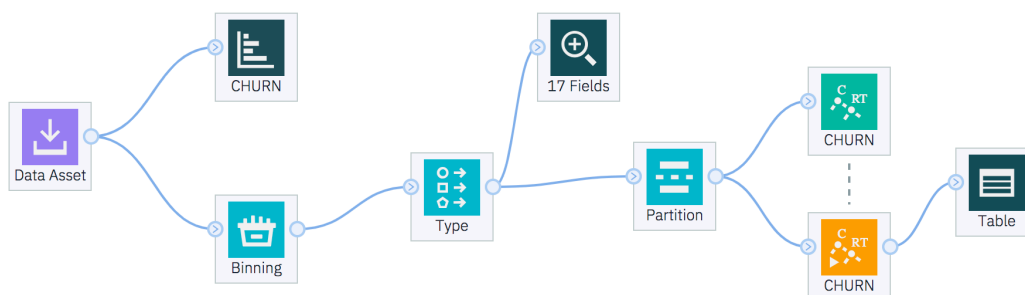
26. The C&R Tree model is generated, added to the workflow and named CHURN.



27. Right-click on the generated model and select View Model to see the outputs. The Model outputs contains Model Information, Predictor Importance, Top Decision Rules and the Tree Diagram. Click the Viewer each one.



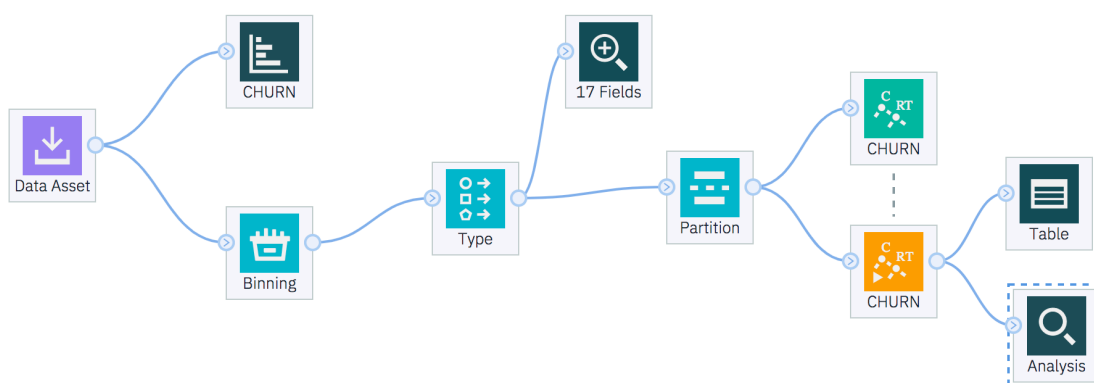
28. To view model output, from the Outputs add a Table node to the generated CHURN model and run it.



29. Look at the last five columns in the table. The fifth from last column is the actual outcome, whether the customer cancelled or is current; the second to last column is the prediction from the model; and the last column is the confidence in the prediction. Note that the model predicted that the customer would cancelled with 78.1% confidence.

LocalBilltype	LongDistanceBilltype	CHURN	Age_BIN	Partition	\$R-CHURN	\$RC-CHURN
Budget	Intl_discount	Cancelled	3.000	1_Training	Cancelled	0.781
FreeLocal	Standard	Current	8.000	1_Training	Current	0.896
FreeLocal	Standard	Current	9.000	1_Training	Cancelled	0.781
Budget	Standard	Current	10.000	2_Testing	Current	0.810

30. To see the overall accuracy of the model, from the Output, add an Analysis node, join it to the generated model and run it.

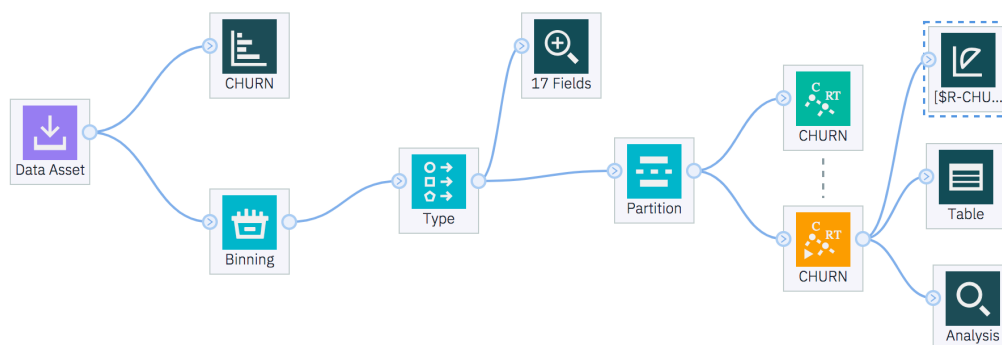


31. The resulting output indicates an overall accuracy of 77.93%. That is, the model predicted with 77.93% accuracy which customers Cancelled or Current the contract.

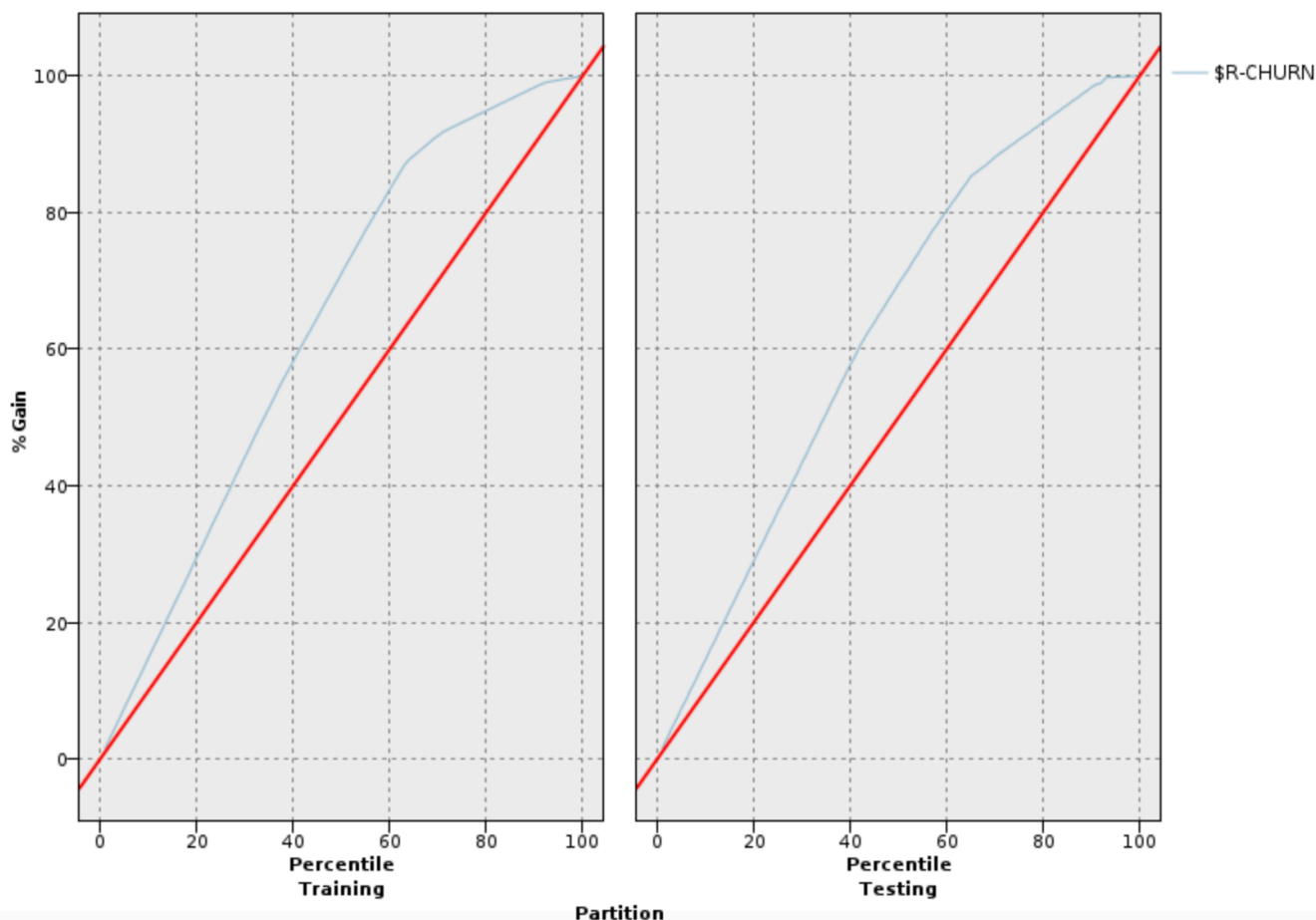
Results for output field CHURN
Comparing \$R-CHURN with CHURN

'Partition'	1_Training		2_Testing	
Correct	1,173	82.49%	505	77.93%
Wrong	249	17.51%	143	22.07%
Total	1,422		648	

32. To further evaluate the model, select an Evaluation node from the Graphs, connect it to the model node and select Run.



33. In the resulting gains chart, the red line reflects what you could expect without Predictive Analytics. The blue line; however, reflects the lift in response you could achieve utilizing Predictive Analytics. Therefore, if you were to randomly select 50% of your client base, you could expect to have captured 50% of those likely cancelled. By using Predictive Analytics, you can more effectively target those 50% of clients and capture almost 80% of those likely to cancel.



34. Finally, to see the relationships between fields, select a Matrix node from the Output and connect it to the model node. Using the drop-down menu, choose “\$R-CHURN” for Rows and “CHURN” for Columns. Select Run.

Matrix of \$R-CHURN by CHURN

\$R-CHURN	Cancelled	Current
Cancelled	580	168
Current	224	1098

Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 738.419, df = 1, probability = 0

Summary

- Prepare data for modeling
- Define which fields to use
- Choose the modeling technique
- Automatically generate a model to identify who has cancelled
- Review results

Over the course of the last 20 minutes, we were able to successfully train a model by exploring SPSS Modeler’s ability to read in data, create new fields via data preparation techniques, choose and run a predictive modeling algorithm, and evaluate the results to accurately identify customer response.