

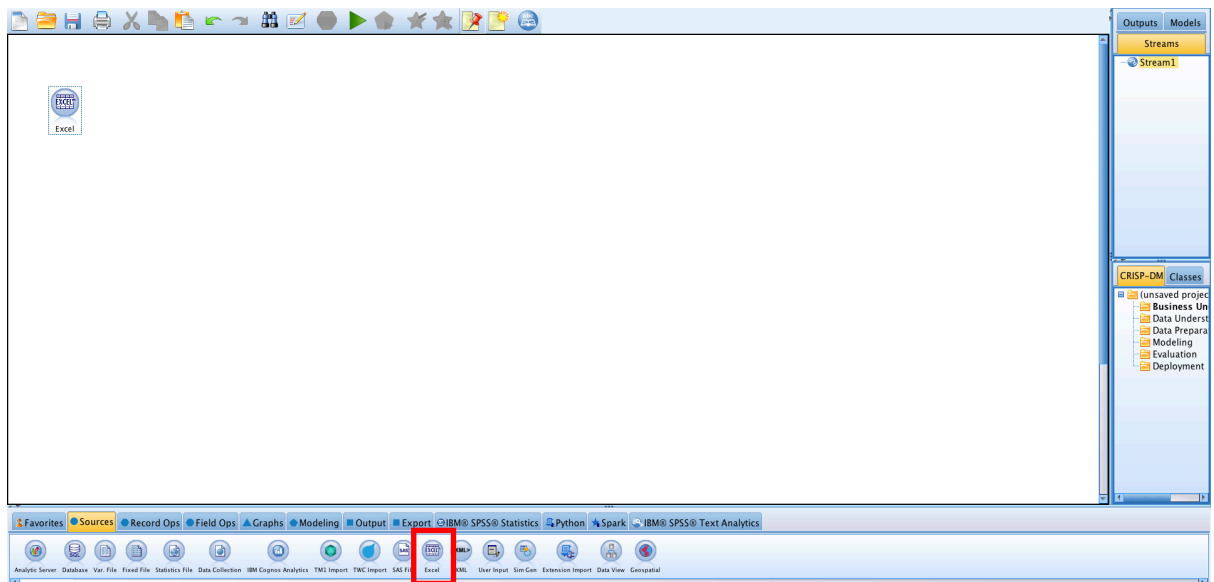
# IBM SPSS Modeler Tutorial Workbook

This tutorial let's you build a Machine Learning (ML) model without coding using IBM SPSS Modeler. In this case you'll build a model that will help to predict the customer churn (customers leaving the company) for a telecommunications company (telco). This tutorial consists of the following steps

- Connect, merge and augment the data sources
- Exploratory Data Analyses (EDA)
- Build the ML model
- Evaluate and visualize ML model performance
- Make predictions on a new data set

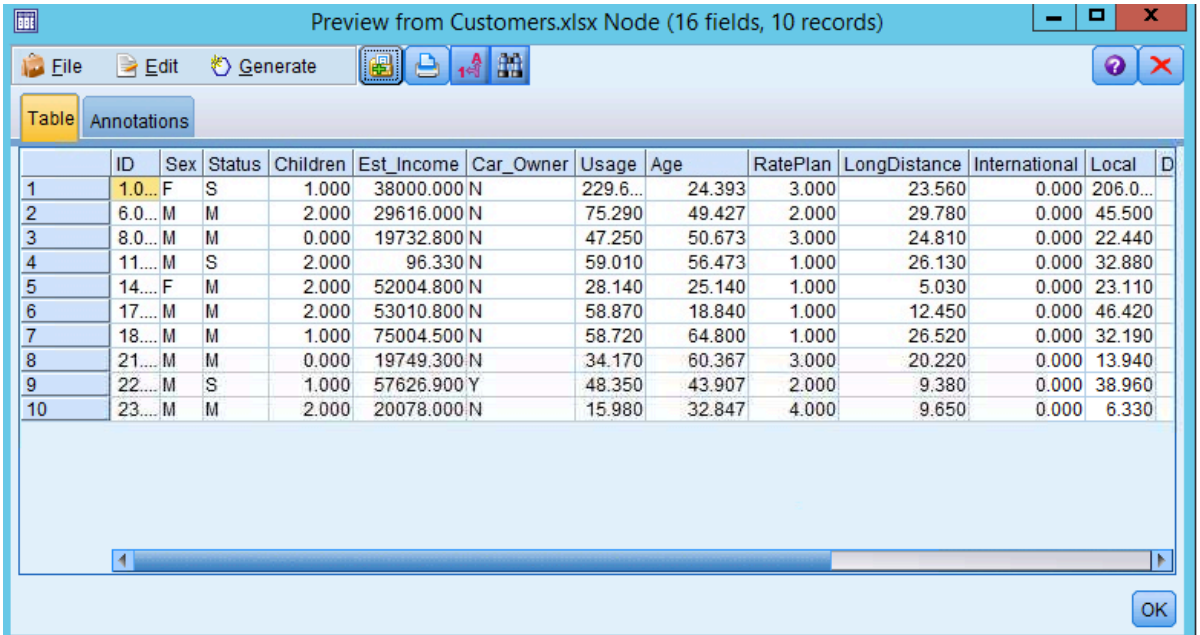
## Connect, merge and augment the data sources

1. Start IBM SPSS Modeler
2. From the **Sources** palette in the bottom of the screen, double-click on the **Excel node** to add it to the canvas.



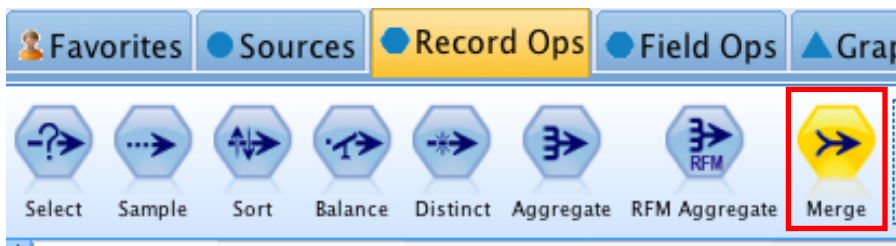
3. Double-click the **Excel node** on the canvas to open a dialog box. Use the data tab to import the **Customers.xlsx** file from the location to which you have downloaded the data files (these can be downloaded [here](#))
4. Click on the **Preview** button at the top of the dialog box to see the first 10 records in the file, an extraction of data from a telco company's CRM system. It includes historical data related to their customers' demographics, purchasing behavior and segments.

5.



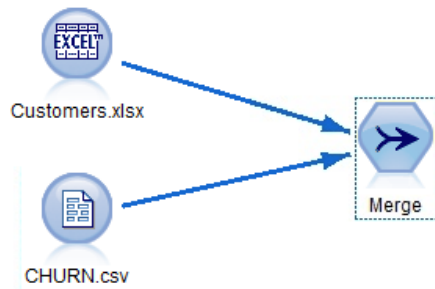
	ID	Sex	Status	Children	Est_Income	Car_Owner	Usage	Age	RatePlan	LongDistance	International	Local	D
1	1.0...	F	S	1.000	38000.000	N	229.6...	24.393	3.000	23.560	0.000	206.0...	
2	6.0...	M	M	2.000	29616.000	N	75.290	49.427	2.000	29.780	0.000	45.500	
3	8.0...	M	M	0.000	19732.800	N	47.250	50.673	3.000	24.810	0.000	22.440	
4	11...	M	S	2.000	96.330	N	59.010	56.473	1.000	26.130	0.000	32.880	
5	14...	F	M	2.000	52004.800	N	28.140	25.140	1.000	5.030	0.000	23.110	
6	17...	M	M	2.000	53010.800	N	58.870	18.840	1.000	12.450	0.000	46.420	
7	18...	M	M	1.000	75004.500	N	58.720	64.800	1.000	26.520	0.000	32.190	
8	21...	M	M	0.000	19749.300	N	34.170	60.367	3.000	20.220	0.000	13.940	
9	22...	M	S	1.000	57626.900	Y	48.350	43.907	2.000	9.380	0.000	38.960	
10	23...	M	M	2.000	20078.000	N	15.980	32.847	4.000	9.650	0.000	6.330	

6. After reviewing the **Preview**, or any subsequent output, click on the red X to close.
7. From the **Sources** palette in the bottom of the screen, double-click on the **Var.file** node to add it to the canvas. Double-click the **Excel** node on the canvas to open a dialog box. Use the data tab to import the **CHURN.csv** file from the location to which you have downloaded the data files (these can be downloaded [here](#))
8. Click on the **Preview** button at the top of the dialog box to see the first 10 records in the file. This file has been prepared by our data engineering department and contains information about customers who have churned and customers who are still active. After reviewing the Preview, or any subsequent output, click on the red X to close.
9. From **Record Ops** tab in the palette drag a **Merge** node to the canvas

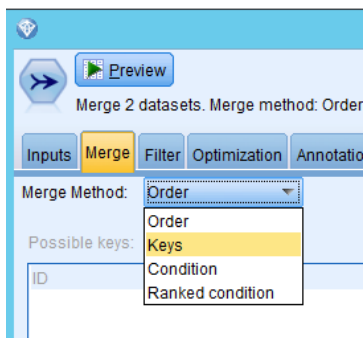


10. Select the **Customer.xlsx** node on the canvas and select the **F2** key on your

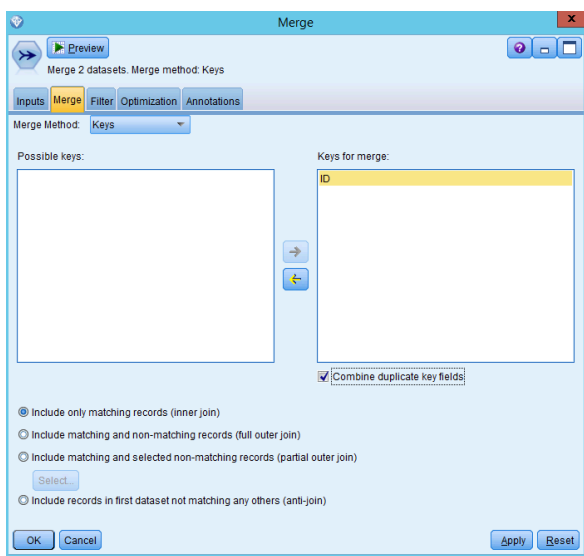
keyboard to connect it to the **Merge** node. Do the same with the **CHURN.csv** node. Once connected you'll see that both source nodes are connected with a blue arrow to the **Merge** node.



11. Double click the **Merge** node on the canvas to set the merge properties. Select **Keys** as the **Merge Method**.

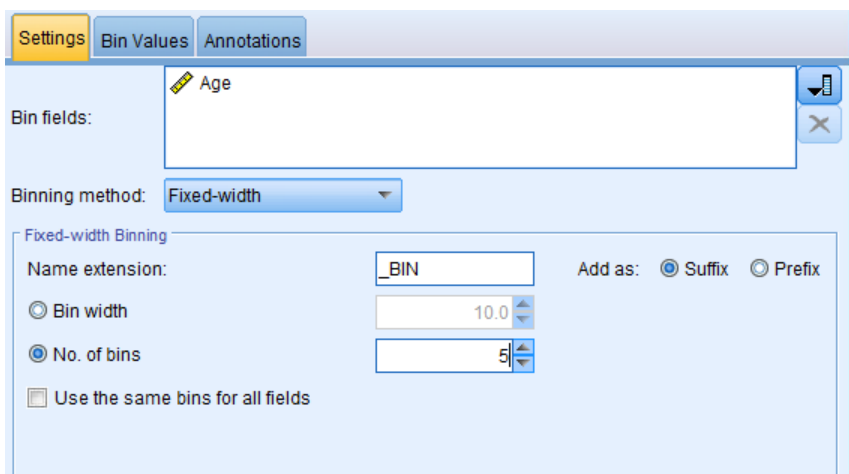
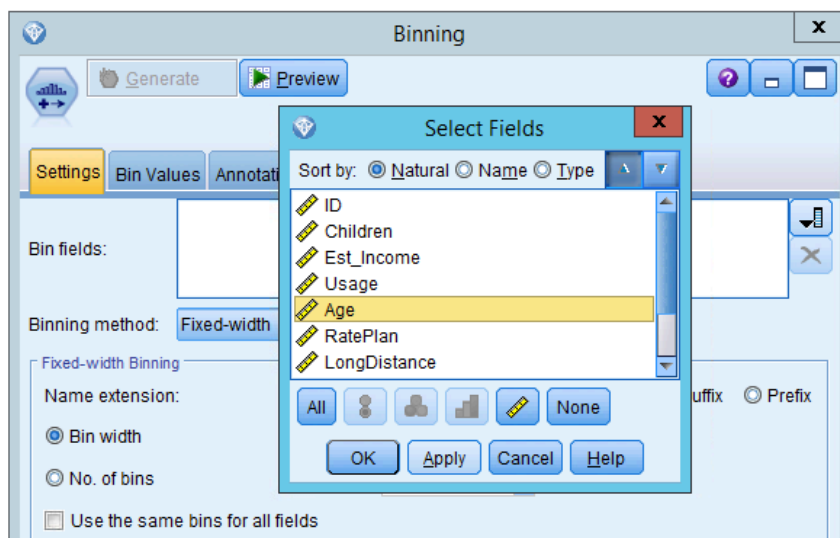


12. Select the **ID** field in the **Possible Fields** and use the **Yellow arrow** in the middle to move this to the **Keys to merge** section. Click **OK** to close the window



13. From the **Field Ops** palette, add a **Binning** node to the canvas 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**.

- Double-click on the Binning node on the canvas to edit the settings.
- Using the **Select field icon**, select **Age** as the **Bin field**. Leaving the **Binning method** at fixed-width, select 5 as the **No of bins** to create, then click **OK**.
- The **Bin Values** tab allows you to see the lower and upper cut points. By selecting **Preview** you can see the appended field **Age\_Bin**, which shows 5 possible categories.

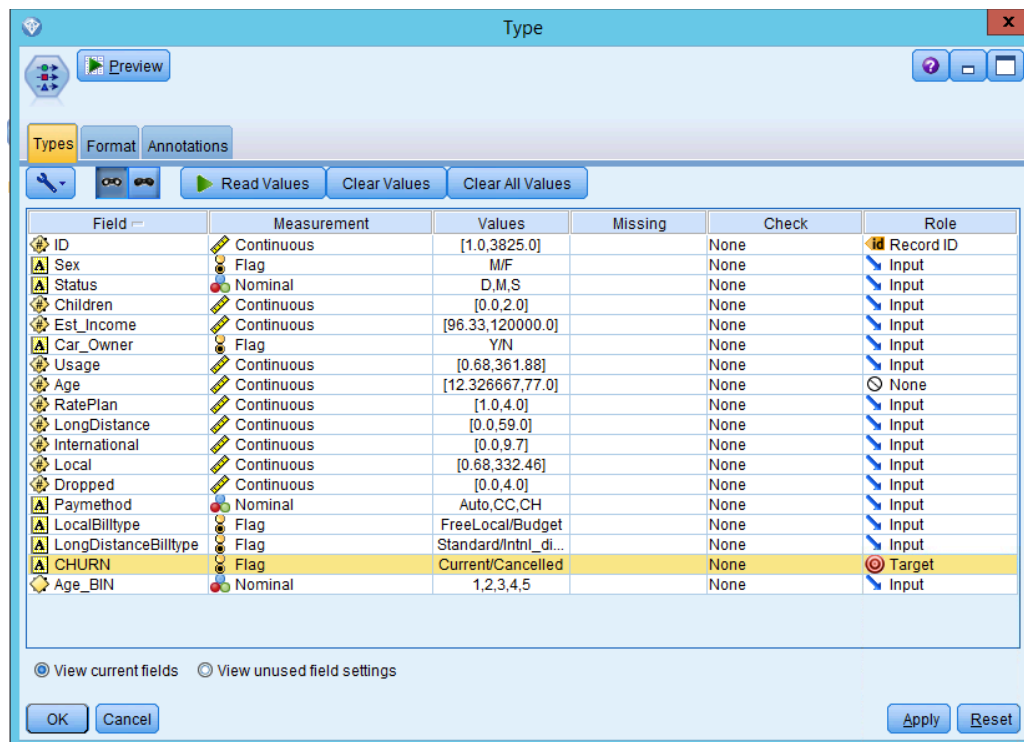


## Exploratory Data Analyses (EDA)

14. From the **Field Ops** palette, add a **Type** node to the canvas and connect it to the **Binning** node. Double-click on the **Type** node and click the **Read Values** button to scan the data as well as to display and update the range of values (we call this *instantiating* the data).

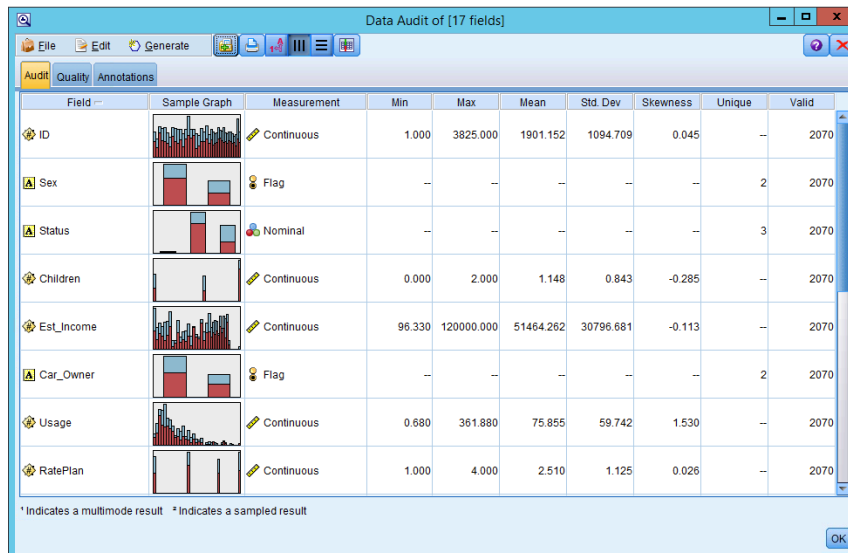
Using the drop-down box under **Role**, modify the following Fields:

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



15. From the **Output** palette, double-click on the **Data Audit** node and connect it to the **Type** node. Right-click the **Data Audit** node on the canvas and select **Run** to inspect the data set. In the **Audit** tab of the resulting output, thumbnail graphs, storage icons, and summary statistics for all fields can be found. Double-clicking on any of the graphs will provide a more detailed outlook of the Field. In the **Quality** tab (not shown), information about outliers, extremes and missing values are shown.

After you are finished reviewing the data you can click **OK** to close the **Data Audit** window



16. After exploration, we will split the data set into a test and a training partition. From the **Field Ops** tab in the palette drag a **Partition** node to the canvas and connect it to the **Type** node.

Double click the **Partition** node on the canvas and set the partition sizes: We use the following split: 70% for training and 30% for testing.

The 'Partition' dialog box contains the following settings:

- Partition field:** Partition
- Partitions:** ☒ Train and test ☐ Train, test and validation
- Training partition size:** 70  Label: Training Value = "1\_Training"
- Testing partition size:** 30  Label: Testing Value = "2\_Testing"
- Validation partition size:** 0  Label: Validation Value = "3\_Validation"
- Total size:** 100%
- Values:** ☐ Use system-defined values ("1", "2" and "3") ☒ Append labels to system-defined values ☐ Use labels as values
- ☒ Repeatable partition assignment
- Seed:** 1234567
- ☐ Use unique field to assign partitions:

Buttons: OK, Cancel, Apply, Reset

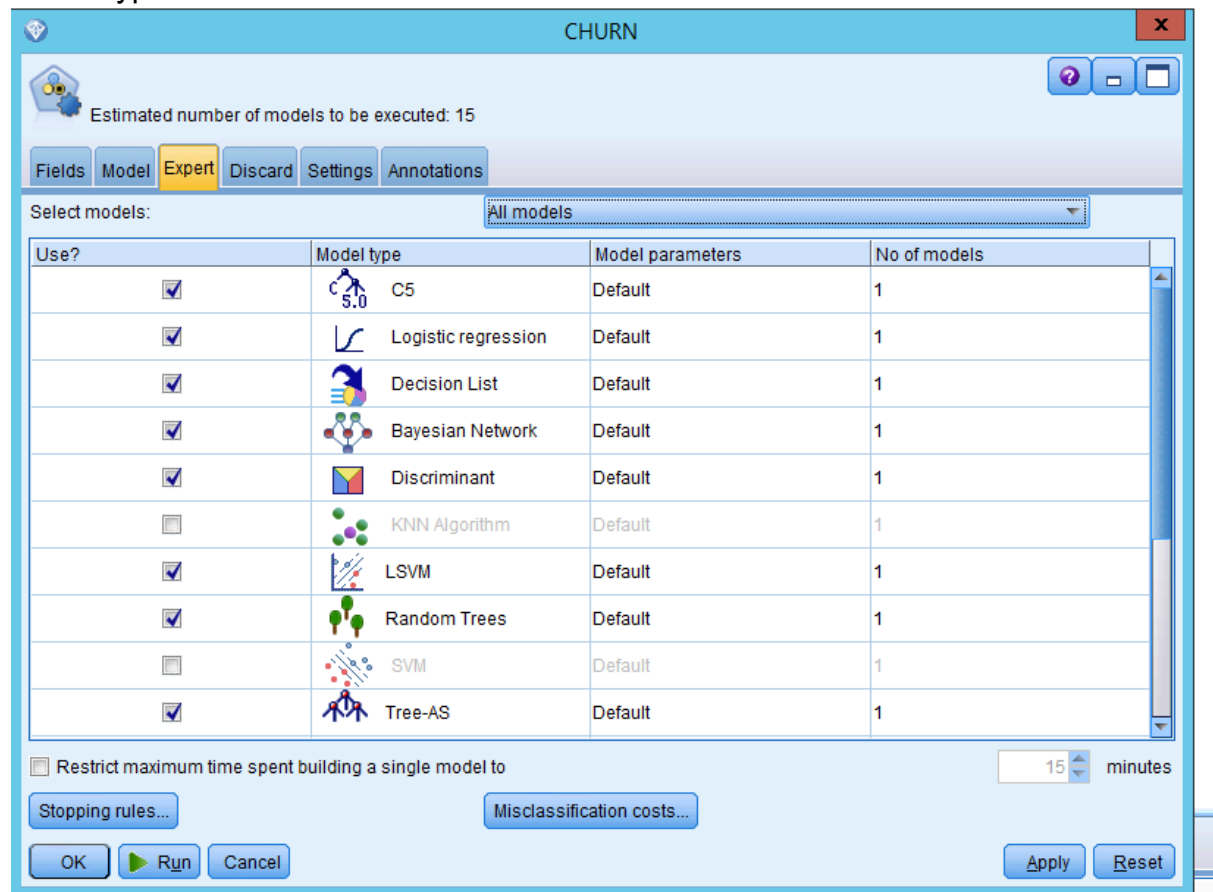
## Build the ML model

17. Now that we have explored our data, we can train a model to uncover the key drivers resulting in customer churn. As we don't exactly know what the best

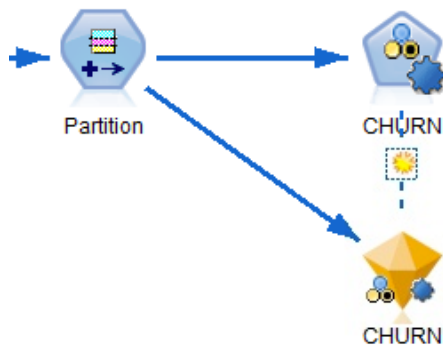
model would be to use we'll get some help from SPSS Modeler.

From the **Modeling** tab drag a **Auto-Classifer** node to the canvas and connect this to the **Partition** node. Once connected the Auto-Classifer node on the canvas will be change to **CHURN** as this is the target we want to predict (we defined this in the Type node).

Double click the **CHURN** node on the canvas to review the properties. On the **Model** tab you can set general properties how to build and rank the various models. We'll leave this as is. Click the **Expert** tab and review all the various model types SPSS Modeler will use. Also leave this as is.



- Click the green **Run** button on the bottom of the window to execute this node. IBM SPSS Modeler will start evaluating all the model types and once finished a "golden nugget" node has been added to the canvas.



19. Double click the generated **CHURN** golden nugget node on the canvas to review the top 3 best performing models. You can review some key evaluation metrics like model lift, model accuracy etc. Click **OK** to close the window

Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift(Top 3...	Overall Accuracy (%)	No. Fields	Area Under
<input checked="" type="checkbox"/>		Random Trees 1	< 1	1,593.846	63	1.554	88.117	15	0.916
<input checked="" type="checkbox"/>		Bayesian Network 1	< 1	1,470.0	62	1.520	84.105	15	0.874
<input checked="" type="checkbox"/>		Neural Net 1	< 1	1,475.0	60	1.513	84.259	15	0.889

## Evaluate and visualize ML model performance

We will evaluate the model performance by using various model evaluation techniques in IBM SPSS Modeler.

20. From the **Output** tab in the palette drag a **Table** node and to the canvas. And connect this to the **CHURN** golden nugget node. Right click the **Table** node and select **Run** from the menu.

Look at the last two columns of the table. The second to last column contains the **predicted response outcomes**, which can be compared to the historical



outcomes in the 5<sup>th</sup> to last column; and the last column contains the **confidence** of that prediction. For example, the first record shows a customer who did, in fact, churn. The appended columns show that the model predicted that the customer would churn with 96.8% confidence.

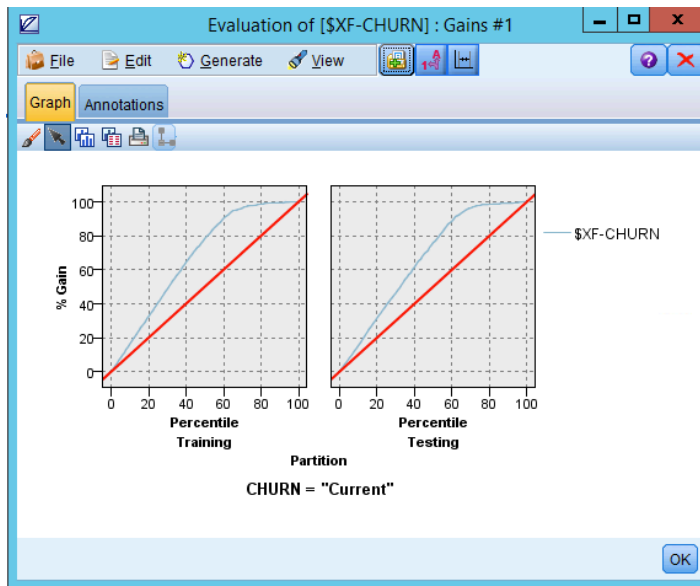
Annotations					
Billtype	CHURN	Age_BIN	Partition	\$XF-CHURN	\$XFC-CHURN
	Cancelled	1	1_Training	Cancelled	0.968
	Current	3	1_Training	Current	0.931
	Current	3	1_Training	Current	0.758
	Current	4	2_Testing	Current	0.857
	Cancelled	1	1_Training	Current	0.716
	Current	1	1_Training	Current	0.994
	Current	5	1_Training	Current	0.930
	Current	4	1_Training	Current	0.645
	Current	3	1_Training	Current	0.768
	Current	2	1_Training	Current	0.959
	Current	2	1_Training	Current	0.992
	Cancelled	1	1_Training	Cancelled	0.962
	Current	3	2_Testing	Current	0.577
	Cancelled	4	1_Training	Cancelled	0.884
	Cancelled	4	2_Testing	Cancelled	0.468
	Cancelled	3	2_Testing	Current	0.806
	Cancelled	1	1_Training	Cancelled	0.968
	Current	3	1_Training	Current	0.965
	Cancelled	4	1_Training	Cancelled	0.638
	Cancelled	3	1_Training	Cancelled	0.906

21. To see the **overall accuracy** of the model, select an **Analysis** node from the **Output** palette, connect it to the **CHURN** golden nugget node. Right click the **Analysis** node and select **Run** from the menu. You can review the accuracy for both the training and the test partition.

Analysis of [CHURN] #4					
File Edit View Help					
Analysis Annotations					
Collapse All Expand All					
Results for output field CHURN					
Comparing \$XF-CHURN with CHURN					
Partition	1_Training		2_Testing		
Correct	1,275	89.66%	567	87.5%	
Wrong	147	10.34%	81	12.5%	
Total	1,422		648		

22. To further evaluate the model, select an **Evaluation** node from the **Graphs** palette, connect it to the **CHURN** golden nugget node. Right click the **Evaluation** node and select **Run** from the menu.

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 this ML model.

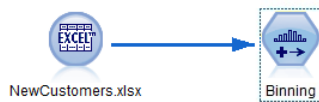


### Make predictions on a new data set

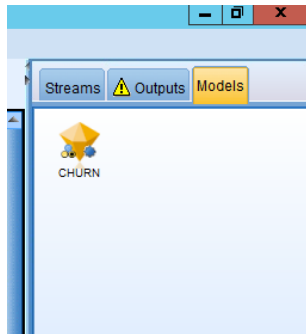
1. We have trained a model using IBM SPSS Modeler. In this last task of this tutorial we are going to apply this to make predictions on a data set with new customers.

From the **Sources** palette drag an **Excel** node to the canvas. Drag it below the model you just created. Double click the Excel node and select the **NewCustomers.xlsx** file from the location to which you have downloaded the data files (these can be downloaded [here](#)) Click **Preview** to preview the data set and **OK** to close after the preview.

2. Since we have binned the **Age** field we need to apply this to our new data set as well. The easiest way is to simply select the **Binning** node on the canvas, copy this (using **Ctrl-C** or via the **Edit** menu) and paste it next to the **NewCustomers.xlsx** node. Connect the **new Binning** node to the **NewCustomers.xlsx** node.



3. Drag from the **Models** tab in the top right corner the **CHURN** golden nugget node and connect this to the **new Binning** node.



4. From the **Output** tab on the palette drag a **Table** node to the canvas and connect this to the **CHURN** golden nugget node.



5. Right click the **Table** node and select **Run** from the menu. Review the last 2 columns of the table with the predicted value and confidence level.

	Billtype	LongDistanceBilltype	Age_BIN	\$XF-CHURN	\$XFC-CHURN
1	Local	Standard	4	Current	0.546
2	Local	Standard	1	Current	0.994
3	Local	Standard	2	Current	0.992
4	jet	Standard	1	Cancelled	0.962
5	jet	Standard	5	Cancelled	0.654
6	jet	Standard	4	Current	0.926
7	Local	Standard	3	Cancelled	0.936
8	jet	Standard	2	Current	0.997
9	Local	Standard	3	Current	0.845
10	Local	Standard	4	Cancelled	0.746
11	Local	Intl_discount	5	Current	0.953
12	Local	Standard	4	Cancelled	0.931
13	Local	Standard	5	Current	0.656
14	Local	Standard	4	Current	0.407
15	jet	Standard	1	Cancelled	0.901
16	Local	Standard	4	Cancelled	0.546
17	jet	Standard	1	Cancelled	0.943
18	jet	Standard	4	Current	0.986
19	Local	Standard	4	Cancelled	0.952
20	jet	Intl_discount	3	Current	0.477

This concludes the IBM SPSS Modeler tutorial. You can find the completed stream (*Tutorial SPSS Modeler - end result.str*) in the [Solutions](#) folder on Github. Normally the next step would be to filter the results on the customer that have the highest propensity to churn. And share these with the marketing and sales departments so they can take action on this. However, this is beyond the scope of this workshop. If you are interested how this might look like please review the *Tutorial – IBM SPSS Modeler Deployment.str* stream in the [Solution](#) folder.

