IBM Data Science - Modeler Flow - Labs instructions

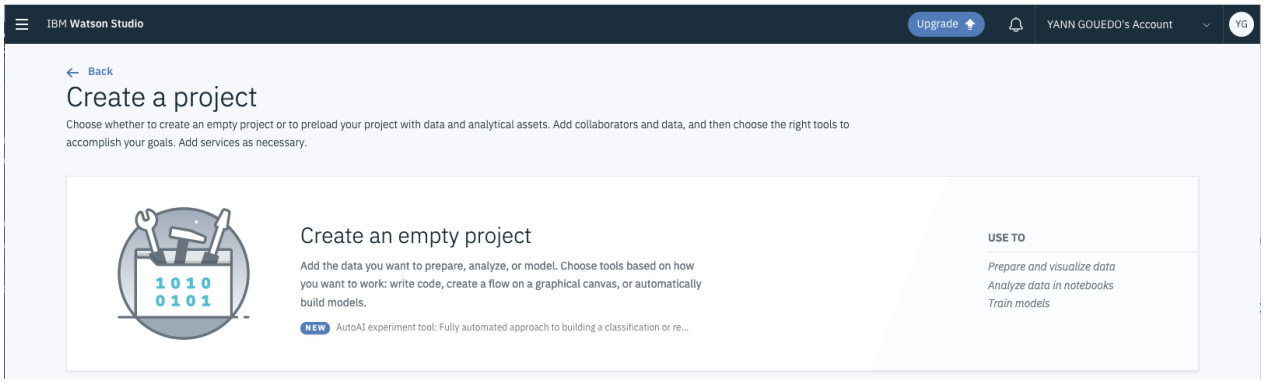
Watson Studio - Modeler Flows - Tutorials

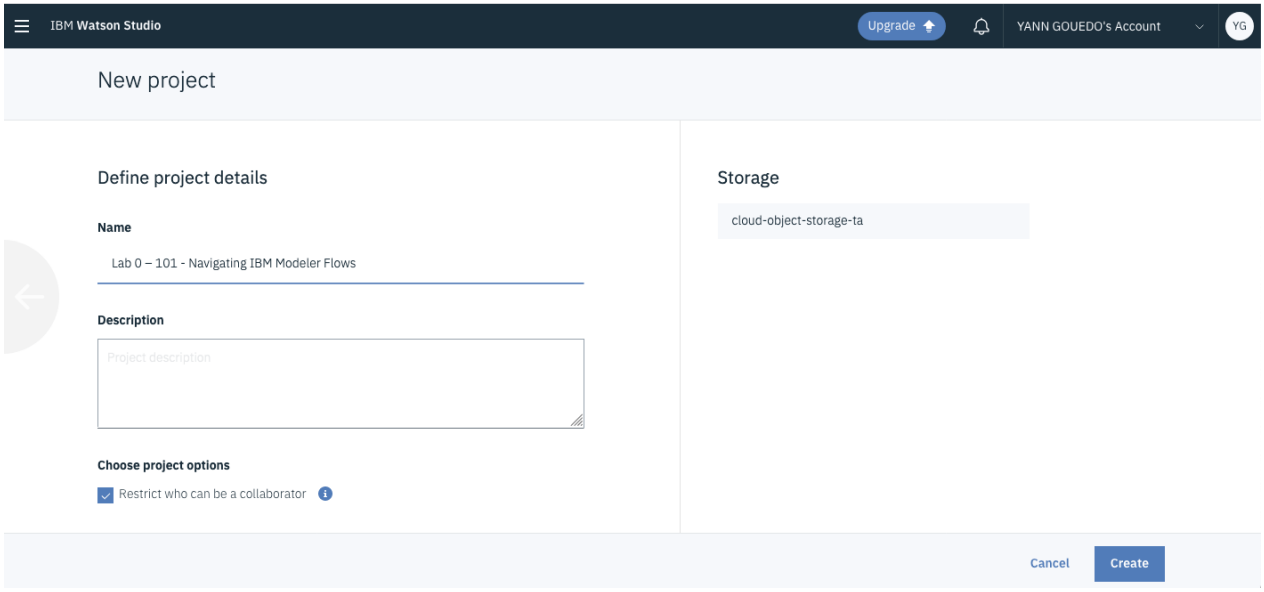
- Watson Studio Modeler Flows: SPSS Runtime (3 minutes 45 secondes)
- Creating SPSS Modeler flows in Watson Studio (60 minutes)

Lab 0 – 101 - Navigating into IBM Modeler Flows

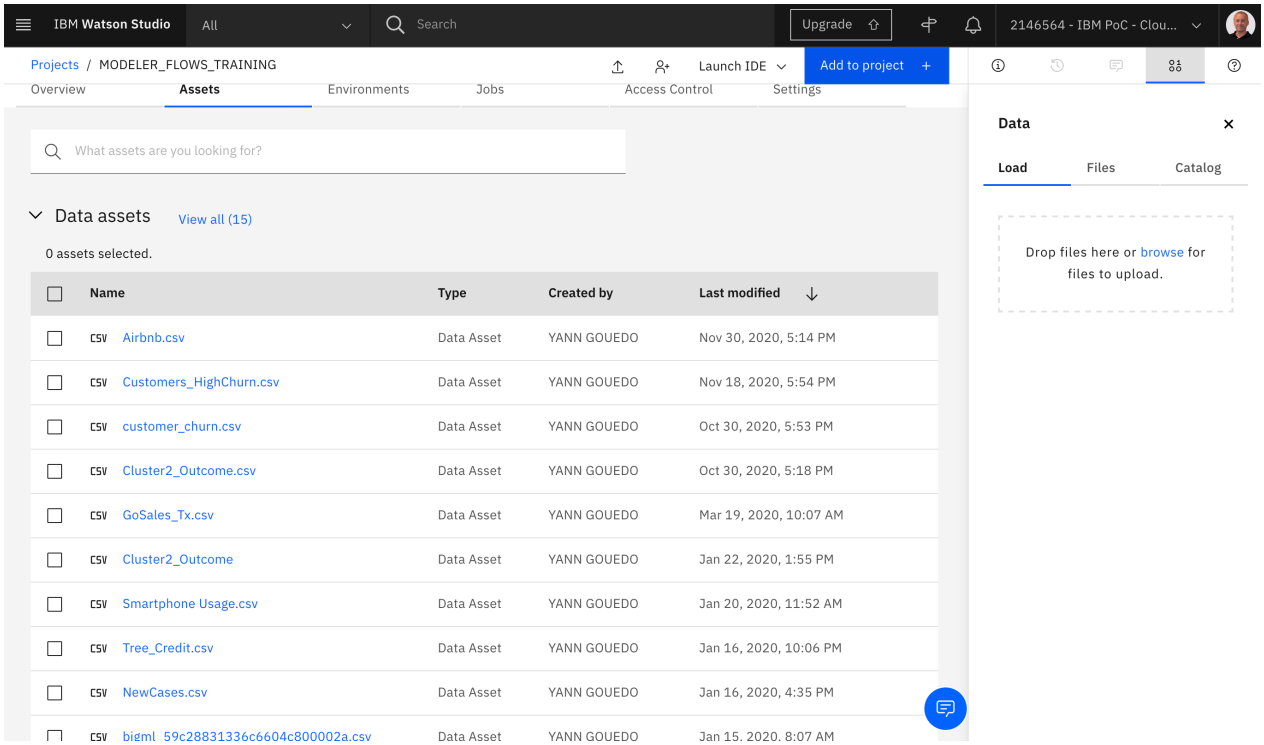
The objective of this lab is to discover Watson Studio / Modeler Flow, which allows to deliver data science activities in a visual programming way. It is the "101" discovery session of the tool.

1 - **Create an empty project** into Watson Studio. You will need to create a cloud object storage as a pre requisite.

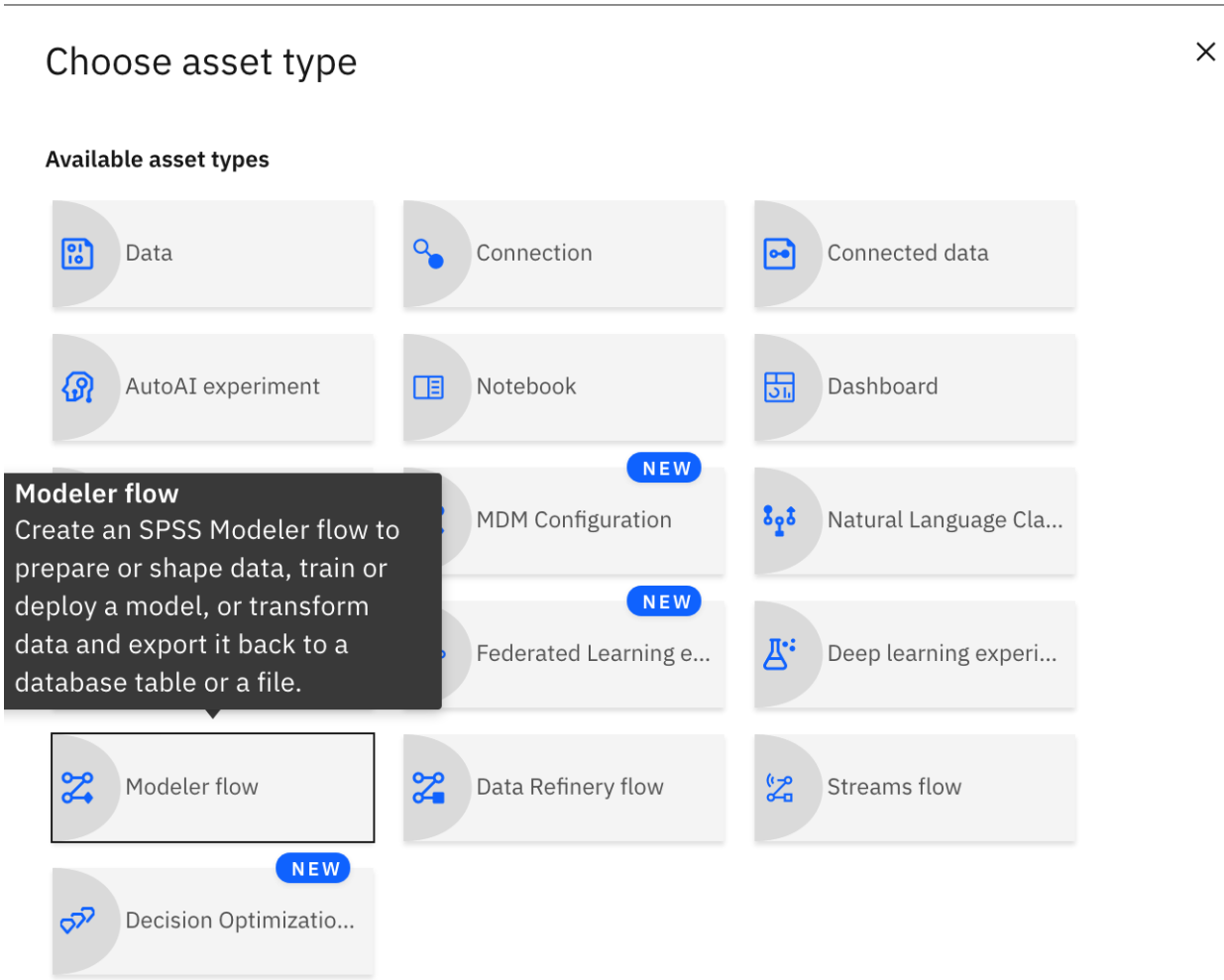


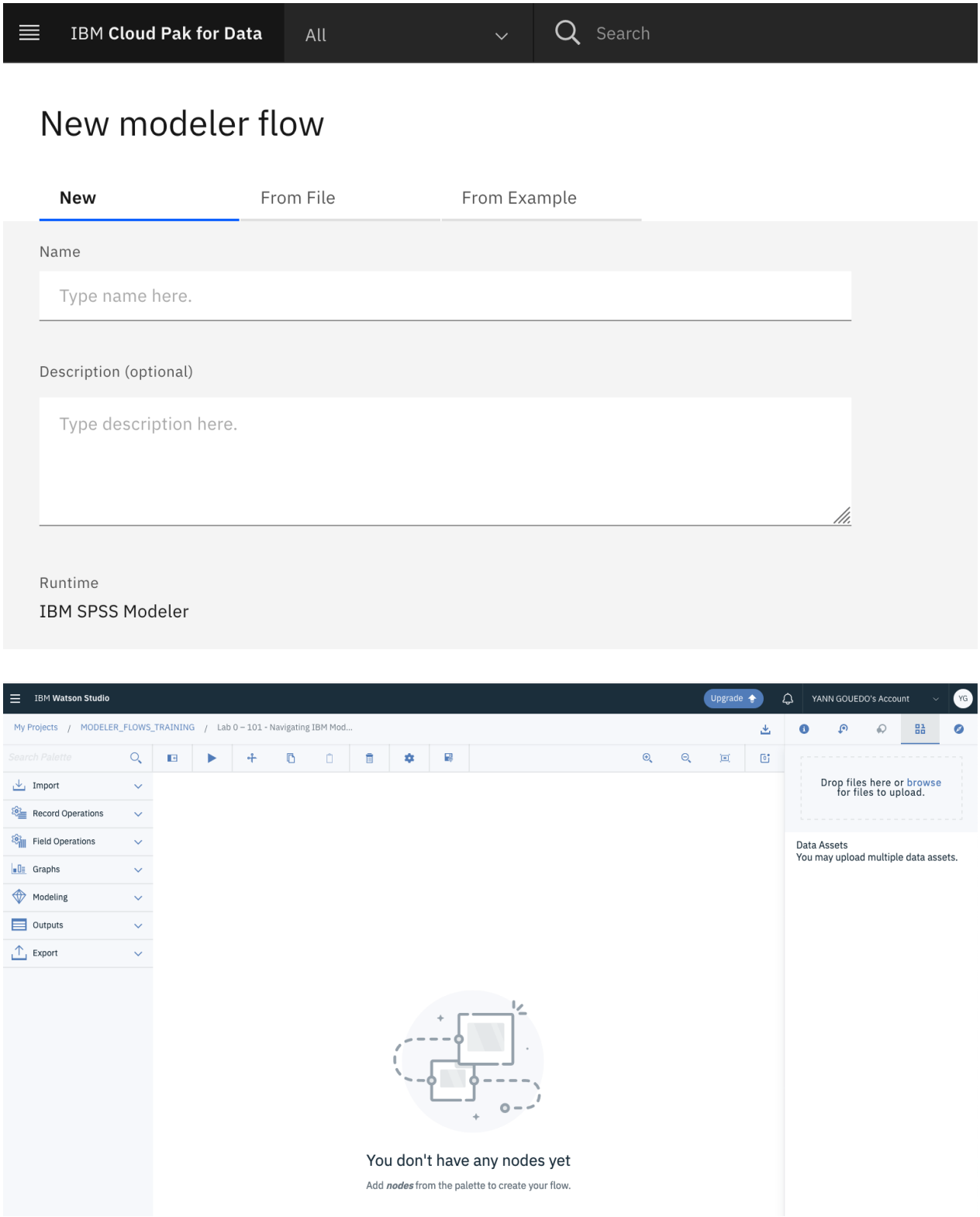


2 - **Upload hands on labs files** into the Data assets section, as the following:

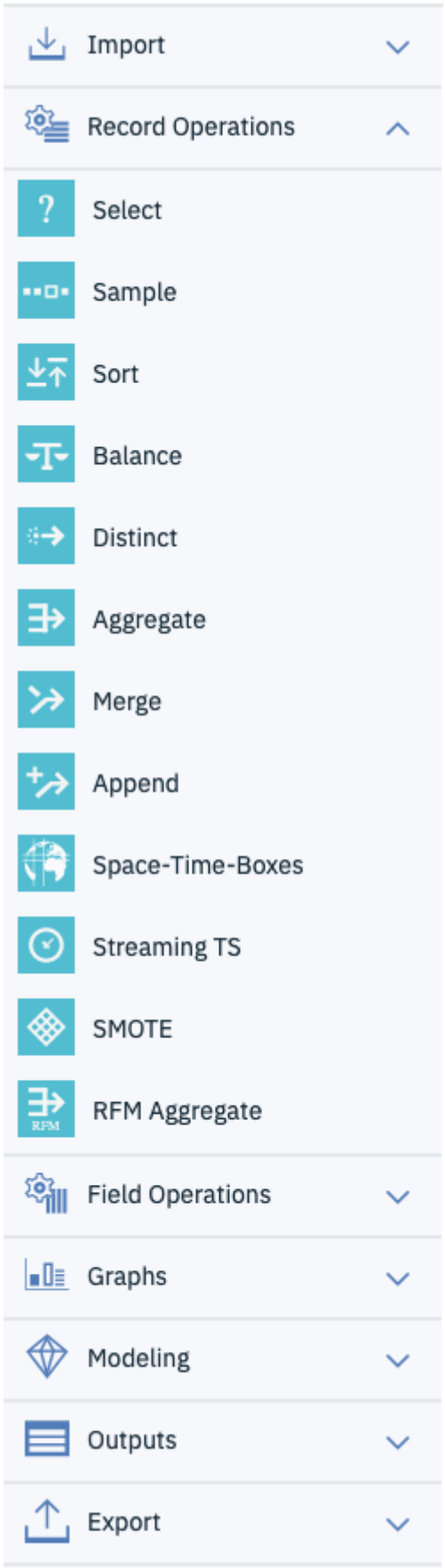


3 - **Add the Modeler Flow** type of asset and **create a new flow**.


Choose asset type
Available asset types
Data
Connection
Connected data
AutoAI experiment
Notebook
Dashboard
NEW
MDM Configuration
Natural Language Cla...
Modeler flow
Create an SPSS Modeler flow to prepare or shape data, train or deploy a model, or transform data and export it back to a database table or a file.
NEW
Federated Learning e...
Deep learning experi...
Modeler flow
Data Refinery flow
Streams flow
NEW
Decision Optimizatio...



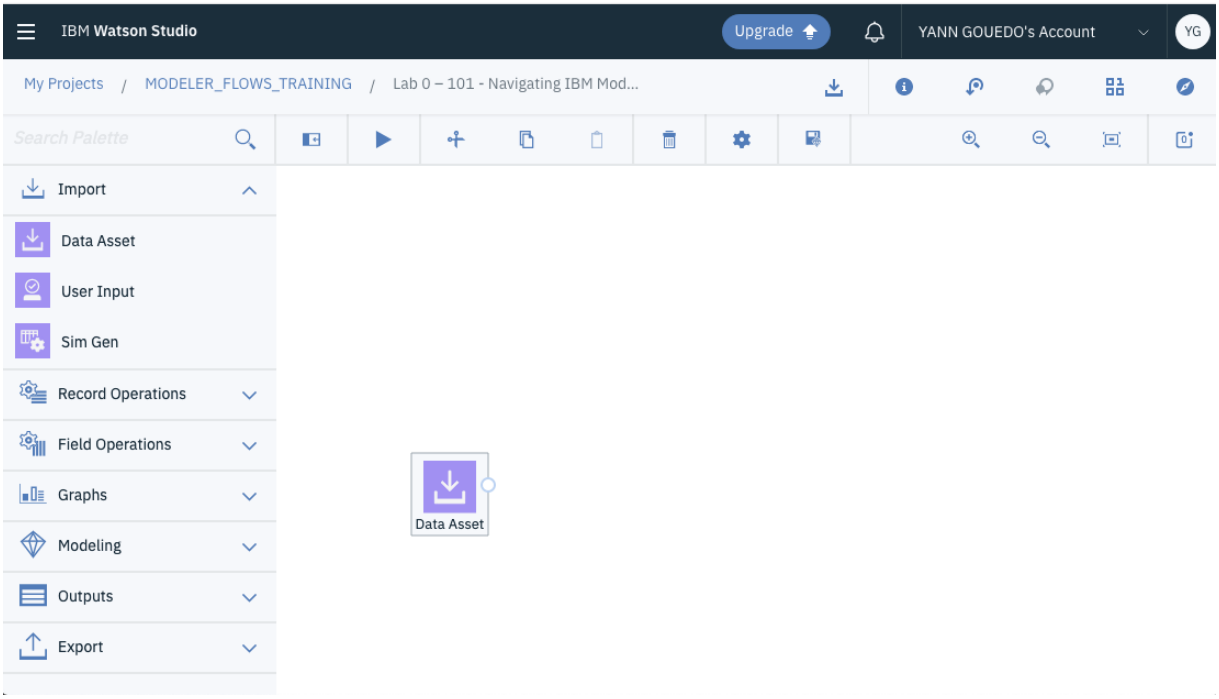
4 - **Review the various palettes** at the left of the screen. The names of the nodes reflect the operation or function accomplished with that node.



5 - From the Import palette, **select the Data Asset node** and **add it to the canvas**. This can be done by the following methods:

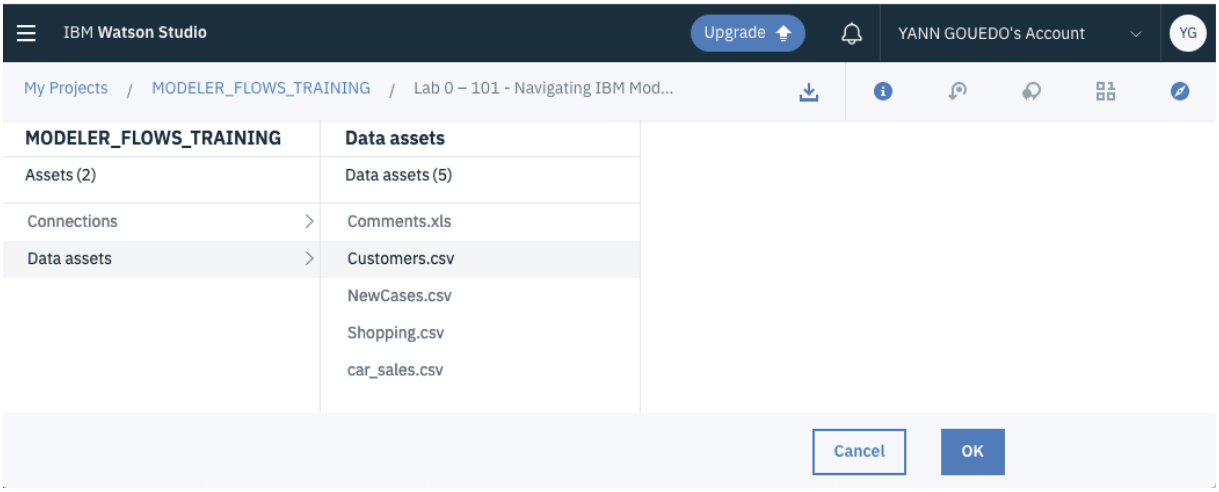
- Double-click the node on the palette.

- Drag and drop the node onto the canvas.

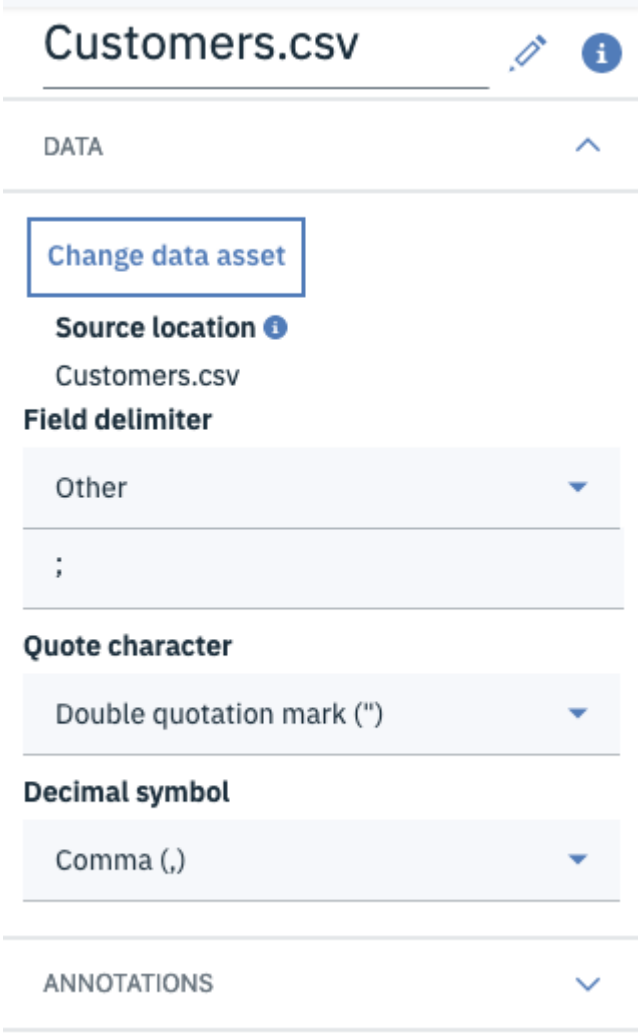


6 - Define the data asset to use

- Double-click on the Data Asset node on the canvas.
- Select *Customers.csv* clicking on Change data asset / Data assets.



- Ensure to have the following parameters:

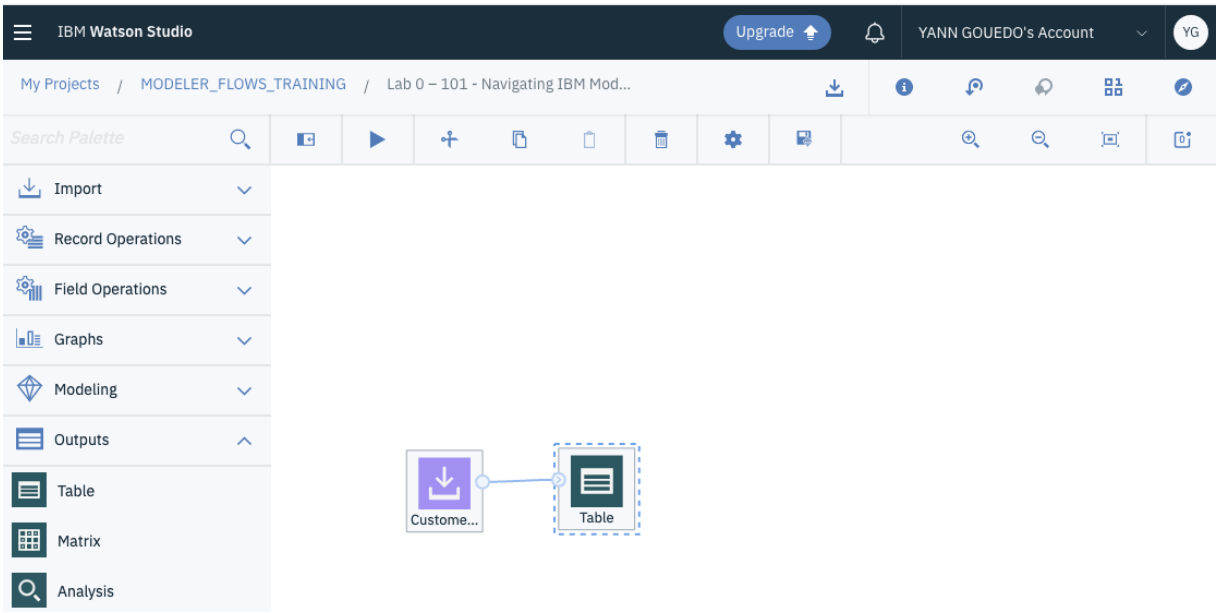


- Save the modification.
- Right click on the Customers.csv file node and select Preview from the menu, to preview the data.

ID	Sex	Status	Children	Est_Income	Car_Owner	Usage	Age	RatePlan	Lon
1	F	S	1	38000.000	N	229.640	24.393	3	23.5
6	M	M	2	29616.000	N	75.290	49.427	2	29.7
8	M	M	0	19732.800	N	47.250	50.673	3	24.8
11	M	S	2	96.330	N	59.010	56.473	1	26.1
14	F	M	2	52004.800	N	28.140	25.140	1	5.03
17	M	M	2	53010.800	N	58.870	18.840	1	12.4
18	M	M	1	75004.500	N	58.720	64.800	1	26.5

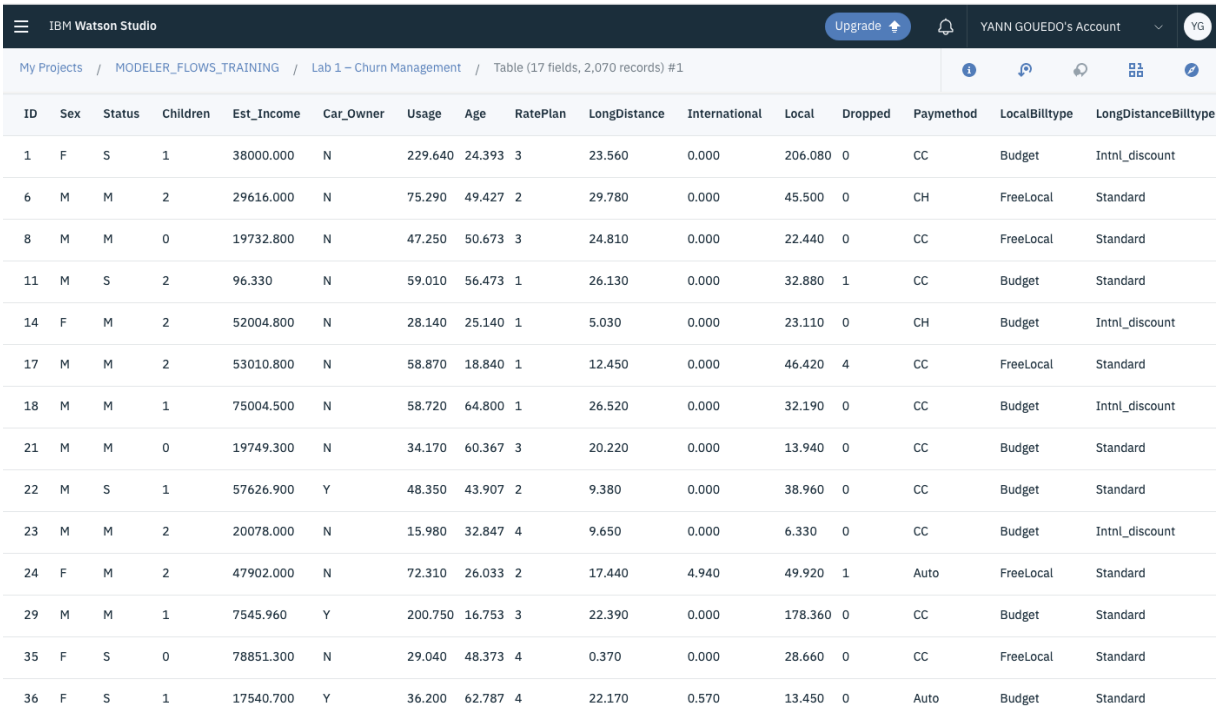
7 - From the Outputs palette, **use the Table node to see the *Customers.csv* data**, using the following methods:

- Add the Table node onto the canvas in clicking twice. The node is automatically add to the flow. From the Customers.csv file node, connect the Table node.
- Drag and drop the Table node to the canvas. From the Customers.csv file node, connect the Table node.
- First, select the Customers.csv file node on the canvas, then add the Table node onto the canvas in clicking twice. The connection is done automatically between the both nodes.



- Right click on the Table node and select Run from the context menu, to generate the table. It contains 2070 records.

- Double click on the table to see the complete data set.



ID	Sex	Status	Children	Est_Income	Car_Owner	Usage	Age	RatePlan	LongDistance	International	Local	Dropped	Paymethod	LocalBilltype	LongDistanceBilltype
1	F	S	1	38000.000	N	229.640	24.393	3	23.560	0.000	206.080	0	CC	Budget	Intnl_discount
6	M	M	2	29616.000	N	75.290	49.427	2	29.780	0.000	45.500	0	CH	FreeLocal	Standard
8	M	M	0	19732.800	N	47.250	50.673	3	24.810	0.000	22.440	0	CC	FreeLocal	Standard
11	M	S	2	96.330	N	59.010	56.473	1	26.130	0.000	32.880	1	CC	Budget	Standard
14	F	M	2	52004.800	N	28.140	25.140	1	5.030	0.000	23.110	0	CH	Budget	Intnl_discount
17	M	M	2	53010.800	N	58.870	18.840	1	12.450	0.000	46.420	4	CC	FreeLocal	Standard
18	M	M	1	75004.500	N	58.720	64.800	1	26.520	0.000	32.190	0	CC	Budget	Intnl_discount
21	M	M	0	19749.300	N	34.170	60.367	3	20.220	0.000	13.940	0	CC	Budget	Standard
22	M	S	1	57626.900	Y	48.350	43.907	2	9.380	0.000	38.960	0	CC	Budget	Standard
23	M	M	2	20078.000	N	15.980	32.847	4	9.650	0.000	6.330	0	CC	Budget	Intnl_discount
24	F	M	2	47902.000	N	72.310	26.033	2	17.440	4.940	49.920	1	Auto	FreeLocal	Standard
29	M	M	1	7545.960	Y	200.750	16.753	3	22.390	0.000	178.360	0	CC	Budget	Standard
35	F	S	0	78851.300	N	29.040	48.373	4	0.370	0.000	28.660	0	CC	FreeLocal	Standard
36	F	S	1	17540.700	Y	36.200	62.787	4	22.170	0.570	13.450	0	Auto	Budget	Standard

8 - **To delete the connection**, right click on the joining arrow and select Delete from the context menu.
Delete the connection.

9 - **To remove a node**, select it and press the Delete button in the toolbar near the top of the screen, or right click on the node and select Delete from the context menu.

Lab 1 – Churn Management

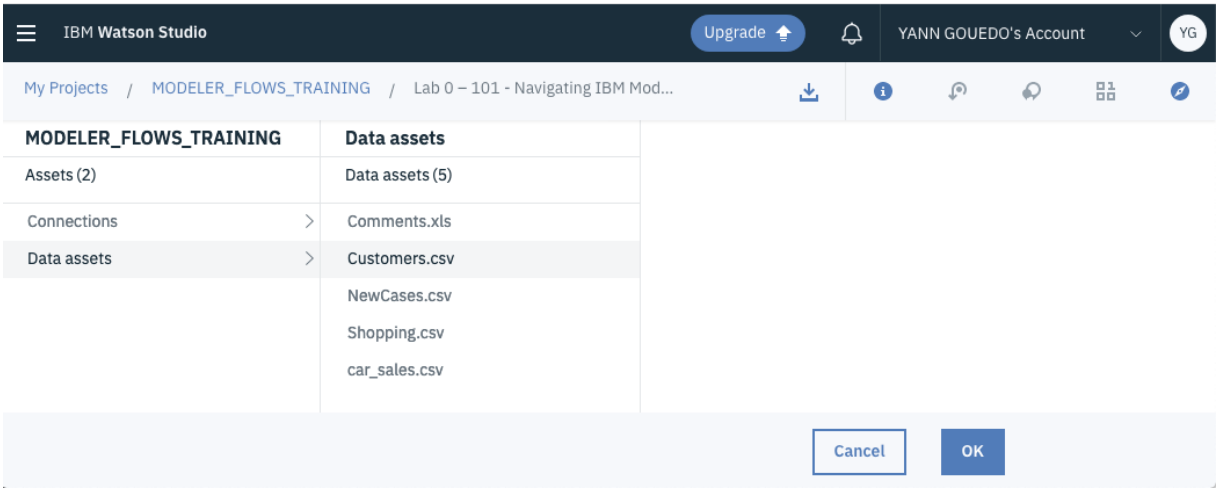
Industry context: telecommunication
Business objective: to predict and explain churners/non churners.
Technical objective: to use machine learning classification capabilities through a visual programming approach.
Data: customer descriptions, contracts

1 - **Create a new Modeler Flow** named Churn Management.

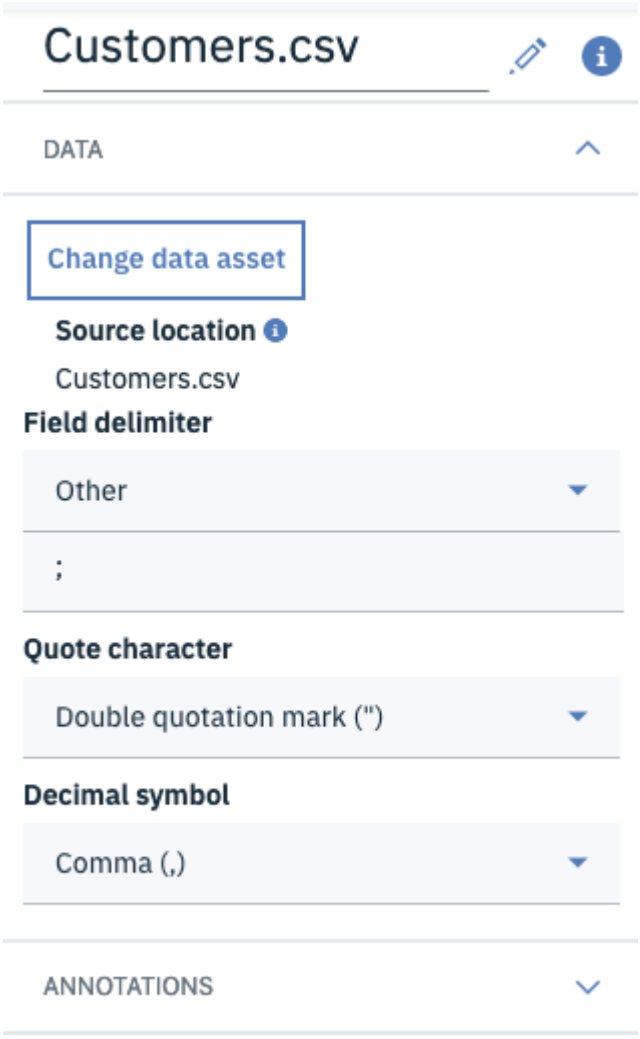
2 - From the Import palette, **add an Data Asset node**.

3 - **Define the data asset** to use:

- Double-click on the Data Asset node on the canvas.
- Select *Customers.csv* clicking on Change data asset / Data assets.



- Ensure to have the following parameters:

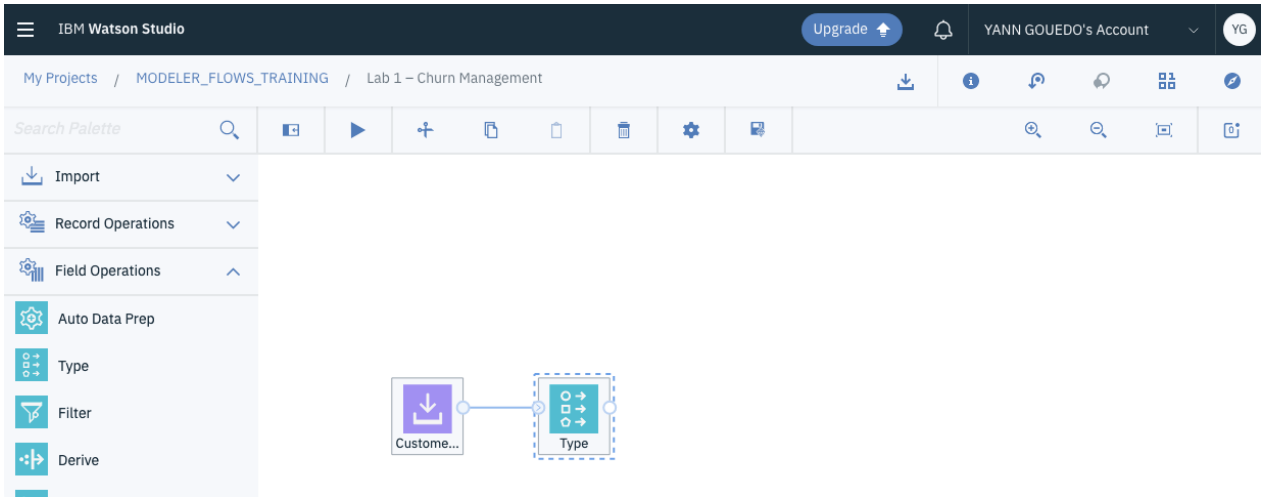


- Save the modification.

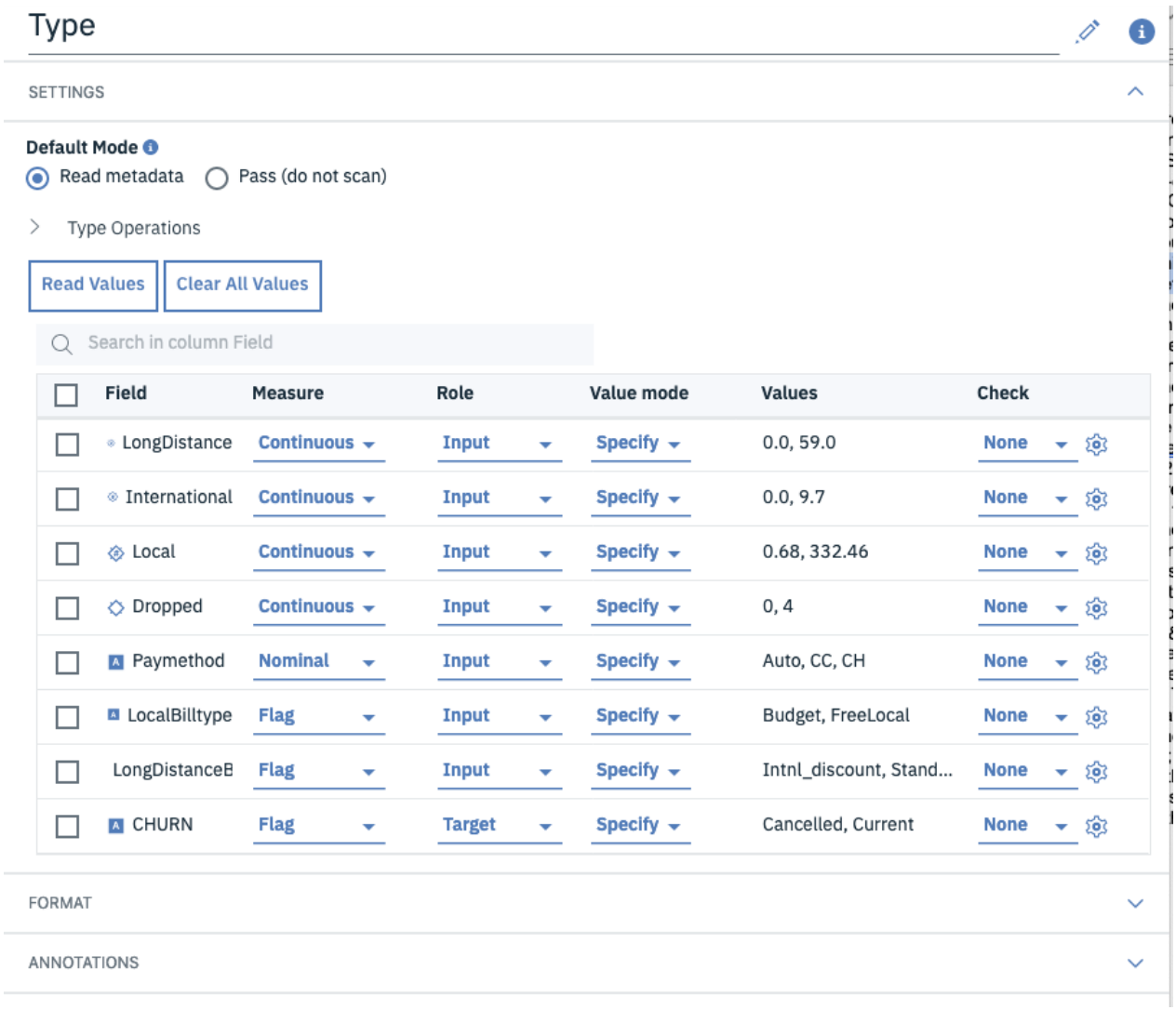
4 - **To preview the data**, right click on the *Customers.csv* file node and select Preview from the context menu. This is an extraction of data from a telecommunications company's CRM system and the dataset includes customers' ID, gender, marital status and information about their services.

ID	Sex	Status	Children	Est_Income	Car_Owner	Usage	Age	RatePlan	Lon
1	F	S	1	38000.000	N	229.640	24.393	3	23.5
6	M	M	2	29616.000	N	75.290	49.427	2	29.7
8	M	M	0	19732.800	N	47.250	50.673	3	24.8
11	M	S	2	96.330	N	59.010	56.473	1	26.1
14	F	M	2	52004.800	N	28.140	25.140	1	5.03
17	M	M	2	53010.800	N	58.870	18.840	1	12.4
18	M	M	1	75004.500	N	58.720	64.800	1	26.5

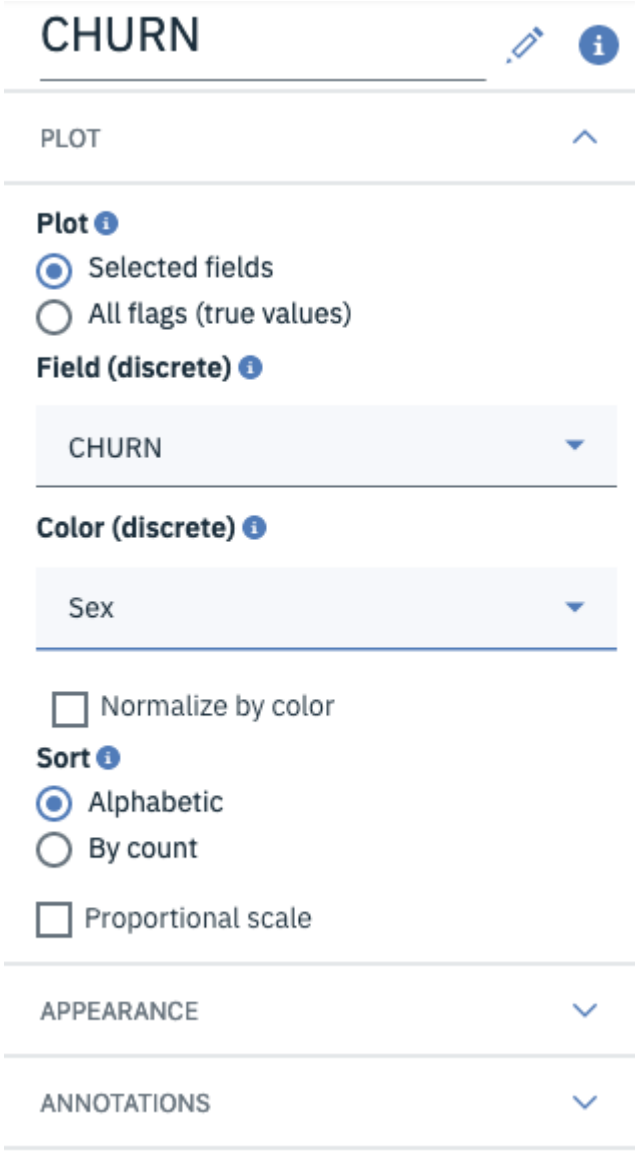
5 - From the Field Operations palette, **add a Type node** to the canvas and connect it to the data source.



6 - **Open the Type node**, clicking twice, and **click the Read Values** button to scan the data as well as to display and update the range of values. **Modify the Role** of the *ID field* to Record ID and *CHURN* field to Target, then save the modification.

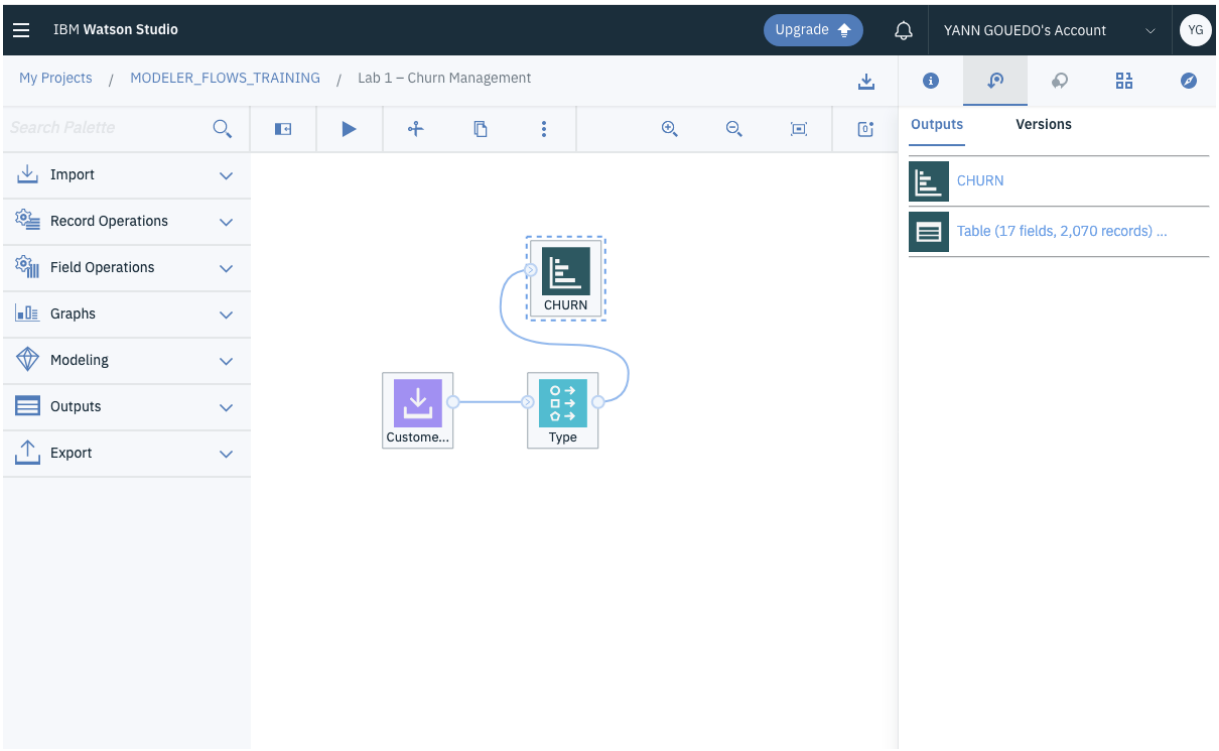


7 - From the Graphs palette, **add a Distribution node** and connect it to the Type node. **Edit the Distribution node** and select *CHURN* for the field and *Sex* for the Color, then save it.

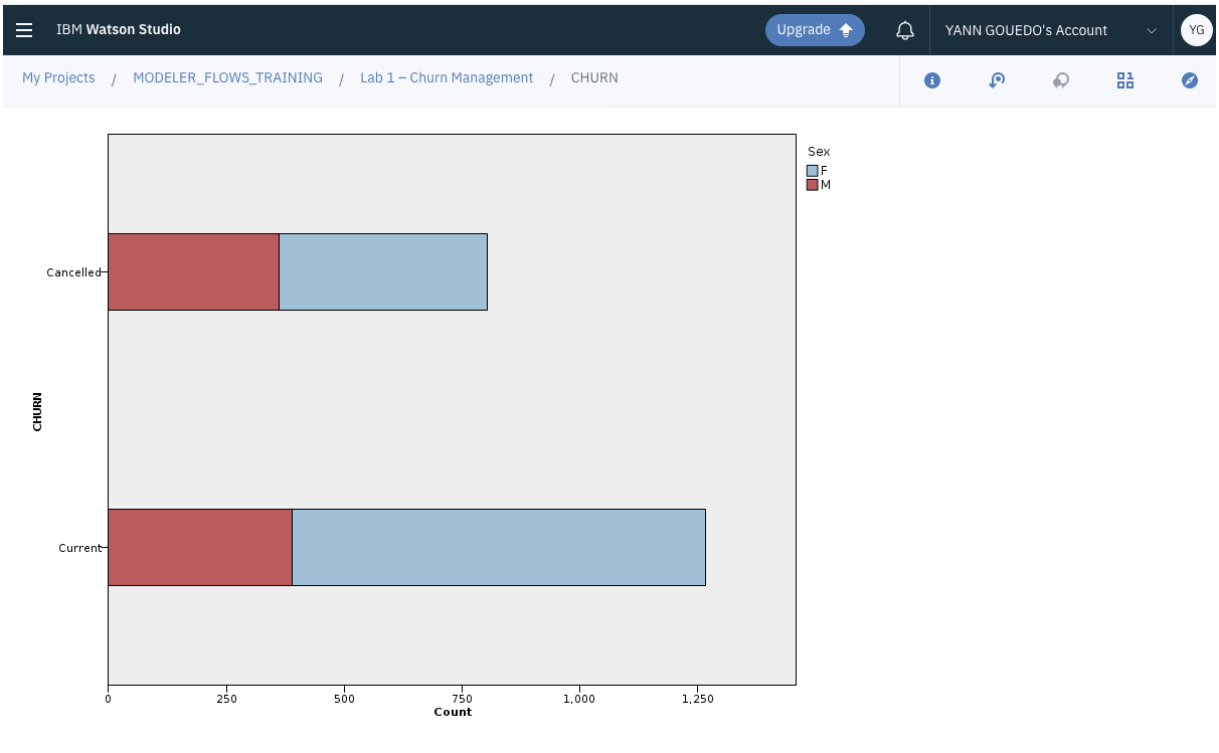


8 - **Produce the distribution graph**:

- Click right on the node and select Run from the context menu, or click the Run button in the toolbar near the top of the screen.

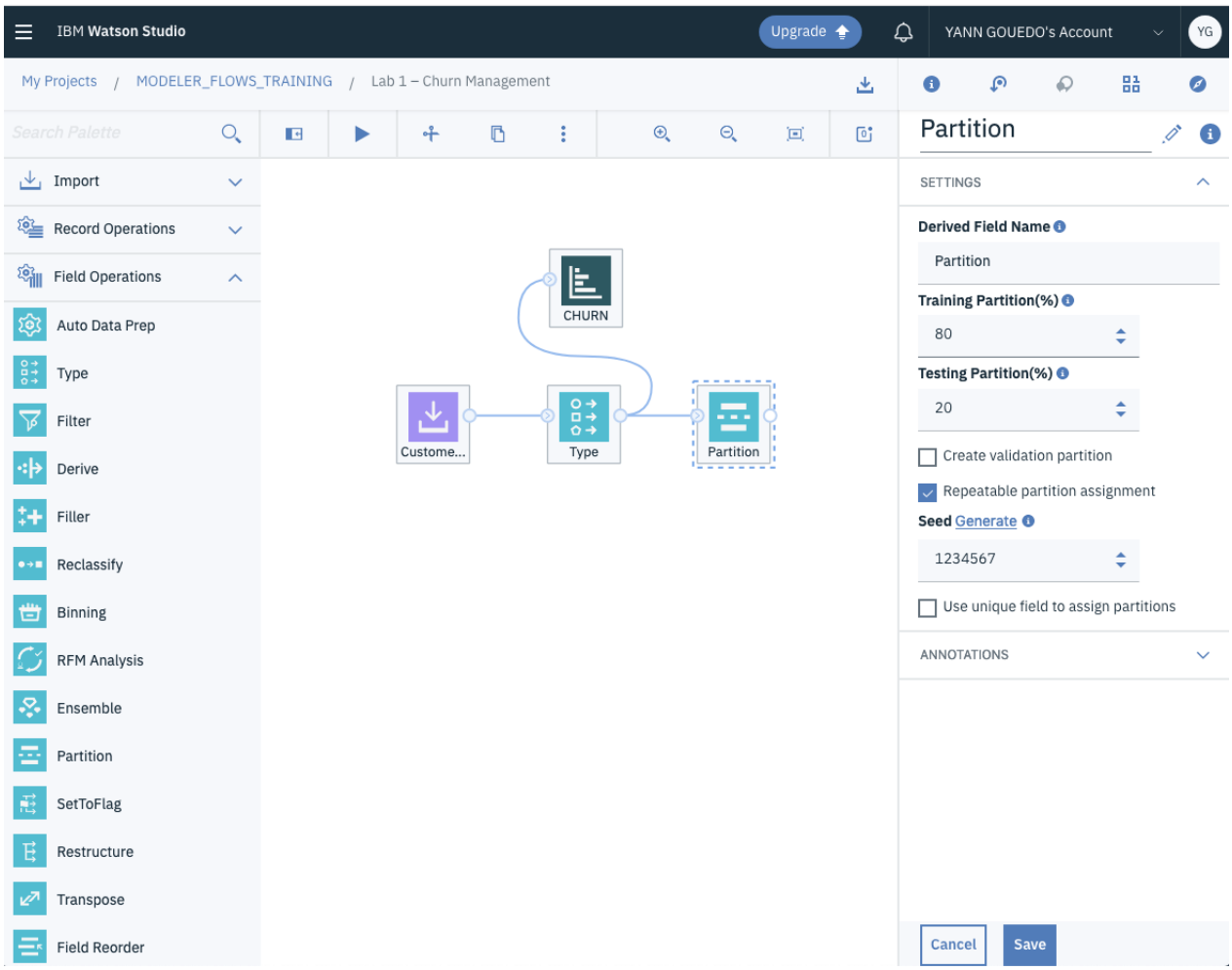


- Double click on the graph CHURN to see the distribution.

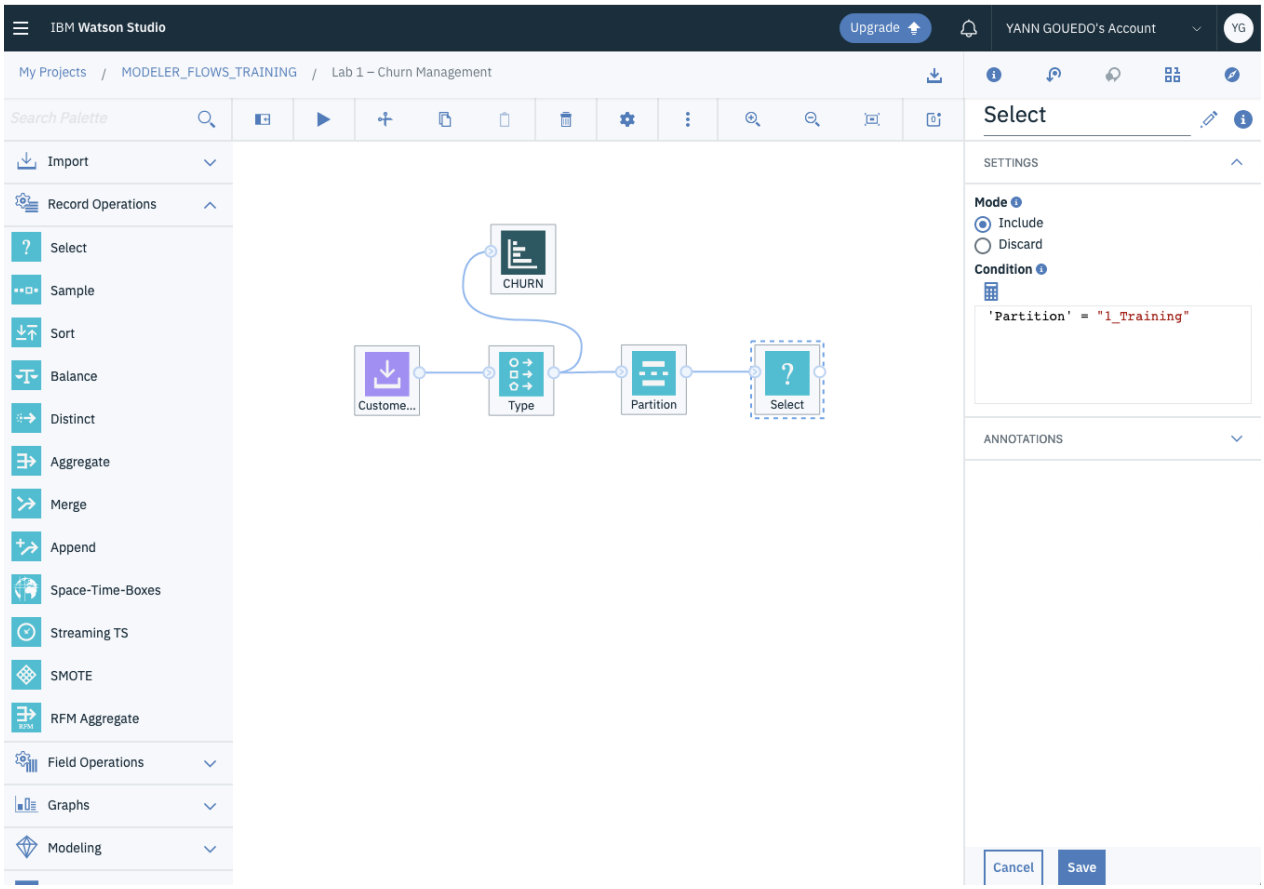


9 - Of the 2,070 customers in this dataset, about 1/3 have cancelled their contract and the remaining 2/3 are current. The task is **to build a model to understand the relationships within the data that led to the canceled their contracts**. A decision tree (like C&R Tree) is needed to do this.

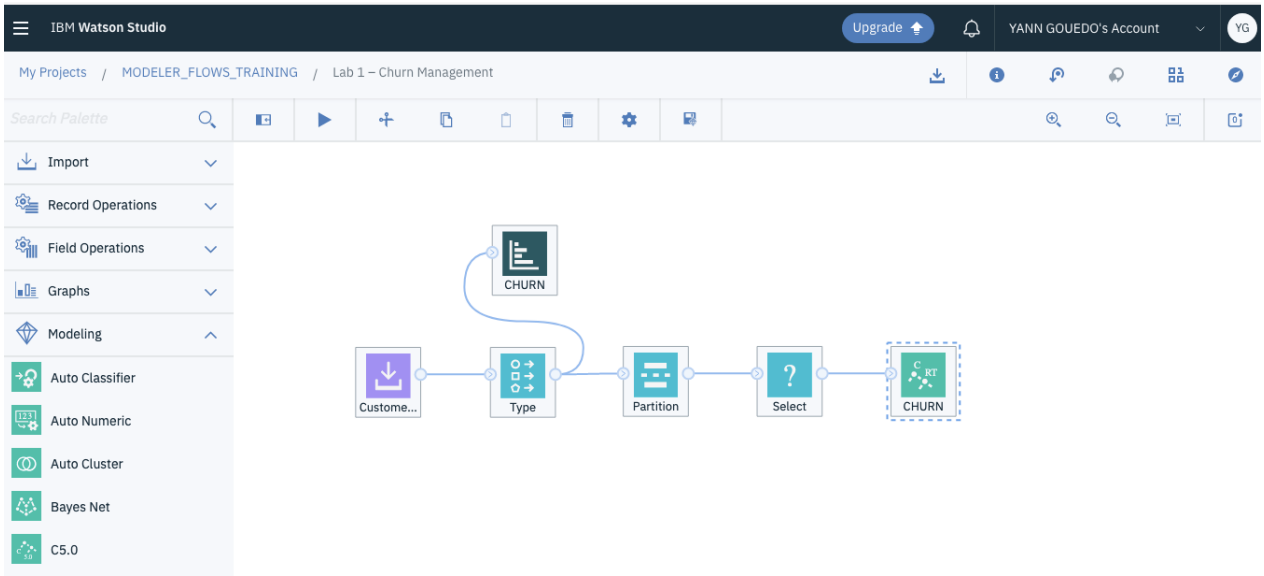
10 - In order to prepare the Modeling stage, a partioning is done (80% for the training stage; 20% for the validation stage). From the Field Operations palette, **add a Partition node** and connect it to the Type node. Edit the Partition node and **change the parameter** : Training Partition (%) = 80 and Testing Partition (%) = 20. Then save it.



11 - From the Record Operations palette, **add a Select node** and connect it to the Partition node. Edit the Select node, then use the Expression Builder **to define the following selection**: '*Partition*' = "*1_Training*". Click OK, then save it.

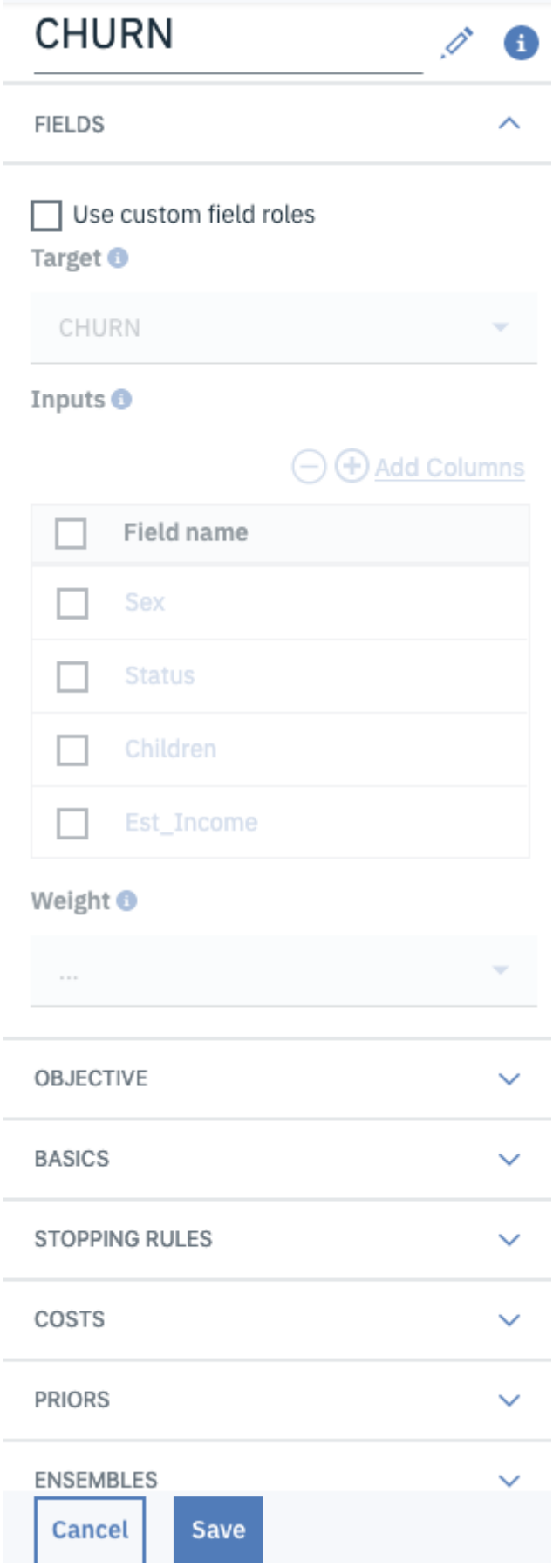


12 - From the Modeling palette, **add the C&R node** to the Select node.
Note that the C&R Tree node name changes to CHURN when it is connected, directly or indirectly, to the Type node (because of the Target Role was defined in Step 6).
Note that the C&R node is only connected to the training data (80%).

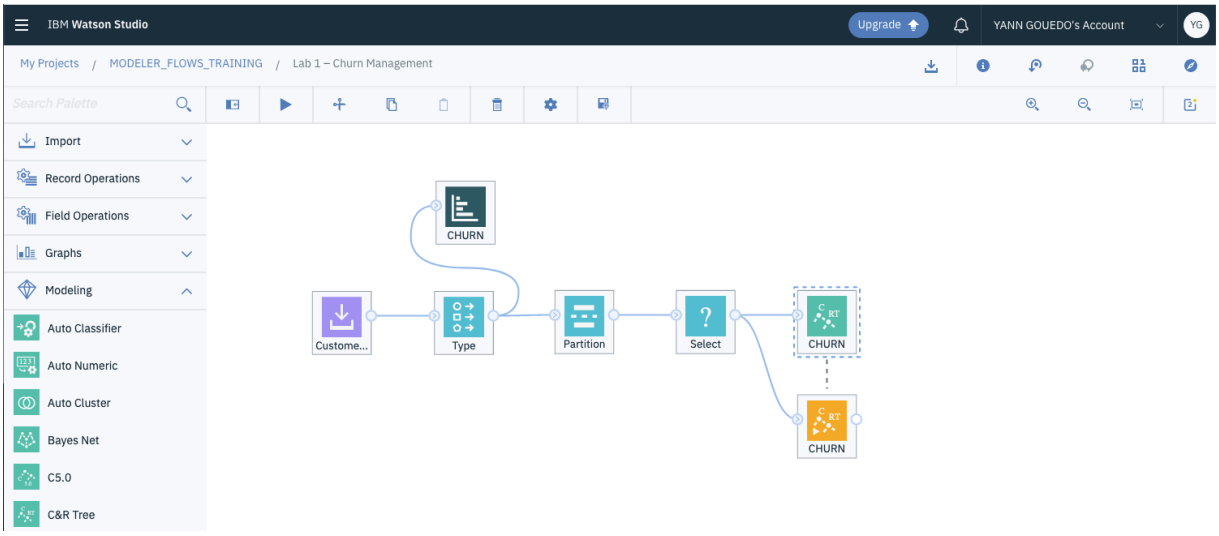


13 - **To do the Modeling stage with the C&R Tree node**:

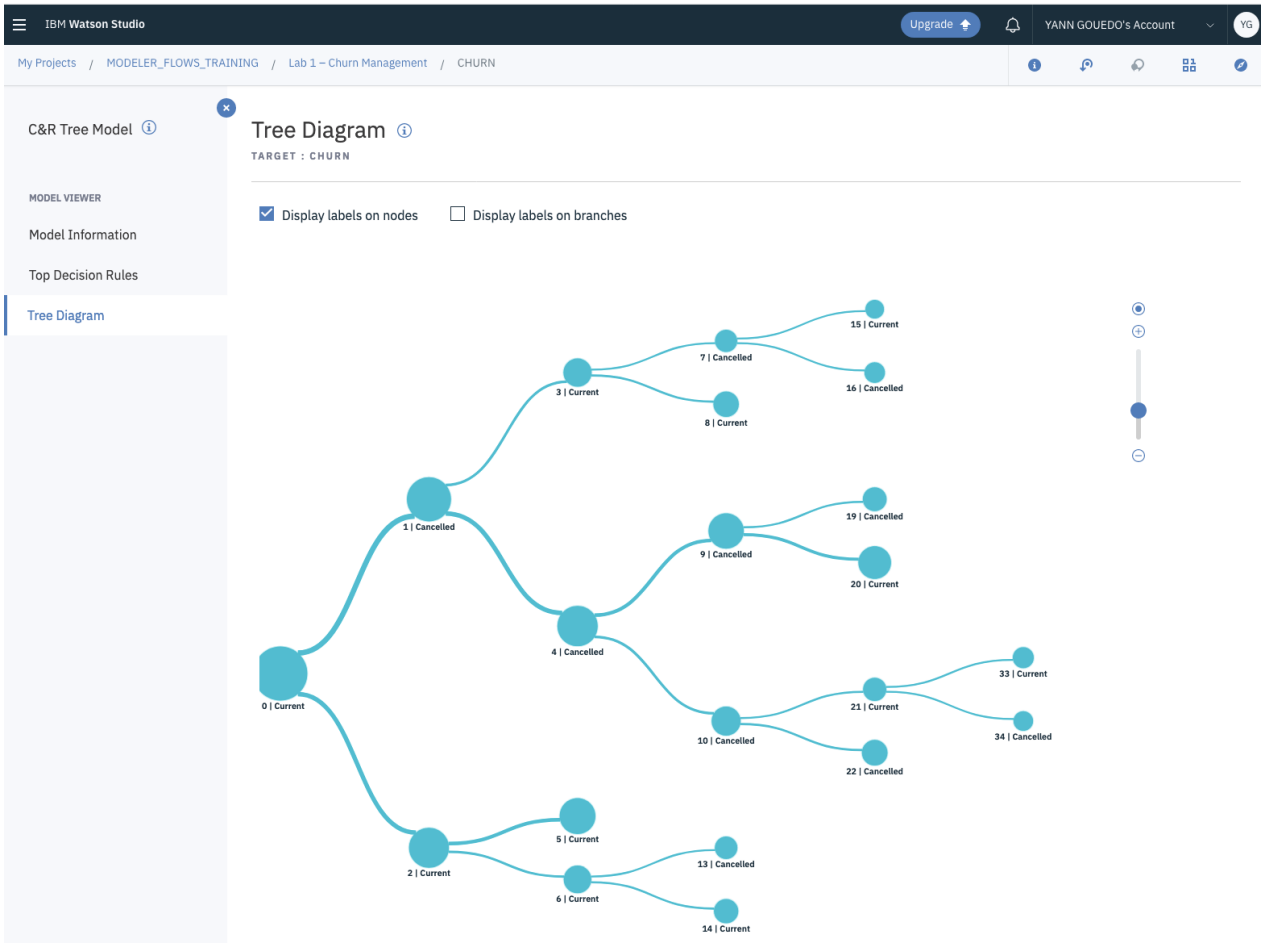
- Open the C&R Tree node to view the settings before running the model. This example uses the defaults.



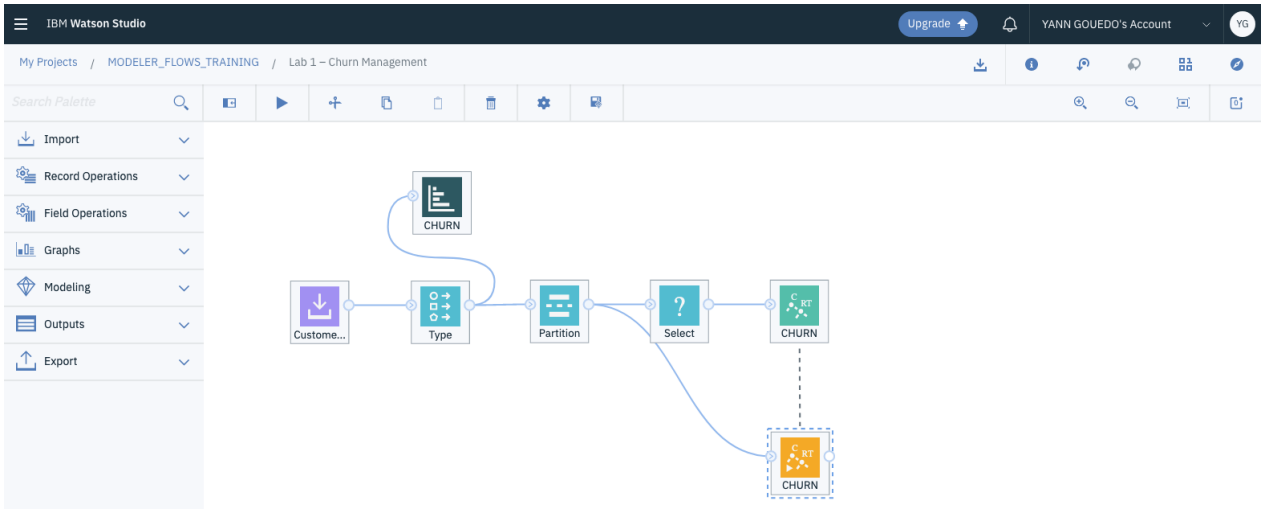
- Run the C&R Tree node from the context menu. The model (named a pattern) is generated (yellow nugget node) from 80% of the data (training dataset), and added automatically to the canvas and connected to the Select node.



14 - Right click on the CHURN model node from the context menu and **select View Model** from the context menu. It displays the decision rules and the tree diagram. Explore the outputs.

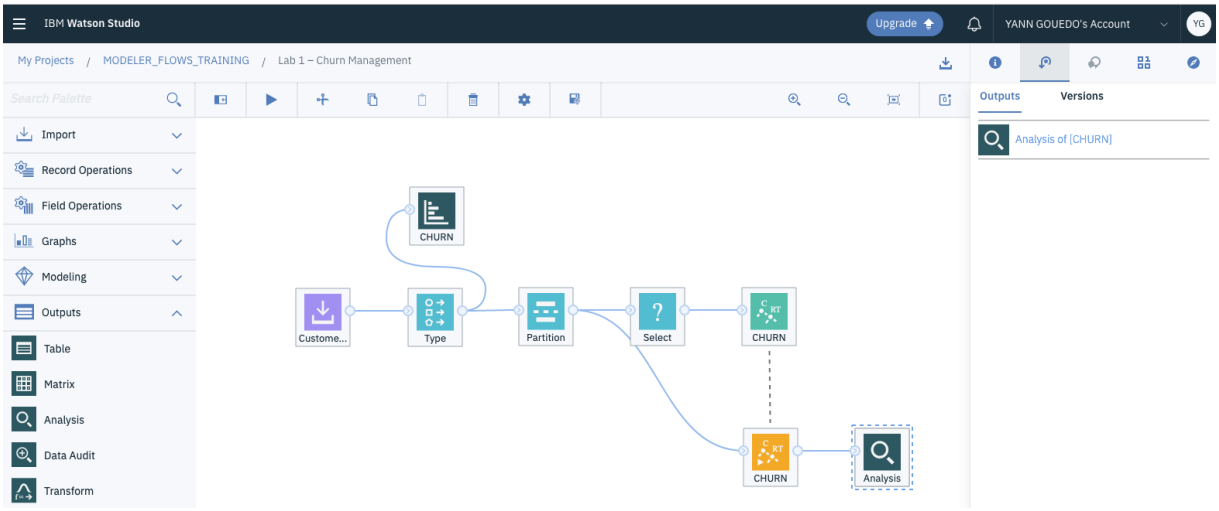


15 - **Connect the CHURN model node** directly to the Partition node (without the selection of 80% of data) as the following:

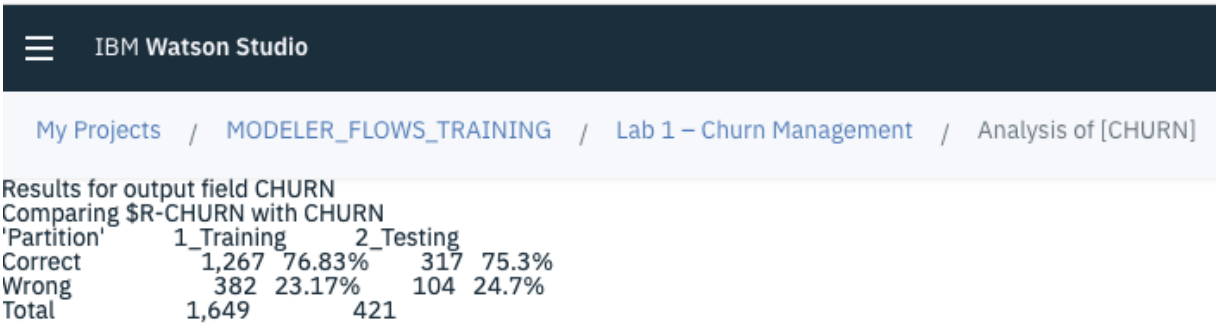


16 - From the Outputs palette, **add a Analysis node** and connect it to the CHURN model node.

- Click right on the node and select Run from the context menu, or click the Run button in the toolbar near the top of the screen.

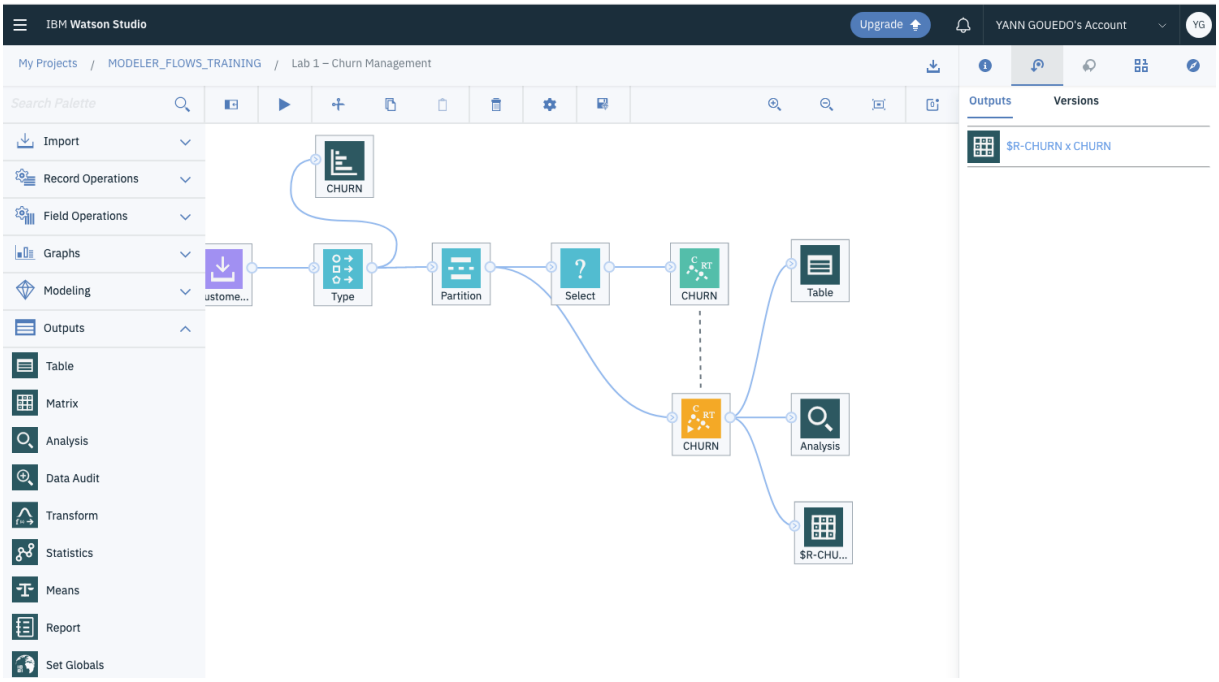


- Double click on Analysis of CHURN to see the overall accuracy of the predictive model.



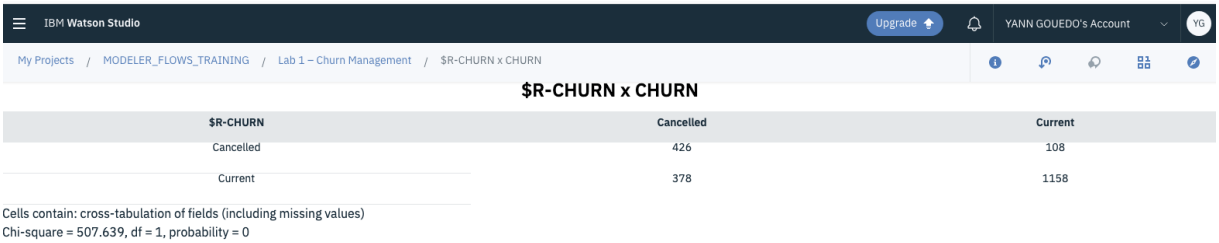
17 - From the Outputs palette, **add a Matrix node** and connect it to the CHURN model node.

- Click twice on the node and select $R-CHURN for Rows and CHURN for Columns, then save it.



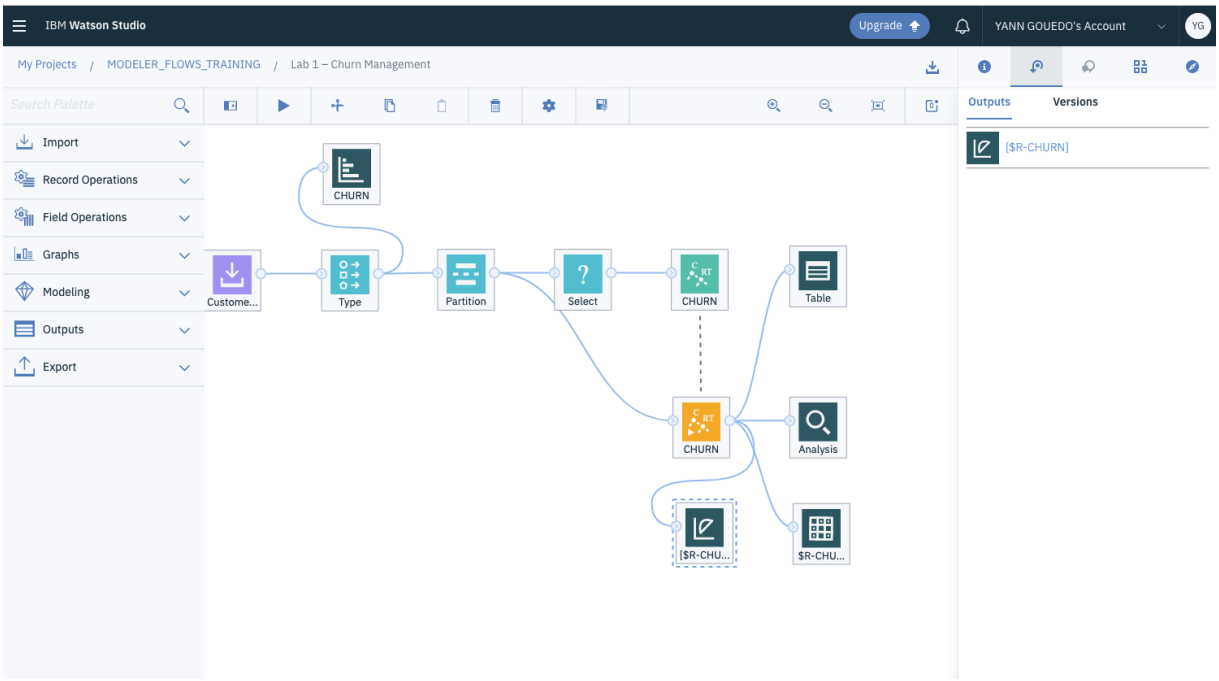
- Run the Matrix node from the context menu, or click the Run button in the toolbar near the top of the screen.

- Double click on $R-CHURN x CHURN to see the confusion matrix.



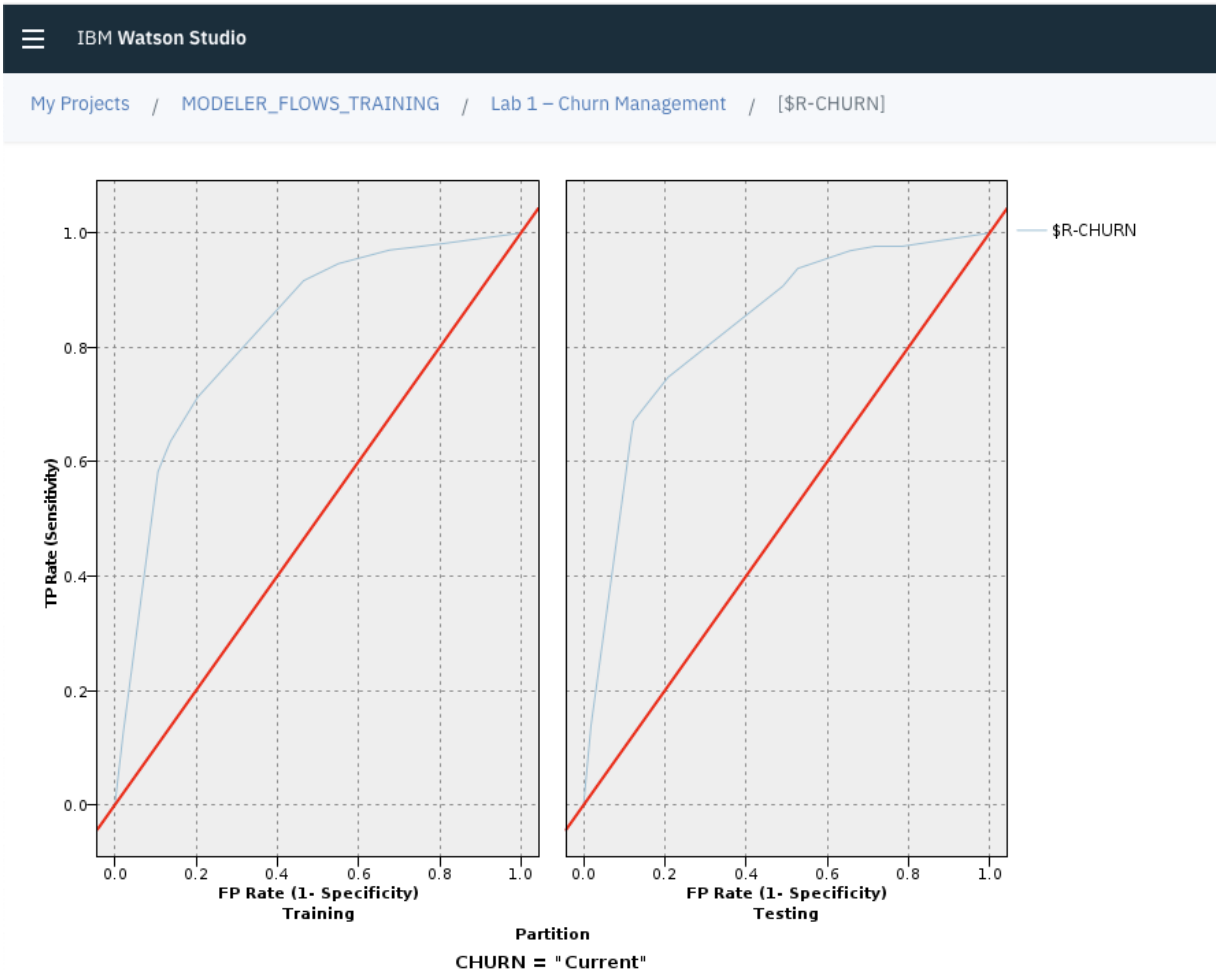
18 - From the Graphs palette, **add an Evaluation node** and connect it to the CHURN model node.

- Click twice on the node.



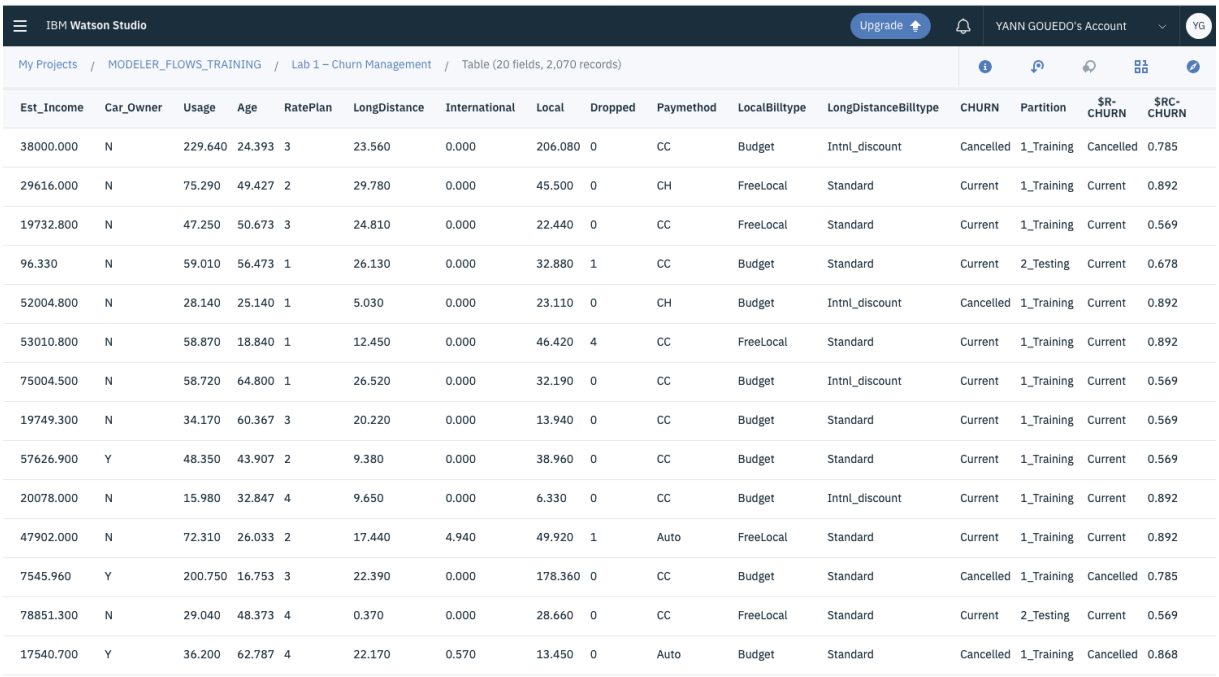
- Edit the Evaluation node, then select the ROC (Receiver Operator Characteristic) in the Chart type menu. Then, run the Evaluation node from the context menu, or click the Run button in the toolbar near the top of the screen.

- Double click on $R-CHURN to see the evaluation output. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR). It helps to see the optimal models.



19 - **Display the data and the predictive outputs**:

- From the Outputs palette, add a Table node and connect it from to the CHURN model node. Then, run it.
- Look at the last two columns. The *CHURN* column is the actual outcome (the customer cancelled or is current); the *$R-CHURN* column is the prediction; and the *$RC-CHURN* column is the confidence in the prediction.



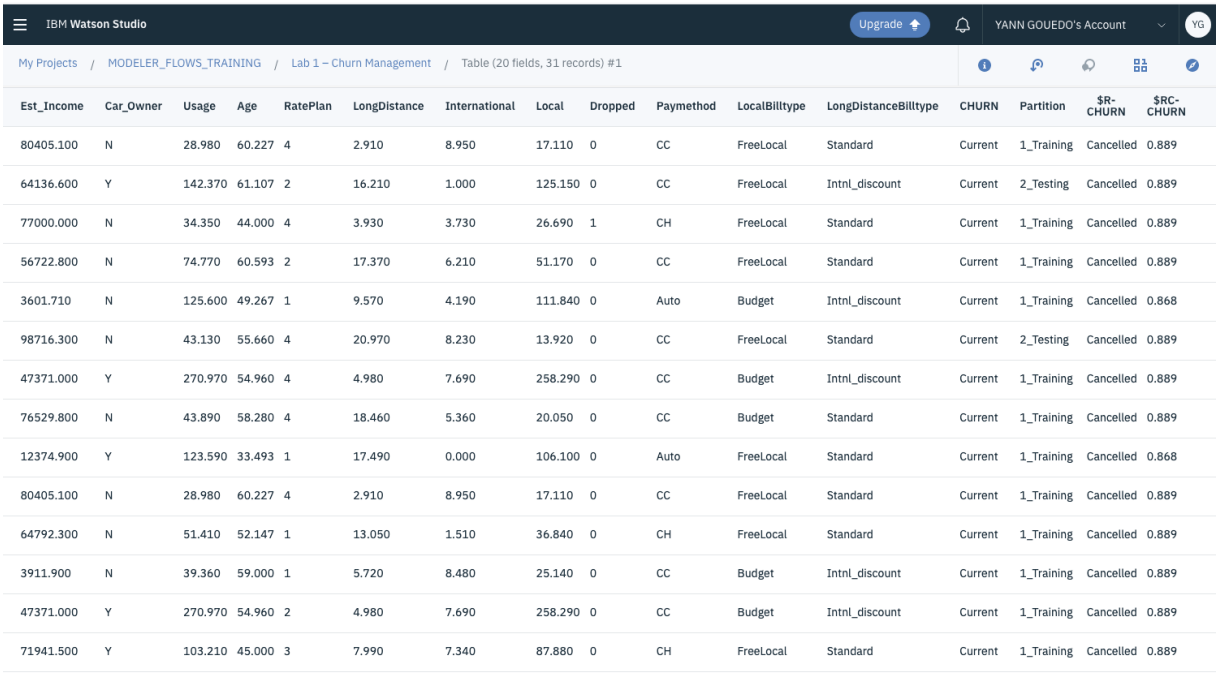
Est_Income	Car_Owner	Usage	Age	RatePlan	LongDistance	International	Local	Dropped	Paymethod	LocalBilltype	LongDistanceBilltype	CHURN	Partition	$R-CHURN	$RC-CHURN
38000.000	N	229.640	24.393	3	23.560	0.000	206.080	0	CC	Budget	Intnl_discount	Cancelled	1_Training	Cancelled	0.785
29616.000	N	75.290	49.427	2	29.780	0.000	45.500	0	CH	FreeLocal	Standard	Current	1_Training	Current	0.892
19732.800	N	47.250	50.673	3	24.810	0.000	22.440	0	CC	FreeLocal	Standard	Current	1_Training	Current	0.569
96.330	N	59.010	56.473	1	26.130	0.000	32.880	1	CC	Budget	Standard	Current	2_Testing	Current	0.678
52004.800	N	28.140	25.140	1	5.030	0.000	23.110	0	CH	Budget	Intnl_discount	Cancelled	1_Training	Current	0.892
53010.800	N	58.870	18.840	1	12.450	0.000	46.420	4	CC	FreeLocal	Standard	Current	1_Training	Current	0.892
75004.500	N	58.720	64.800	1	26.520	0.000	32.190	0	CC	Budget	Intnl_discount	Current	1_Training	Current	0.569
19749.300	N	34.170	60.367	3	20.220	0.000	13.940	0	CC	Budget	Standard	Current	1_Training	Current	0.569
57626.900	Y	48.350	43.907	2	9.380	0.000	38.960	0	CC	Budget	Standard	Current	1_Training	Current	0.569
20078.000	N	15.980	32.847	4	9.650	0.000	6.330	0	CC	Budget	Intnl_discount	Current	1_Training	Current	0.892
47902.000	N	72.310	26.033	2	17.440	4.940	49.920	1	Auto	FreeLocal	Standard	Current	1_Training	Current	0.892
7545.960	Y	200.750	16.753	3	22.390	0.000	178.360	0	CC	Budget	Standard	Cancelled	1_Training	Cancelled	0.785
78851.300	N	29.040	48.373	4	0.370	0.000	28.660	0	CC	FreeLocal	Standard	Current	2_Testing	Current	0.569
17540.700	Y	36.200	62.787	4	22.170	0.570	13.450	0	Auto	Budget	Standard	Cancelled	1_Training	Cancelled	0.868

20 - **Selection and display of the customers with a high churn rate**:

- From the Record Operations palette, add a Select node and connect it to the CHURN model node. Edit the Select node, then use the Expression Builder to define the following selection: *'CHURN'* = "*Current*" *and* *'$R-CHURN'* = "*Cancelled*" *and* *'$RC-CHURN'* >=*0.85*

  Click OK, then save it.

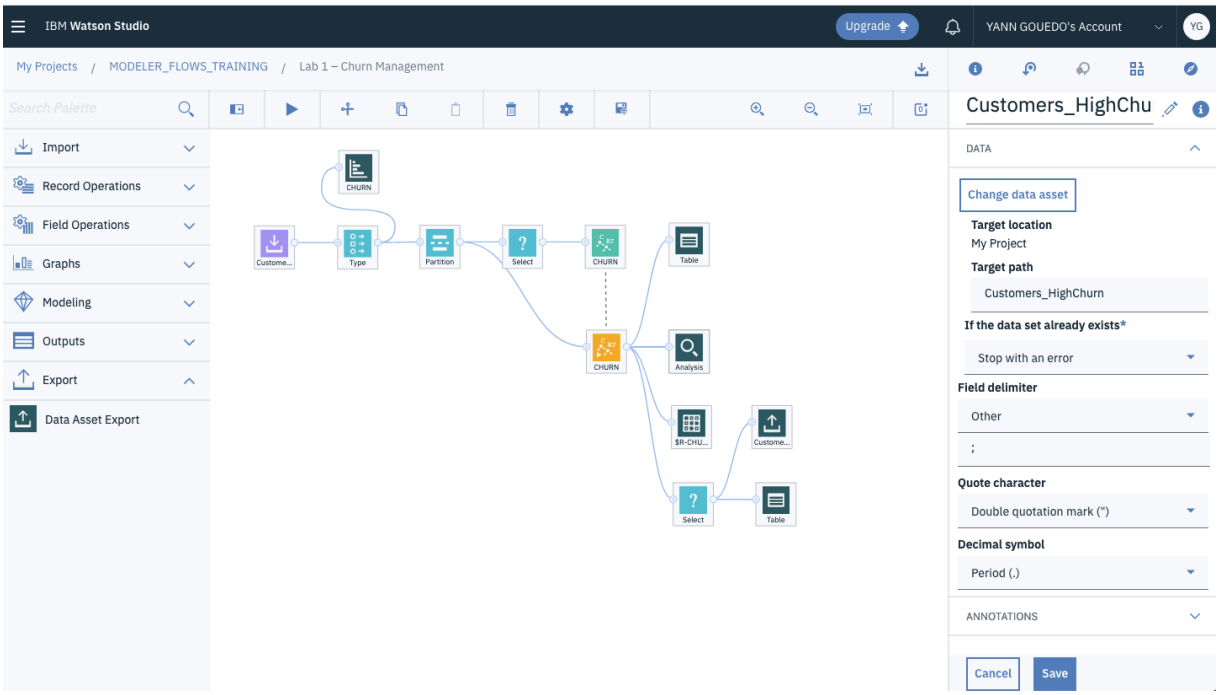
- From the Outputs palette, add a Table node and connect it to the Select node. Then, run it. It displays the list of the customers with a high risk to churn.



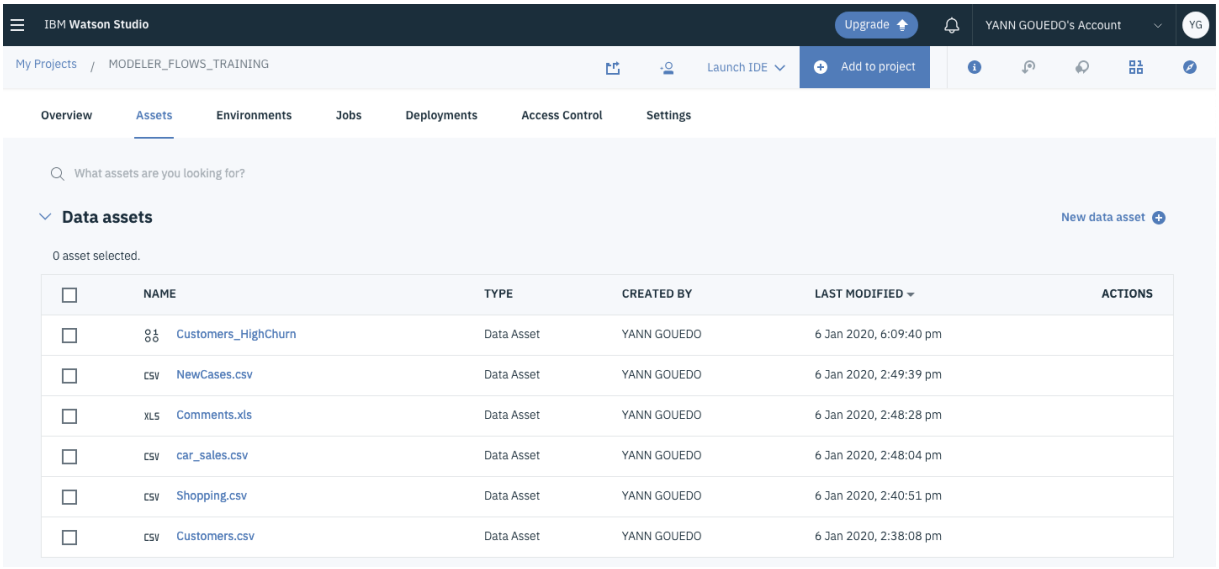
Est_Income	Car_Owner	Usage	Age	RatePlan	LongDistance	International	Local	Dropped	Paymethod	LocalBilltype	LongDistanceBilltype	CHURN	Partition	$R-CHURN	$RC-CHURN
80405.100	N	28.980	60.227	4	2.910	8.950	17.110	0	CC	FreeLocal	Standard	Current	1_Training	Cancelled	0.889
64136.600	Y	142.370	61.107	2	16.210	1.000	125.150	0	CC	FreeLocal	Intnl_discount	Current	2_Testing	Cancelled	0.889
77000.000	N	34.350	44.000	4	3.930	3.730	26.690	1	CH	FreeLocal	Standard	Current	1_Training	Cancelled	0.889
56722.800	N	74.770	60.593	2	17.370	6.210	51.170	0	CC	FreeLocal	Standard	Current	1_Training	Cancelled	0.889
3601.710	N	125.600	49.267	1	9.570	4.190	111.840	0	Auto	Budget	Intnl_discount	Current	1_Training	Cancelled	0.868
98716.300	N	43.130	55.660	4	20.970	8.230	13.920	0	CC	FreeLocal	Standard	Current	2_Testing	Cancelled	0.889
47371.000	Y	270.970	54.960	4	4.980	7.690	258.290	0	CC	Budget	Intnl_discount	Current	1_Training	Cancelled	0.889
76529.800	N	43.890	58.280	4	18.460	5.360	20.050	0	CC	Budget	Standard	Current	1_Training	Cancelled	0.889
12374.900	Y	123.590	33.493	1	17.490	0.000	106.100	0	Auto	FreeLocal	Standard	Current	1_Training	Cancelled	0.868
80405.100	N	28.980	60.227	4	2.910	8.950	17.110	0	CC	FreeLocal	Standard	Current	1_Training	Cancelled	0.889
64792.300	N	51.410	52.147	1	13.050	1.510	36.840	0	CH	FreeLocal	Standard	Current	1_Training	Cancelled	0.889
3911.900	N	39.360	59.000	1	5.720	8.480	25.140	0	CC	Budget	Intnl_discount	Current	1_Training	Cancelled	0.889
47371.000	Y	270.970	54.960	2	4.980	7.690	258.290	0	CC	Budget	Intnl_discount	Current	1_Training	Cancelled	0.889
71941.500	Y	103.210	45.000	3	7.990	7.340	87.880	0	CH	FreeLocal	Standard	Current	1_Training	Cancelled	0.889

21 - **Export the customers with a high churn rate**:

- From the Export palette, add a Data Asset Export node and connect it to the previous Select node.
- Change the name, change the parameters, and define a Target Path, as the following. Then save it and run it.



- Go back to the Data assets section view. The export file is here.

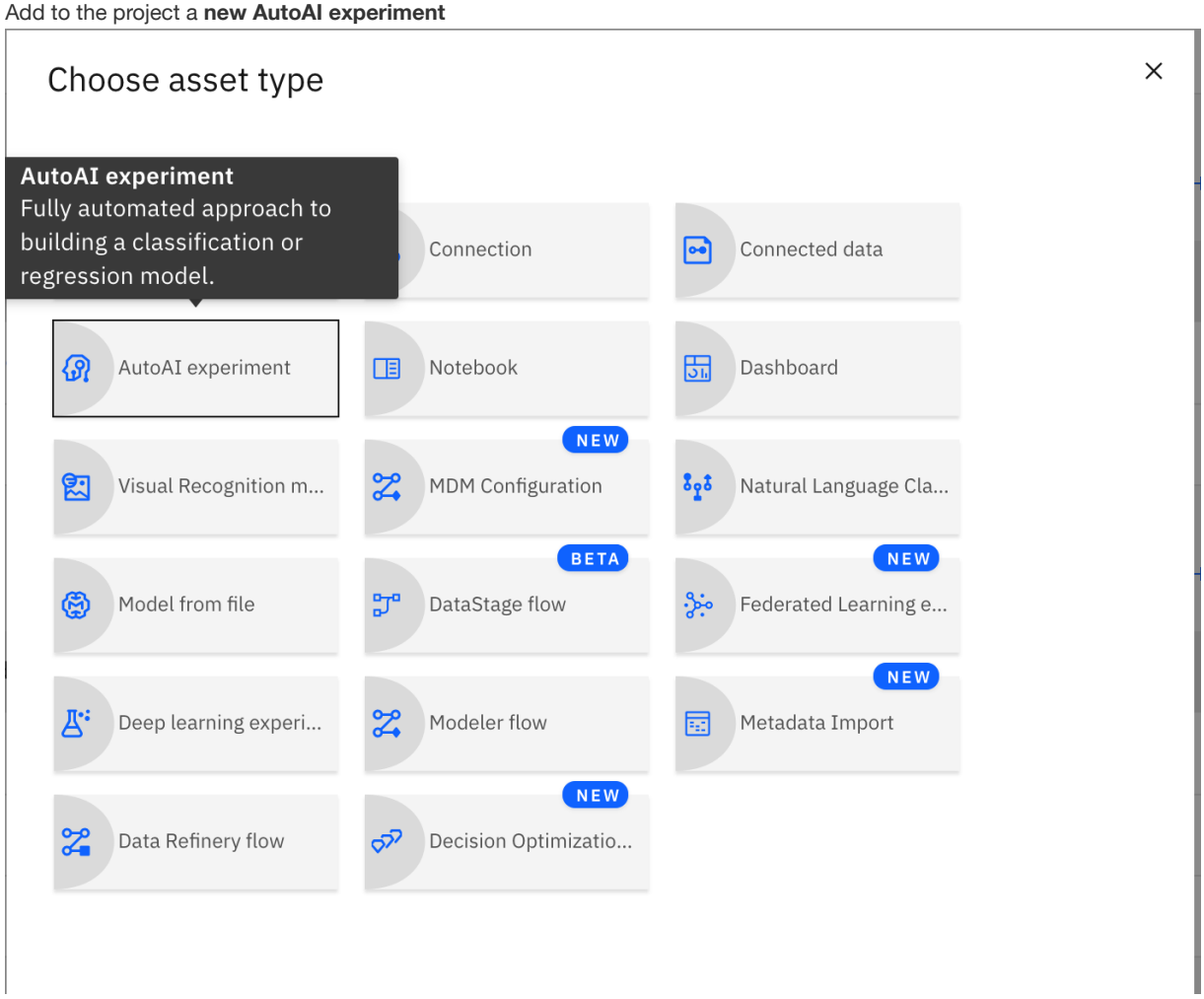


Lab 2 – Churn Management with Automated Machine Learning (with AutoAI)

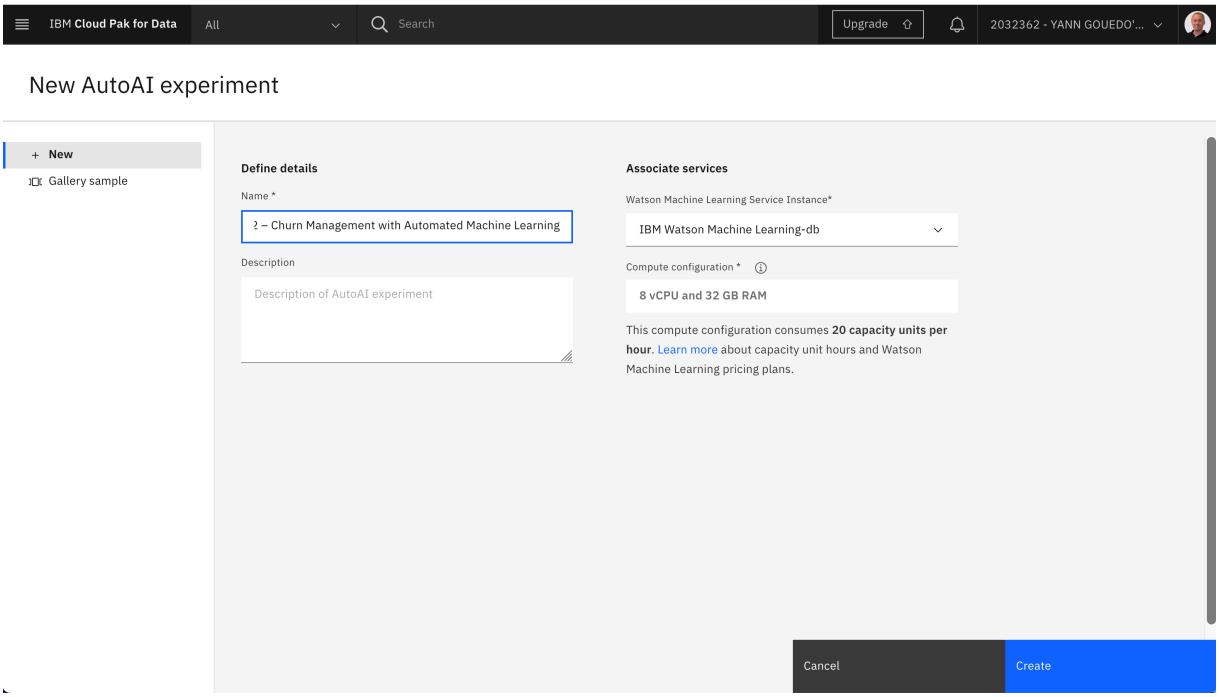
Industry context: telecommunication
Business objective: to predict and explain churners/non churners.
Technical objective: to use automatic machine learning data science capabilities.
Data: customer descriptions, contracts

1 - Create a **new AutoAI experiment**, which will allow to detect the different models that predict & explain churners / non churners.

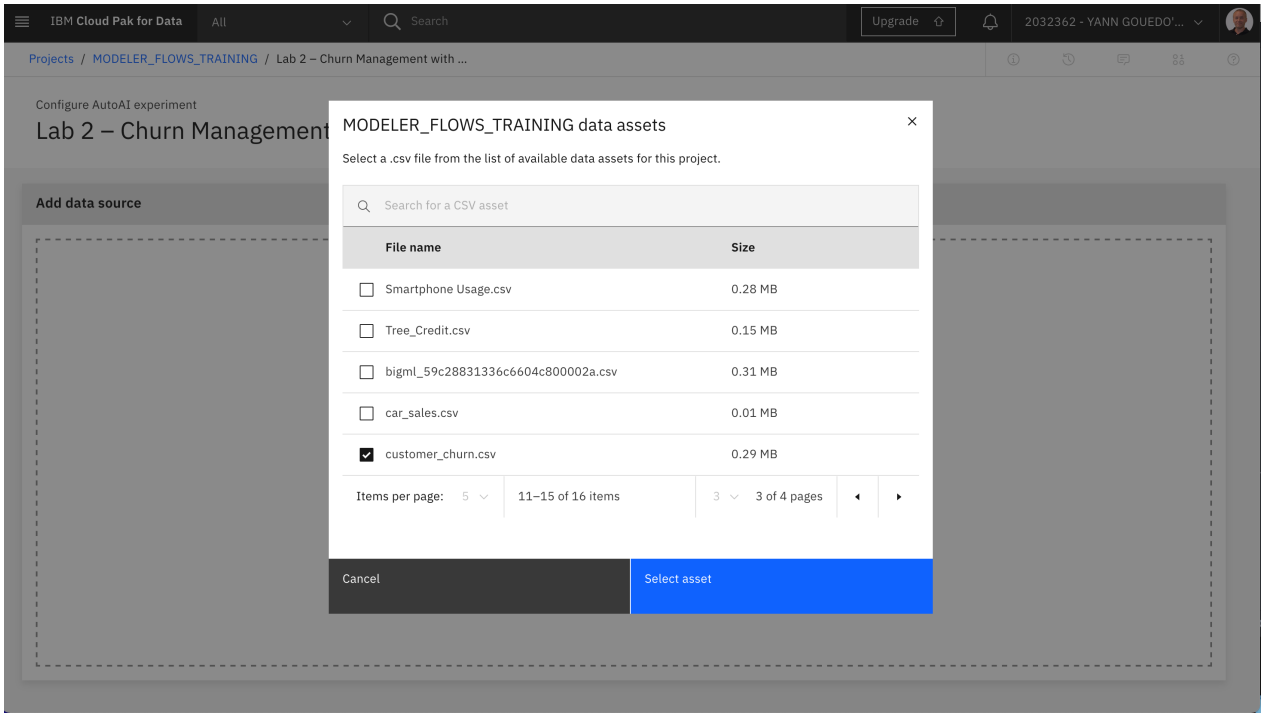
- Add to the project a **new AutoAI experiment**



- Name it "Lab 2 – Churn Management with Automatic Machine Learning.", for instance, and verify that a Watson Machine Learning service instance is listed in the windows, as below

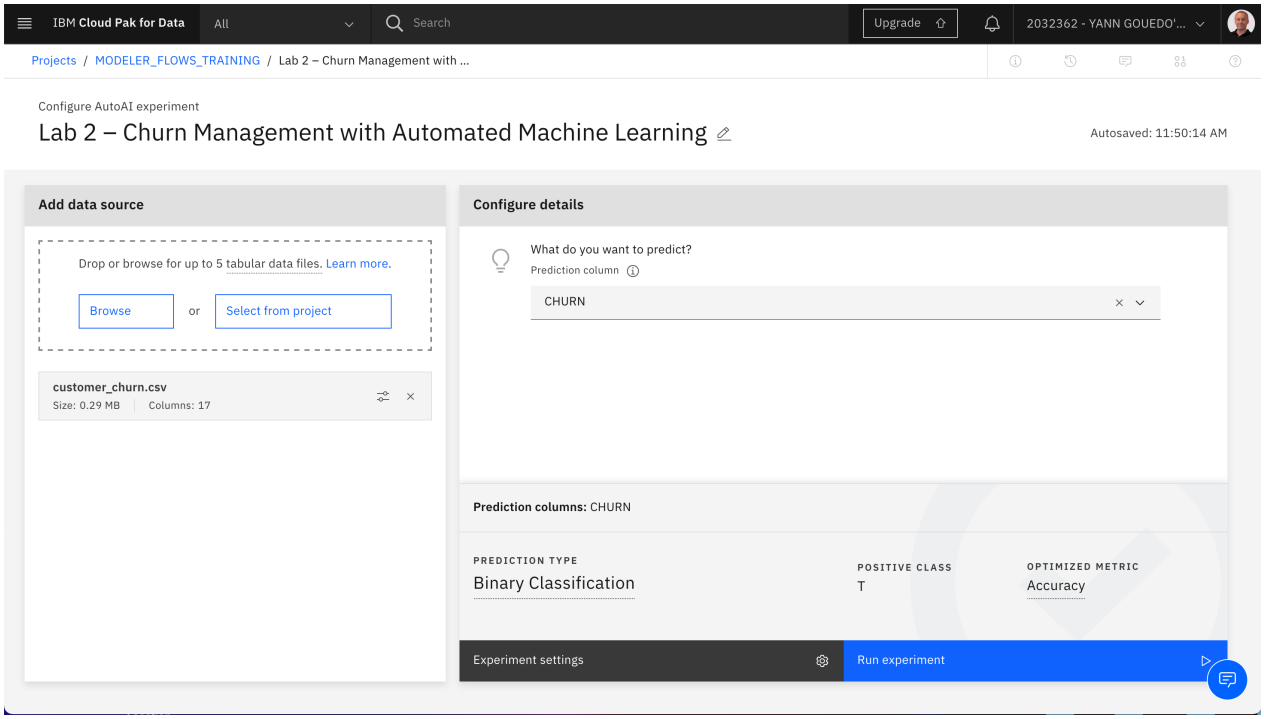


2 - **Select the *customer_churn.csv* Data Asset** already available in the project



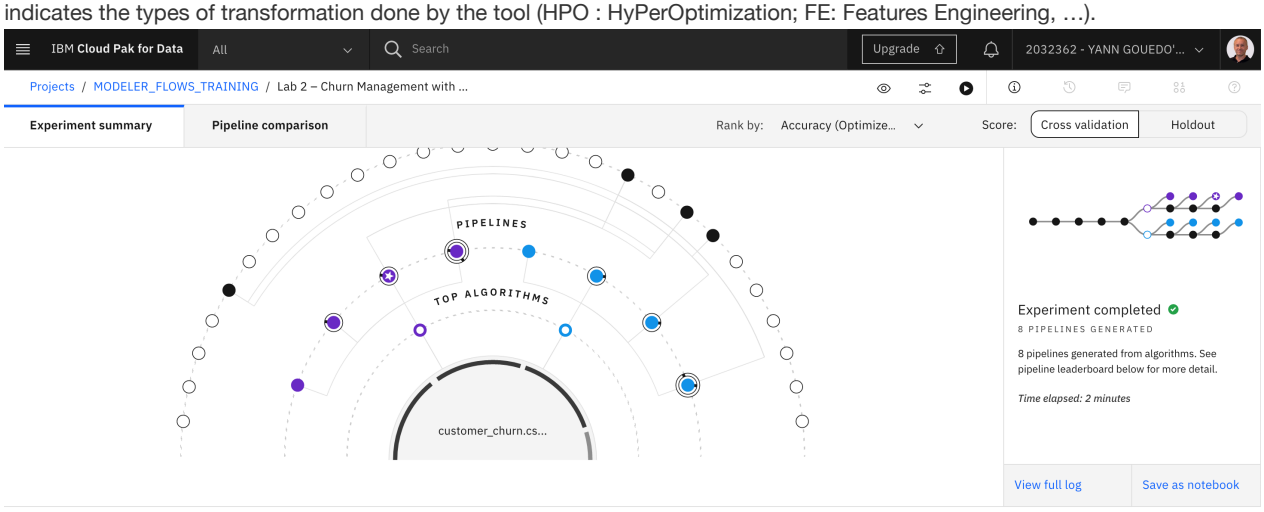
3 - **Select the *CHURN* variable as the field to predict. The objective is to predict churners and non churners, using the other variables available into the dataset. The prediction type is a binary classification, and the quality metric is the accuracy (based

on the confusion matrix). You could refine these parameters, if necessary.



4 - **Launch the experiment**, then wait for the calculations to be completed

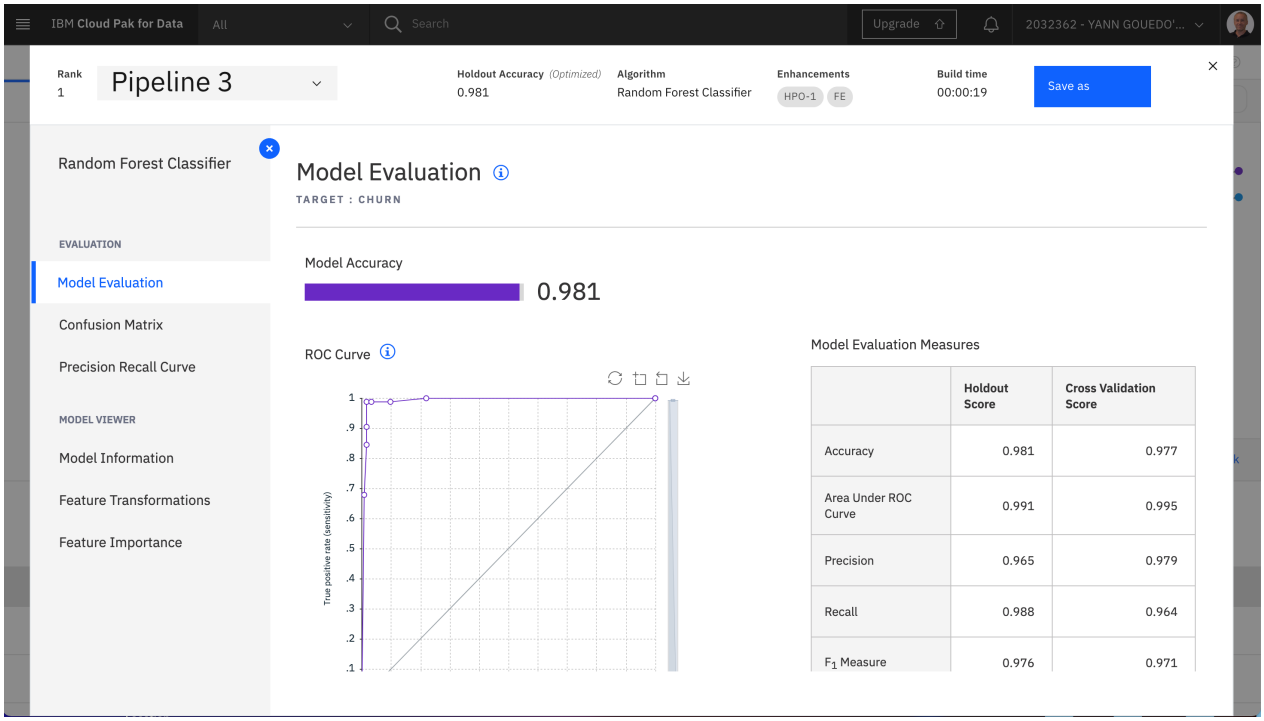
5 - **Explore the different models outcomes**, displayed in the windows outcome. These models ouctomes are automatically generated by the solution. The accuracy value helps to compare the different models together. The enhancements information indicates the types of transformation done by the tool (HPO : HyPerOptimization; FE: Features Engineering, …).



Pipeline leaderboard

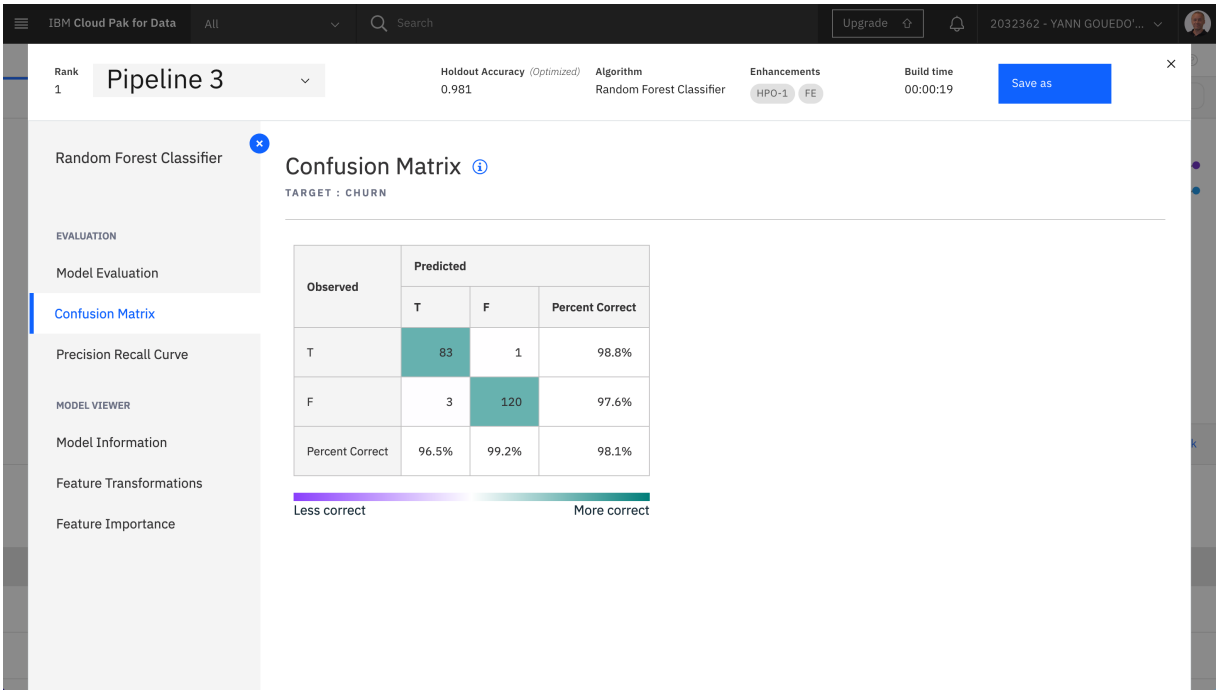
Rank	Name	Algorithm	Accuracy (Optimized)	Enhancements	Build time
★ 1	Pipeline 3	Random Forest Classifier	0.977	HPO-1 FE	00:00:19
2	Pipeline 4	Random Forest Classifier	0.977	HPO-1 FE HPO-2	00:00:18

6 - **Select the first model** in clicking on the model/algorithm name. It is the model with the best accuracy.

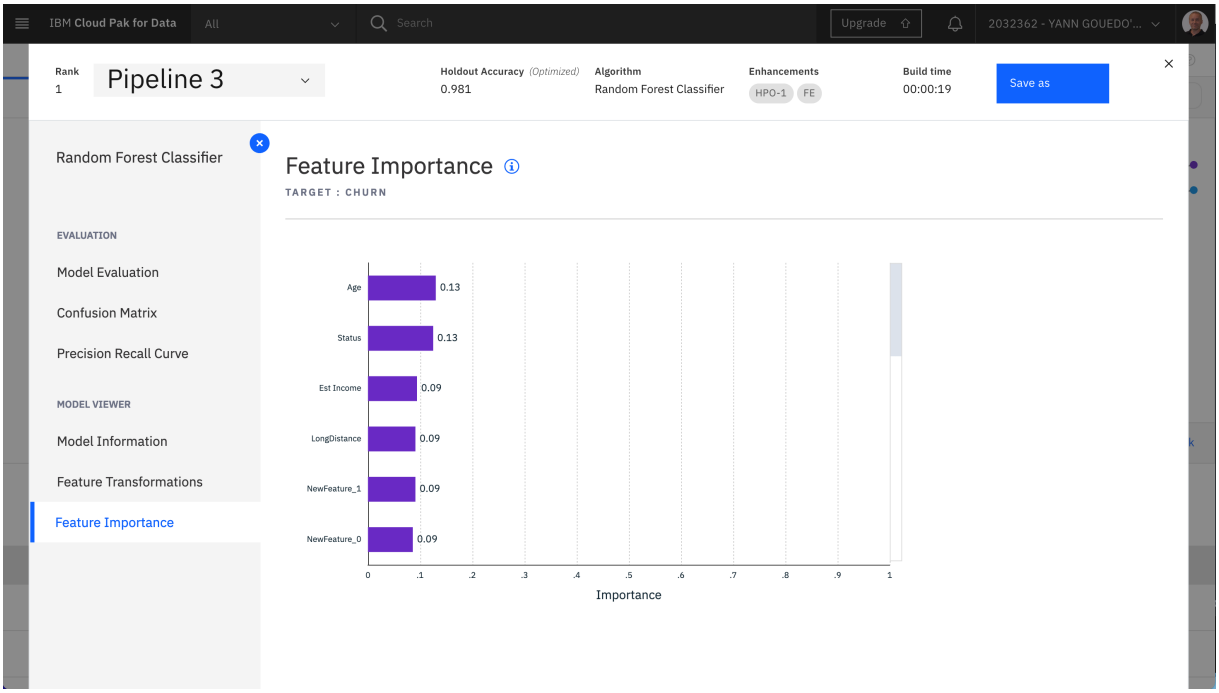


7 - **Navigate** into the different sections to explore the characteristics of this model, as the following

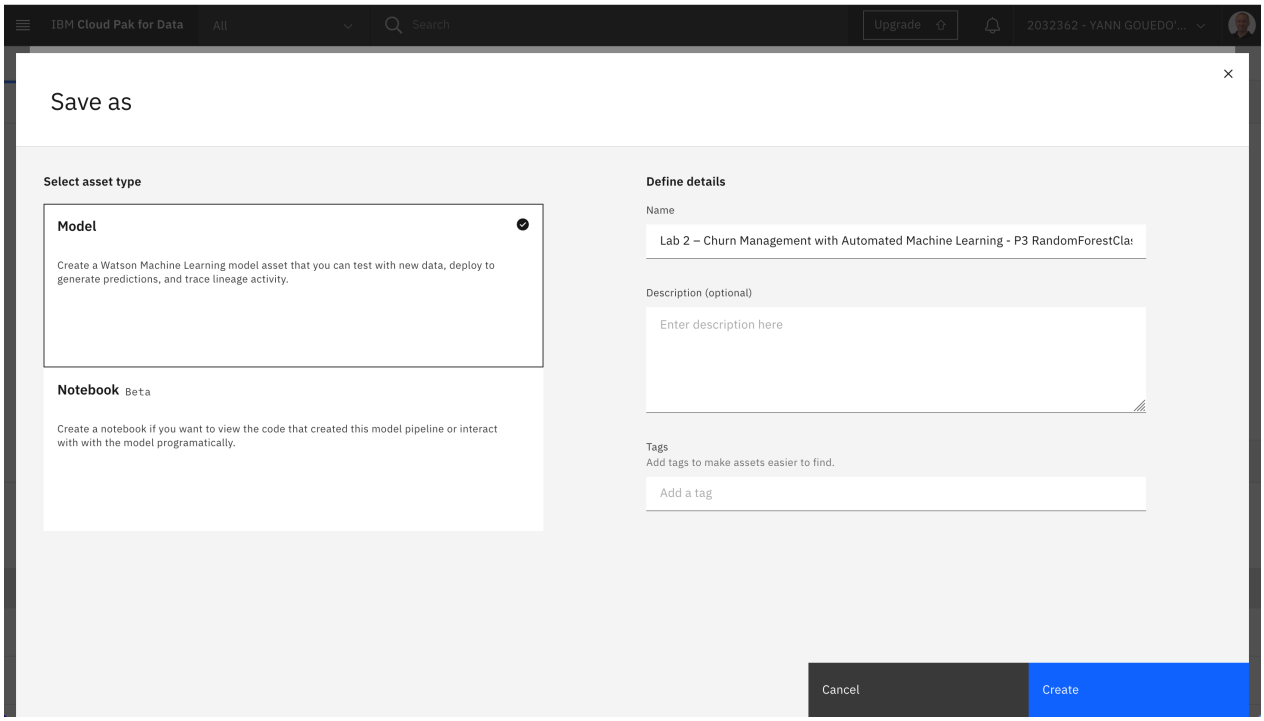
- Confusion matrix



- Feature importance



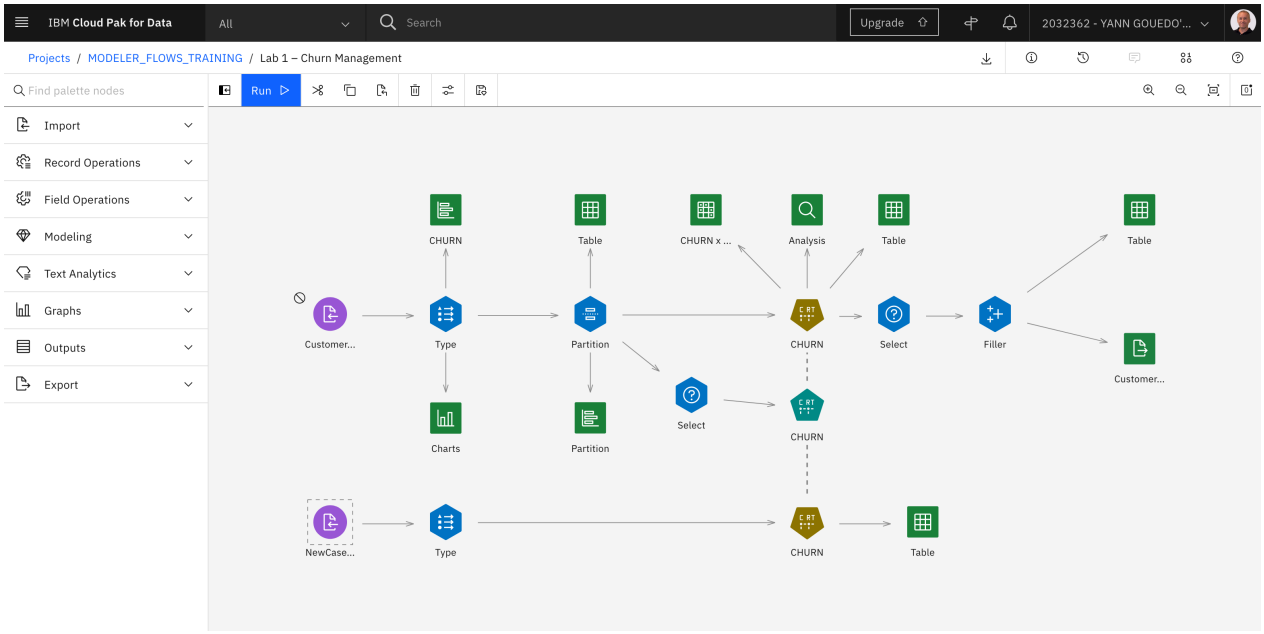
8 - **Click on the Save menu** to see the different options to prepare the deployment step.



Lab 3 – Deployment

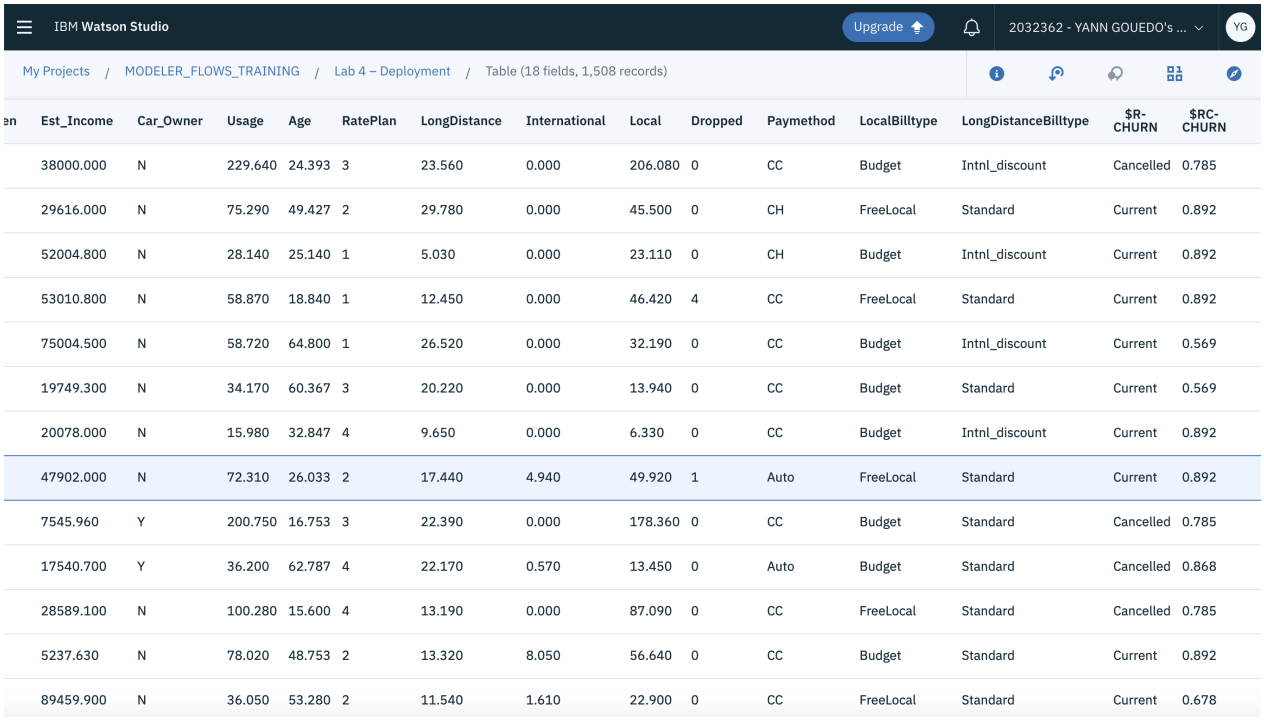
Industry context: telecommunication
Business objective: to deploy a churn model on a new customer list.
Technical objective: to use predefined deployment capabilities.
Data: customer descriptions, contracts, model outcome.

1 - Open the "Lab 1 - Churn Management" and **Modify the stream** adding a new branch below, using the *NewCases.csv* file as the following, with the objective to apply the Churn model node on new cases in order to identify if a customer will churn or not. Note that the historical *CHURN* information is no more present in the dataset.



Below the updates:

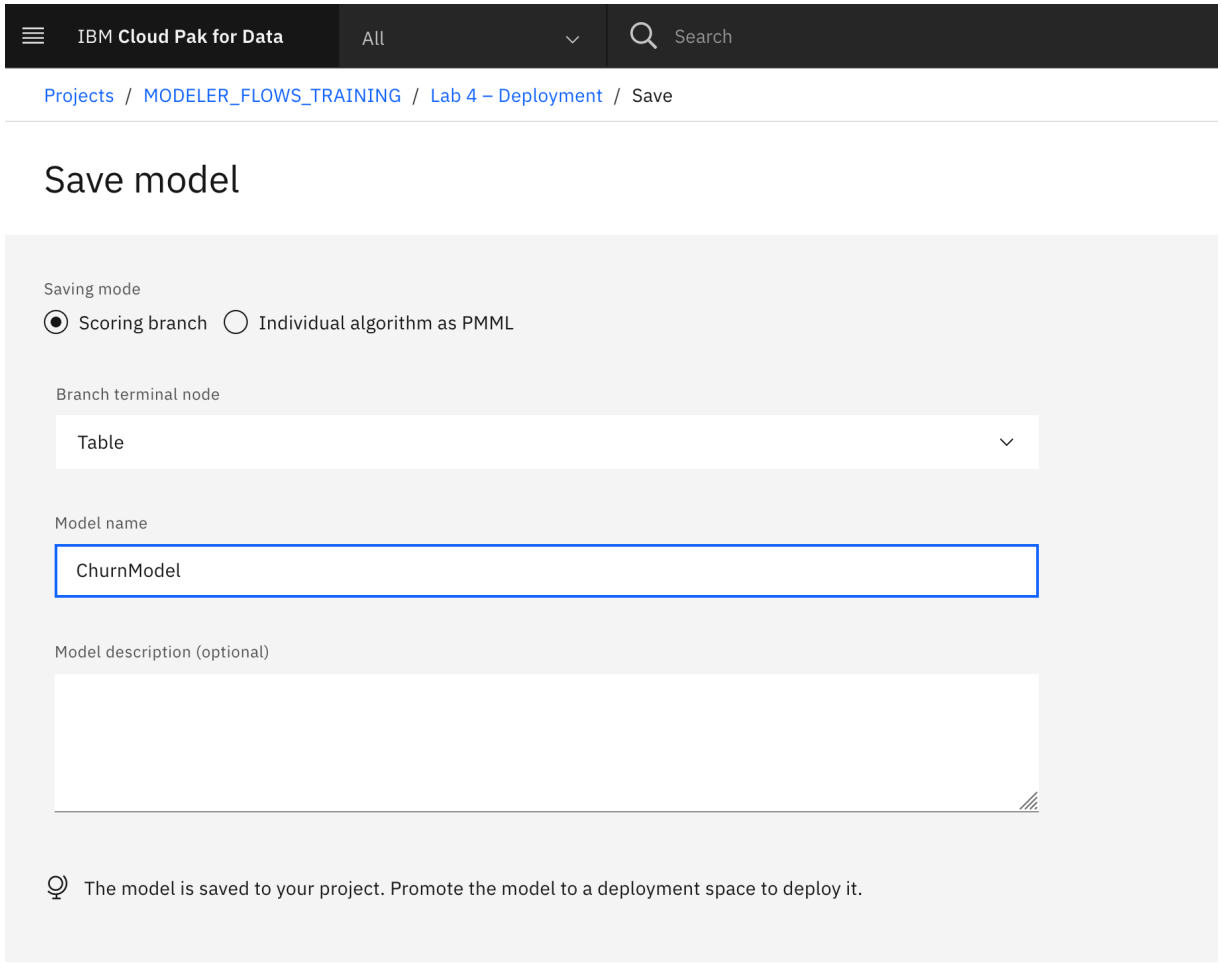
2 - **Display the predictive CHURN outputs**, in running the stream.



Est_Income	Car_Owner	Usage	Age	RatePlan	LongDistance	International	Local	Dropped	Paymethod	LocalBilltype	LongDistanceBilltype	$R-CHURN	$RC-CHURN
38000.000	N	229.640	24.393	3	23.560	0.000	206.080	0	CC	Budget	Intnl_discount	Cancelled	0.785
29616.000	N	75.290	49.427	2	29.780	0.000	45.500	0	CH	FreeLocal	Standard	Current	0.892
52004.800	N	28.140	25.140	1	5.030	0.000	23.110	0	CH	Budget	Intnl_discount	Current	0.892
53010.800	N	58.870	18.840	1	12.450	0.000	46.420	4	CC	FreeLocal	Standard	Current	0.892
75004.500	N	58.720	64.800	1	26.520	0.000	32.190	0	CC	Budget	Intnl_discount	Current	0.569
19749.300	N	34.170	60.367	3	20.220	0.000	13.940	0	CC	Budget	Standard	Current	0.569
20078.000	N	15.980	32.847	4	9.650	0.000	6.330	0	CC	Budget	Intnl_discount	Current	0.892
47902.000	N	72.310	26.033	2	17.440	4.940	49.920	1	Auto	FreeLocal	Standard	Current	0.892
7545.960	Y	200.750	16.753	3	22.390	0.000	178.360	0	CC	Budget	Standard	Cancelled	0.785
17540.700	Y	36.200	62.787	4	22.170	0.570	13.450	0	Auto	Budget	Standard	Cancelled	0.868
28589.100	N	100.280	15.600	4	13.190	0.000	87.090	0	CC	FreeLocal	Standard	Cancelled	0.785
5237.630	N	78.020	48.753	2	13.320	8.050	56.640	0	CC	Budget	Standard	Current	0.892
89459.900	N	36.050	53.280	2	11.540	1.610	22.900	0	CC	FreeLocal	Standard	Current	0.678

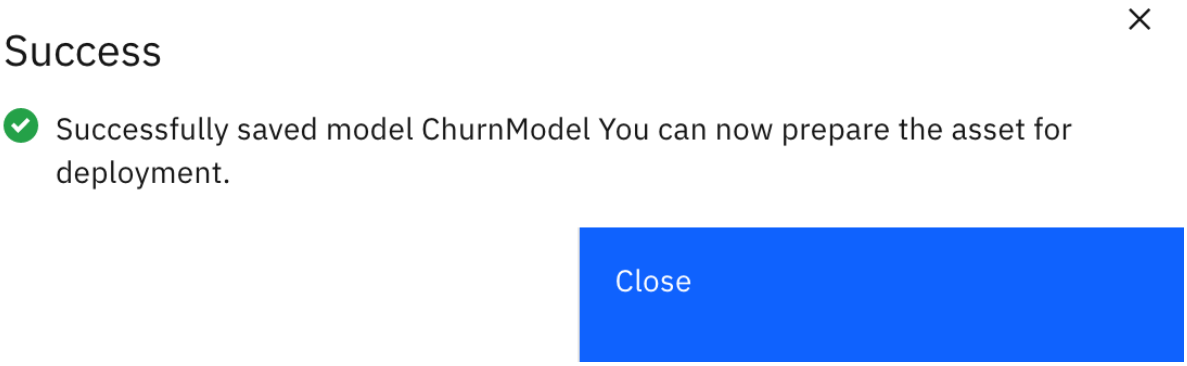
3 - **Save the model**:

- Click right on the Table node from the context menu to enable the stream for deployment, then select Save branch as a model. The following windows appears:

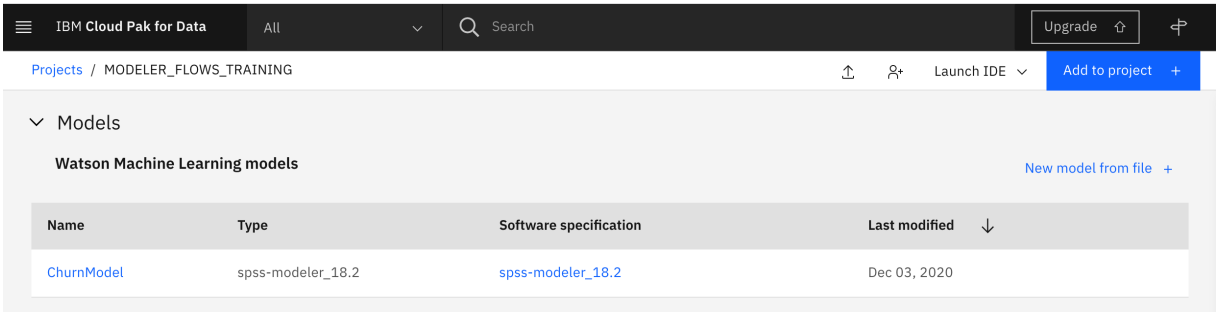


- Define a Model name

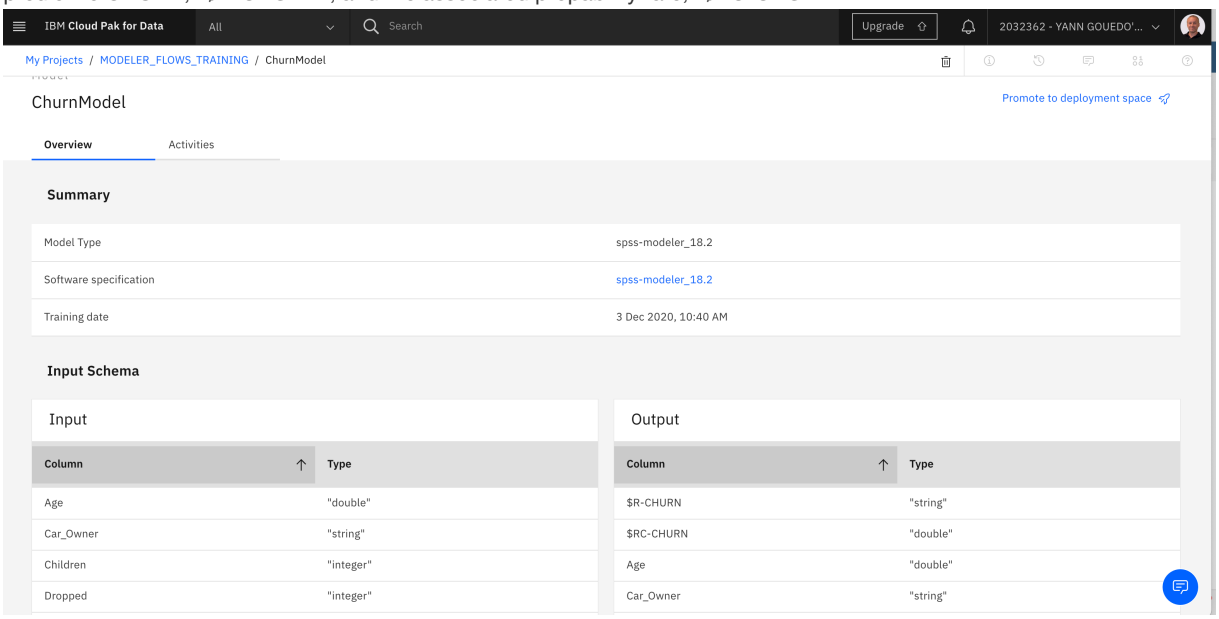
- Select a Machine Learning service or create a new one.
- Save it.



- Go back to the Assets menu, then to the Model section.



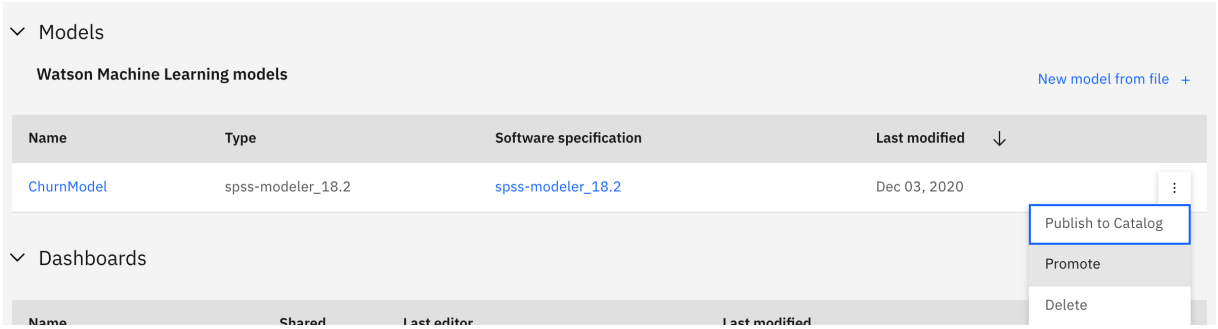
- Click on the model to see its characteristics as below. Note that there is not the *CHURN* variable in the "Input" data (because of this is what we want to obtain from the model), and there are two new columns in the "Output" section: the predictive CHURN, "*$R-CHURN*", and the associated propability rate, "*$RC-CHURN*"



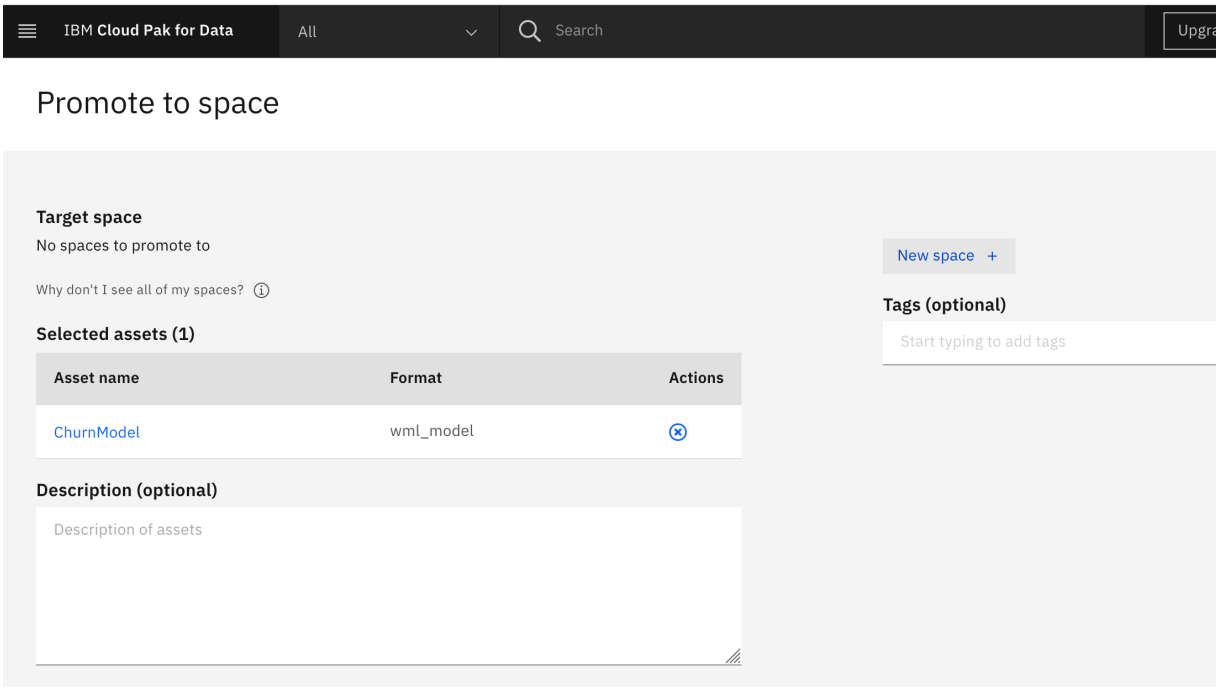
- Go back to the Model section.

4 - **Promote the Churn model to the deployment space**

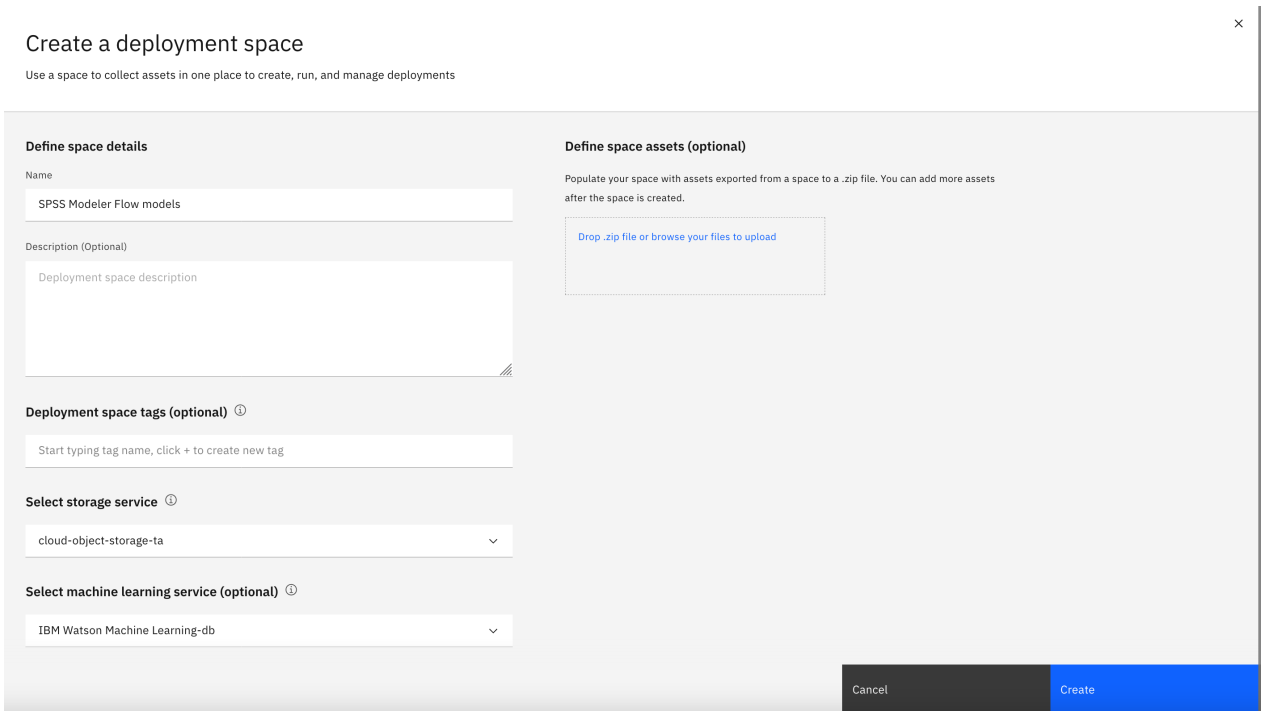
- Go back to the Assets menu, then select Promote in clicking on the 3 dots.

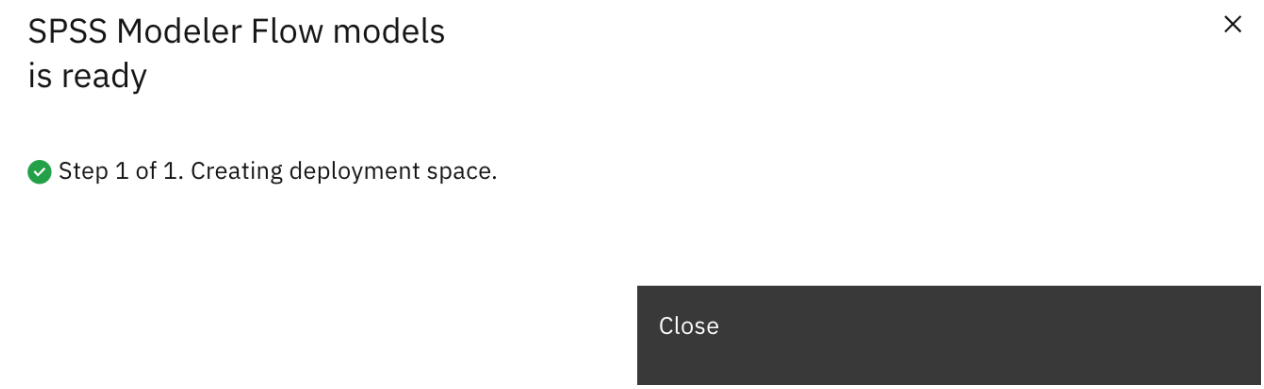


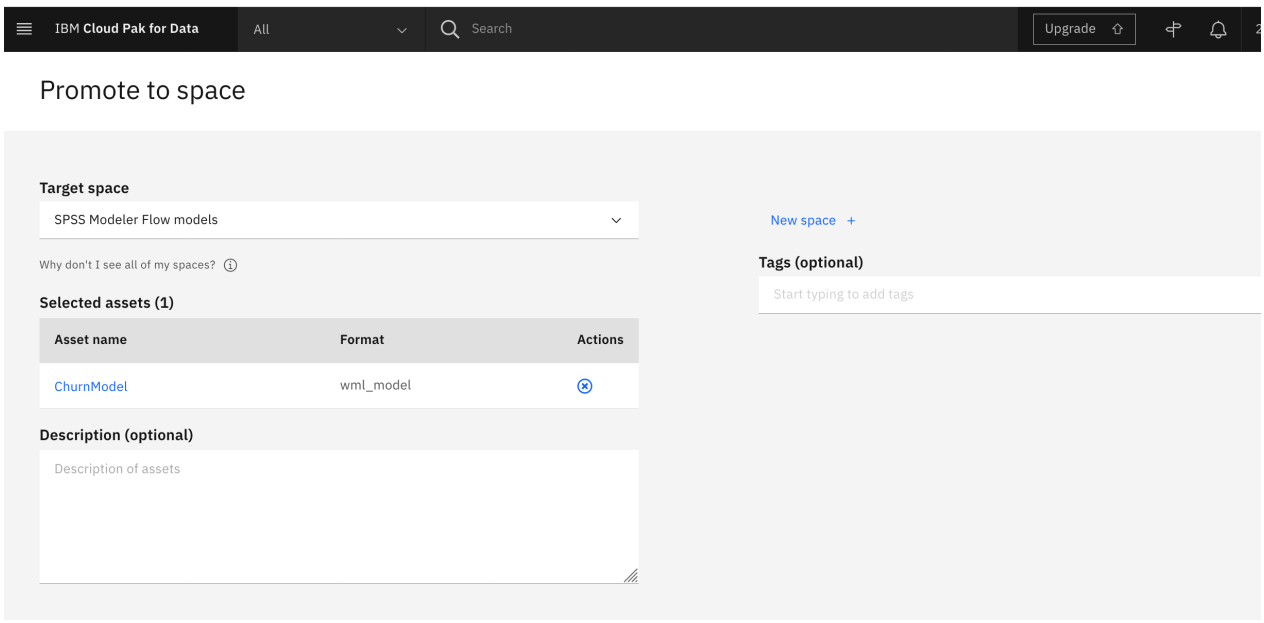
- Click on New space menu, then define a space



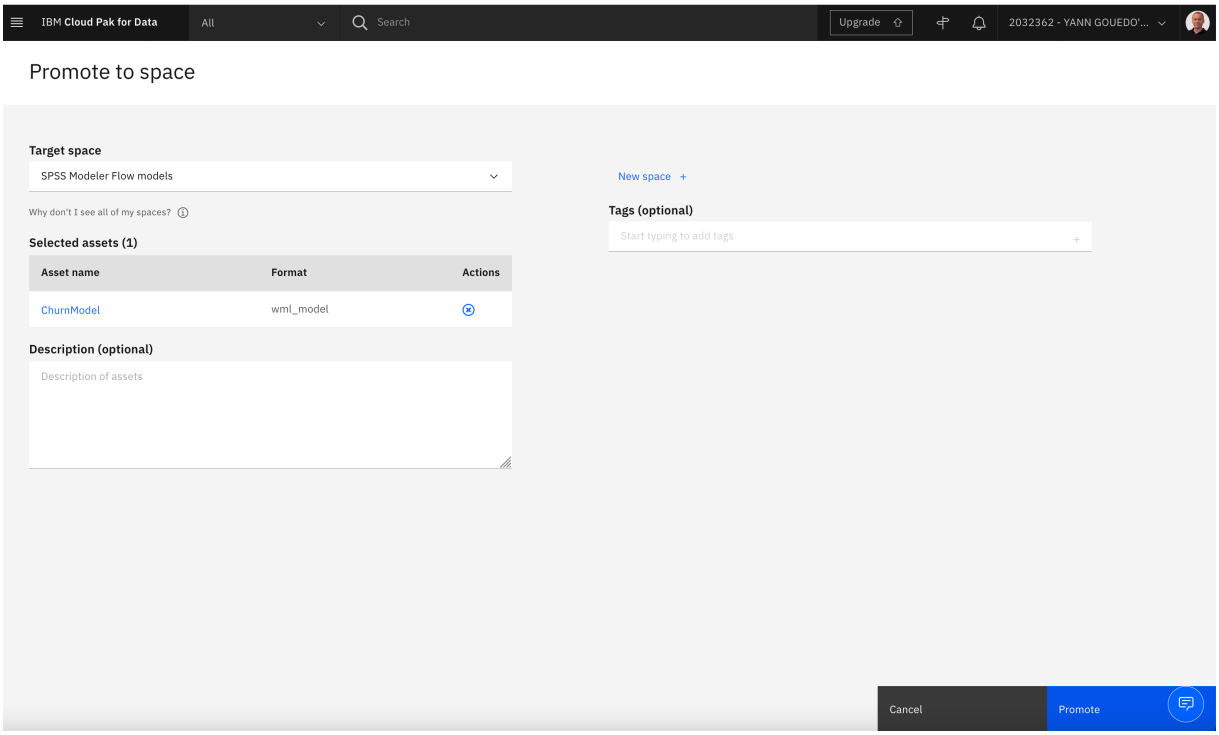
- Define a space name, select a cloud object storage, and a machine learning service (create it if you if you don't already have one), then create the space


Create a deployment space
Use a space to collect assets in one place to create, run, and manage deployments
Define space details
Name
SPSS Modeler Flow models
Description (Optional)
Deployment space description
Deployment space tags (optional)
Start typing tag name, click + to create new tag
Select storage service
cloud-object-storage-ta
Select machine learning service (optional)
IBM Watson Machine Learning-db
Define space assets (optional)
Populate your space with assets exported from a space to a .zip file. You can add more assets after the space is created.
Drop .zip file or browse your files to upload
Cancel
Create


SPSS Modeler Flow models is ready
Step 1 of 1. Creating deployment space.
Close

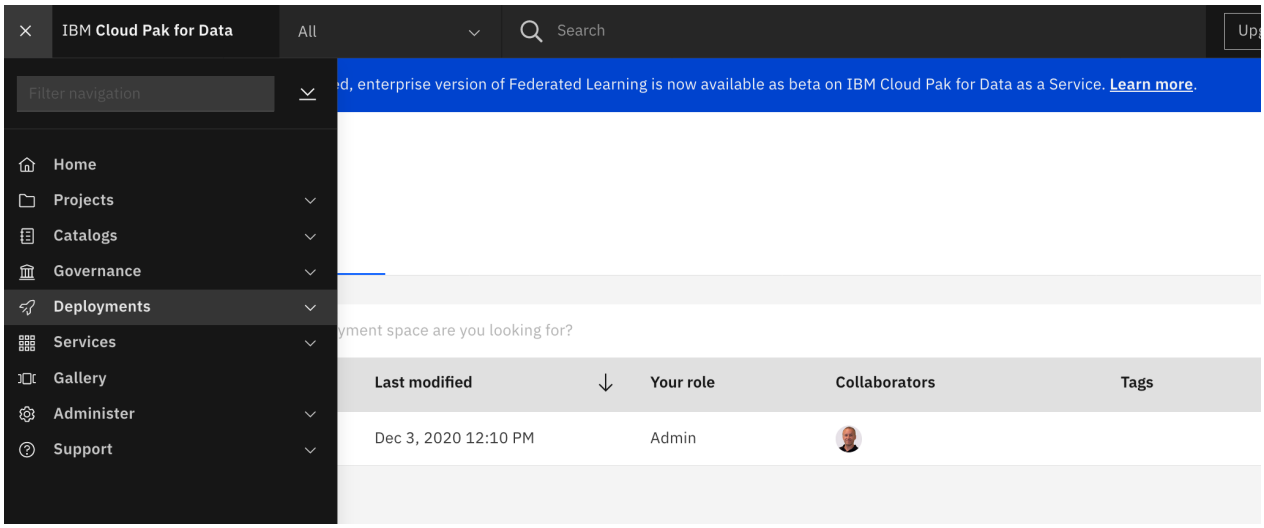

IBM Cloud Pak for Data
All
Search
Upgrade
Promote to space
Target space
SPSS Modeler Flow models
New space +
Why don't I see all of my spaces?
Tags (optional)
Start typing to add tags
Selected assets (1)
Asset name
Format
Actions
ChurnModel
wml_model
Description (optional)
Description of assets

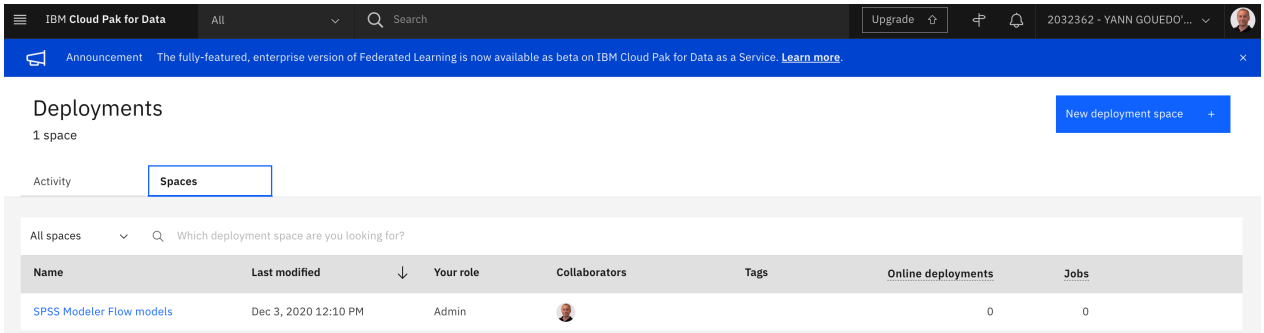
- The target space is automatically associated, then promote the **ChurnModel**

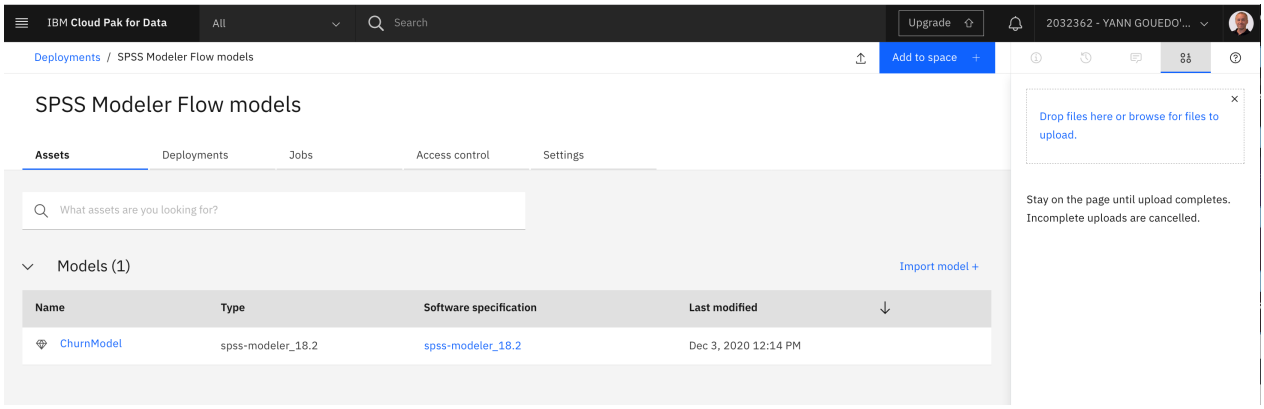


5 - **Prepare the deployment**

- Duplicate the existing Watson Studio internet page to another tab, in order to access to the Watson Studio deployment space while keeping the Watson Studio project page.



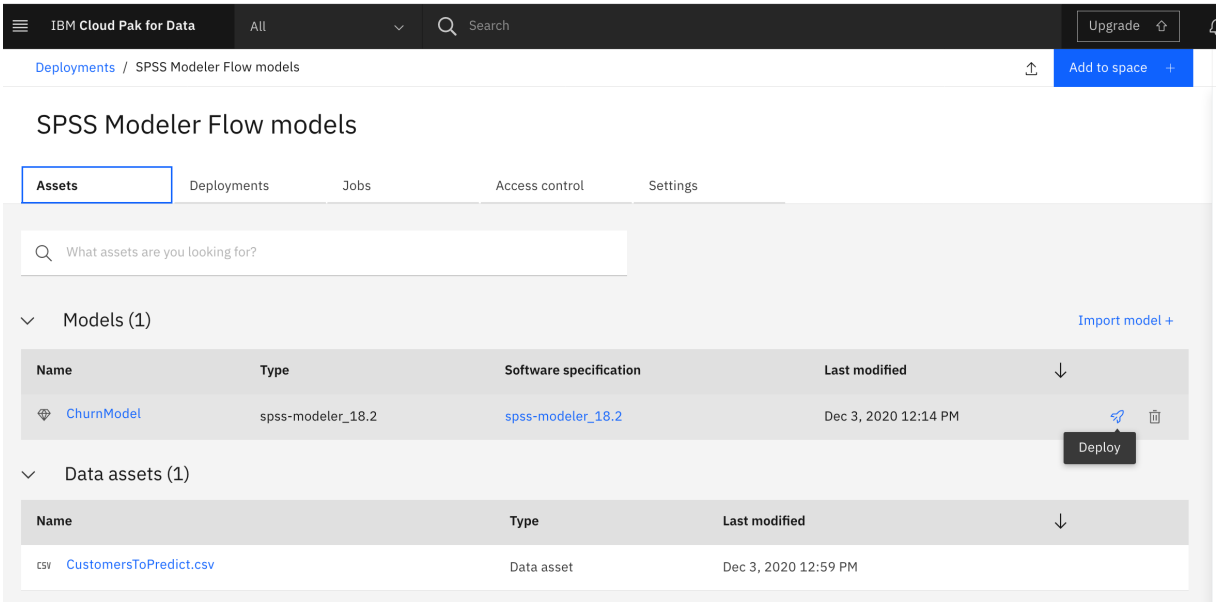




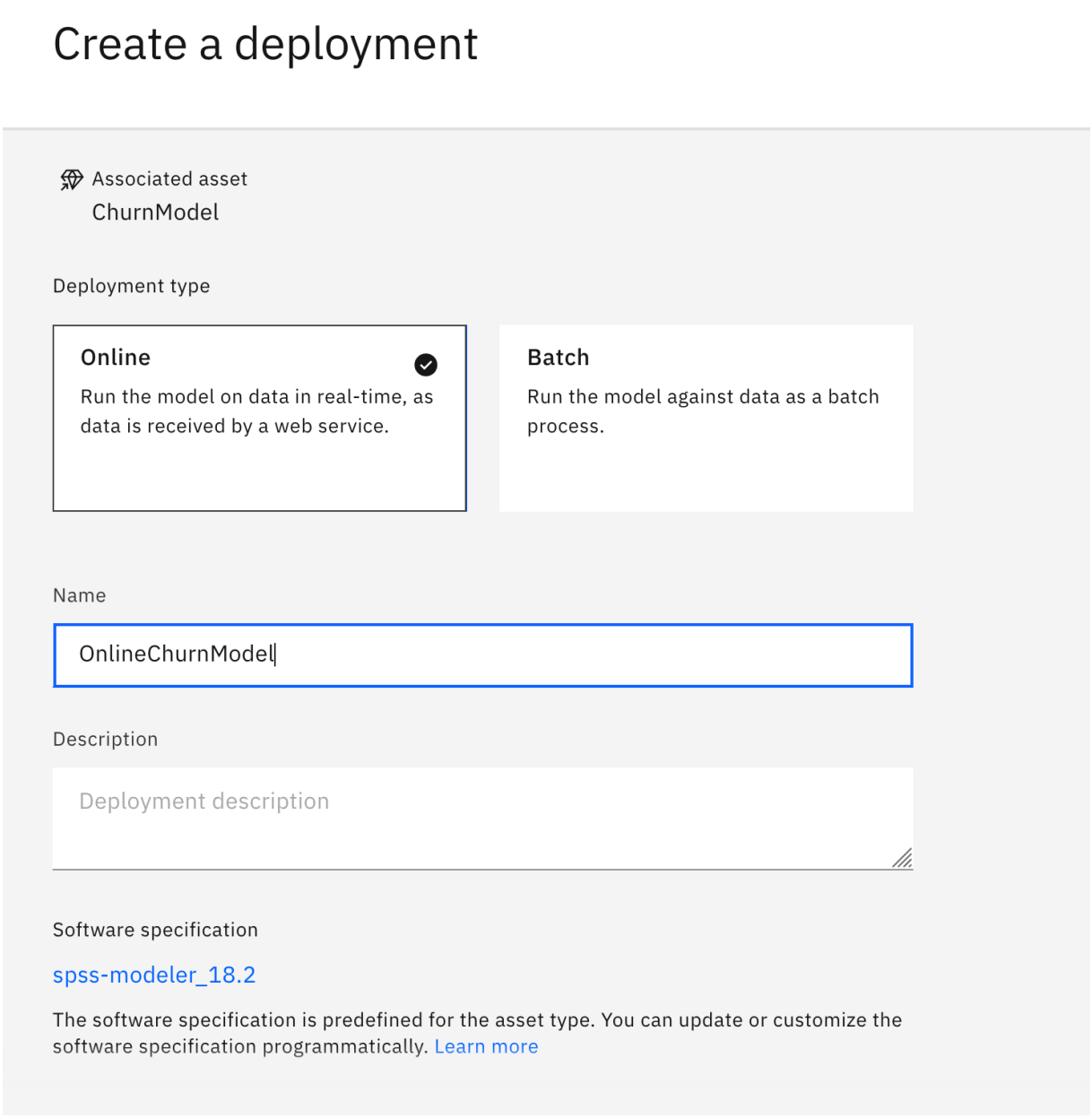
- Upload the file which contains the list of customers for which you want to know if they present a risk to churn, "*CustomersToPredict.csv*"

6 - **Use the deployment functions**

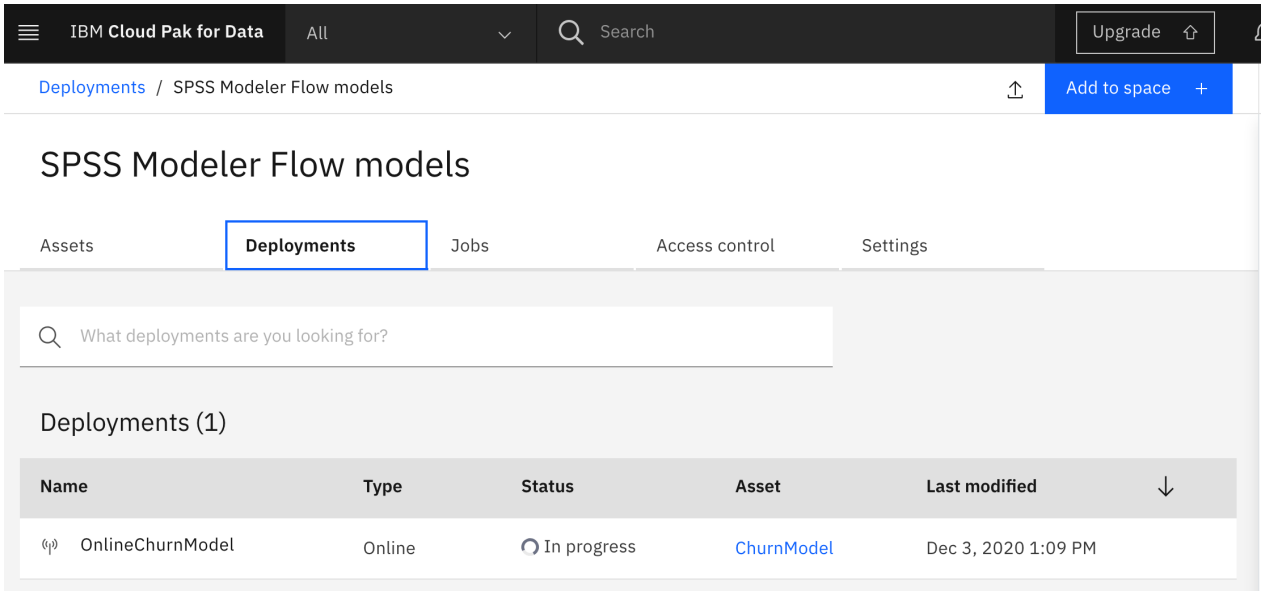
- Deploy the *ChurnModel*

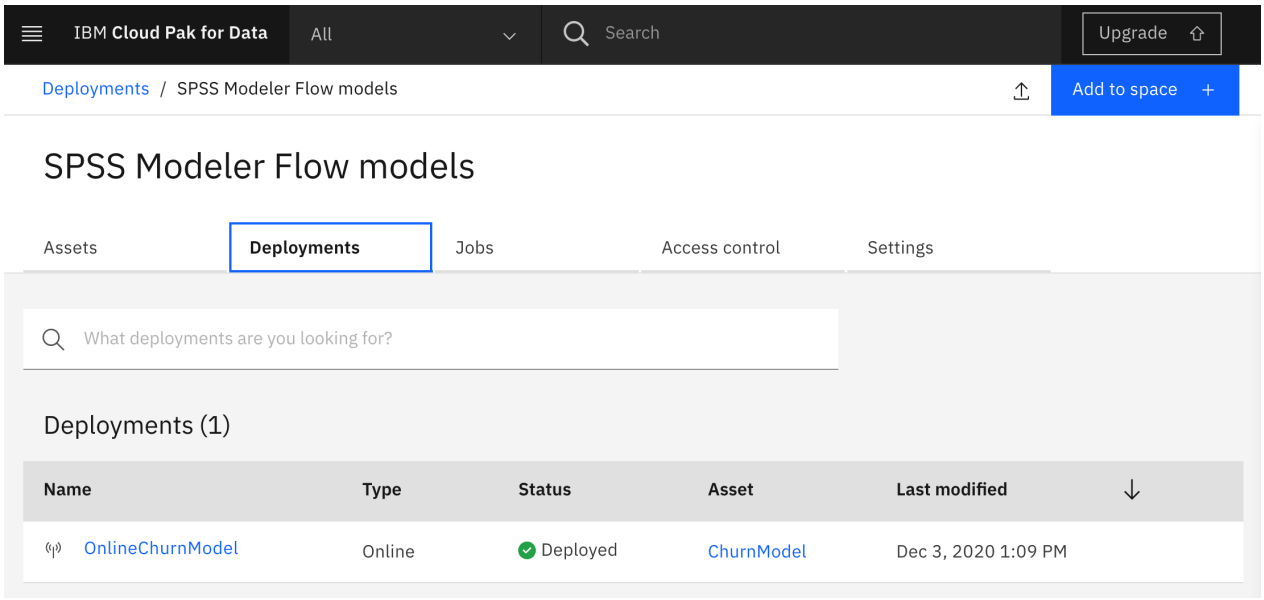


- Select the type of deployment, online or batch, choose the "online" option



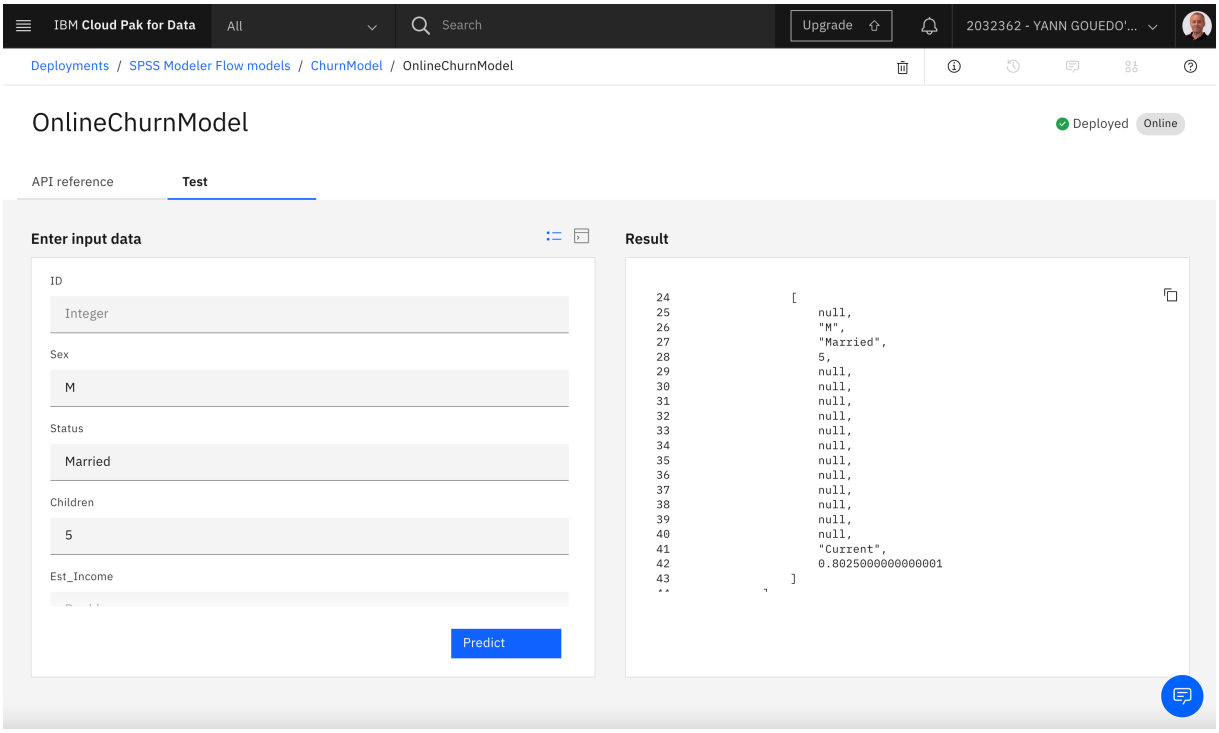
- Go to the "Deployments" tab, you will see the status "In progress", then "Deployed"





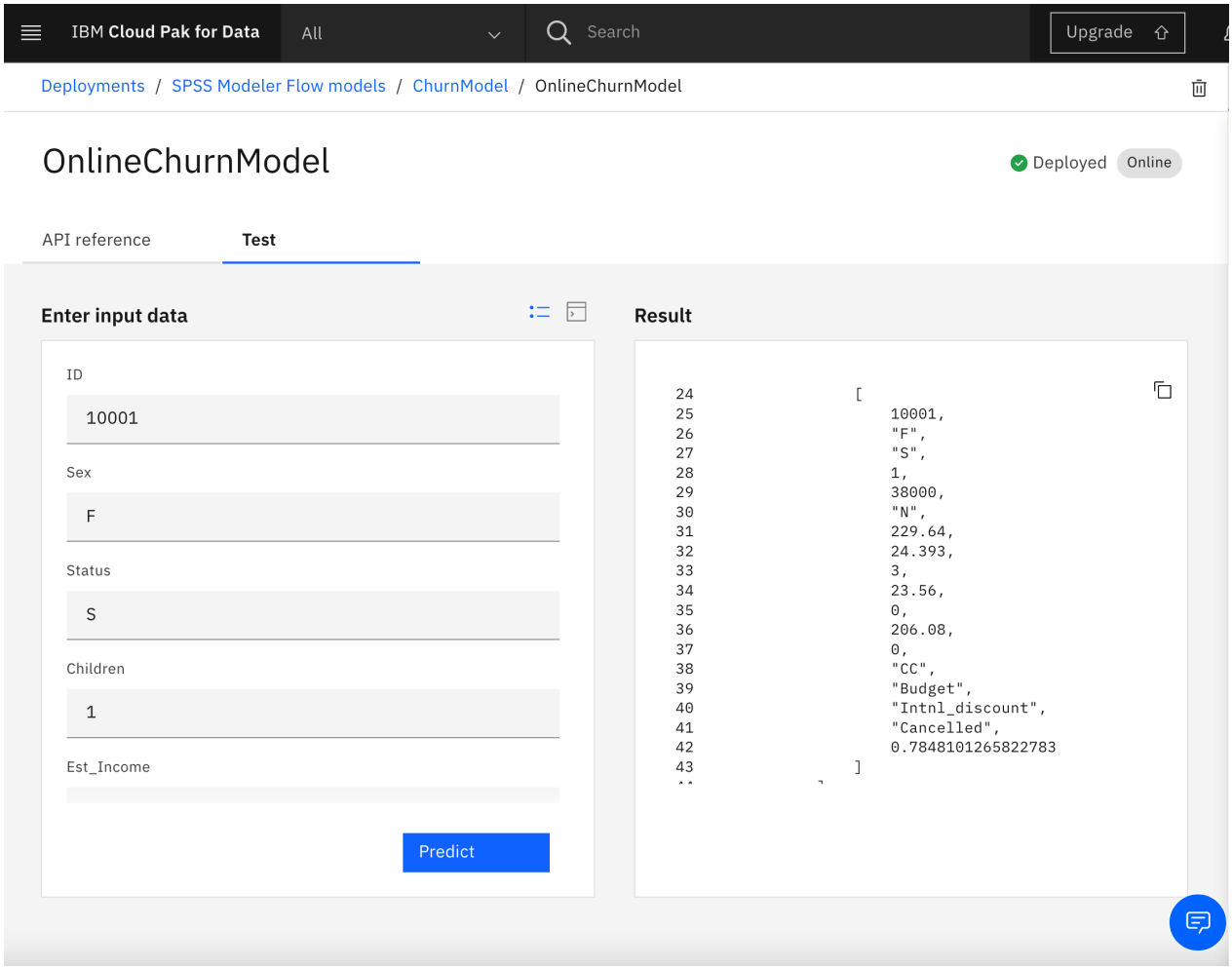
- Click on the "OnlineChurnModel", you will see that some codes (in - Click on the "OnlineChurnModel", you will see codes lines (written in different languages), useful in case you want to use this online Churn model in any external operational process. Clicking on the "Test" tab, you will be able to predict the Churn, and its probability rate, at the customer level, or at higher levels, filling the modalities of one or several variables. Try different tests.

- First example: Male customers, married, and with 3 children will not churn with a probability rate = 80,25%



- Second example: The customer with ID = "10001", will churn with a probability rate = 78,48%. You must fill each attributes for the customer ID = "10001", as below.

ID	Sex	Status	Children	Est_Income	Car_Owner	Usage	Age	RatePlan	LongDistance	International	Local	Dropped	Paymethod	LocalBilltype	LongDistanceBilltype
10001	F	S	1	38000.000	N	229.640	24.393	3	23.560	0.000	206.080	0	CC	Budget	Intnl_discount


IBM Cloud Pak for Data
All
Search
Upgrade
Deployments / SPSS Modeler Flow models / ChurnModel / OnlineChurnModel
OnlineChurnModel
Deployed
Online
API reference
Test
Enter input data
ID
10001
Sex
F
Status
S
Children
1
Est_Income
Predict
Result
[
10001,
"F",
"S",
1,
38000,
"N",
229.64,
24.393,
3,
23.56,
0,
206.08,
0,
"CC",
"Budget",
"Intnl_discount",
"Cancelled",
0.7848101265822783
]

Lab 4 - Market Basket Analysis

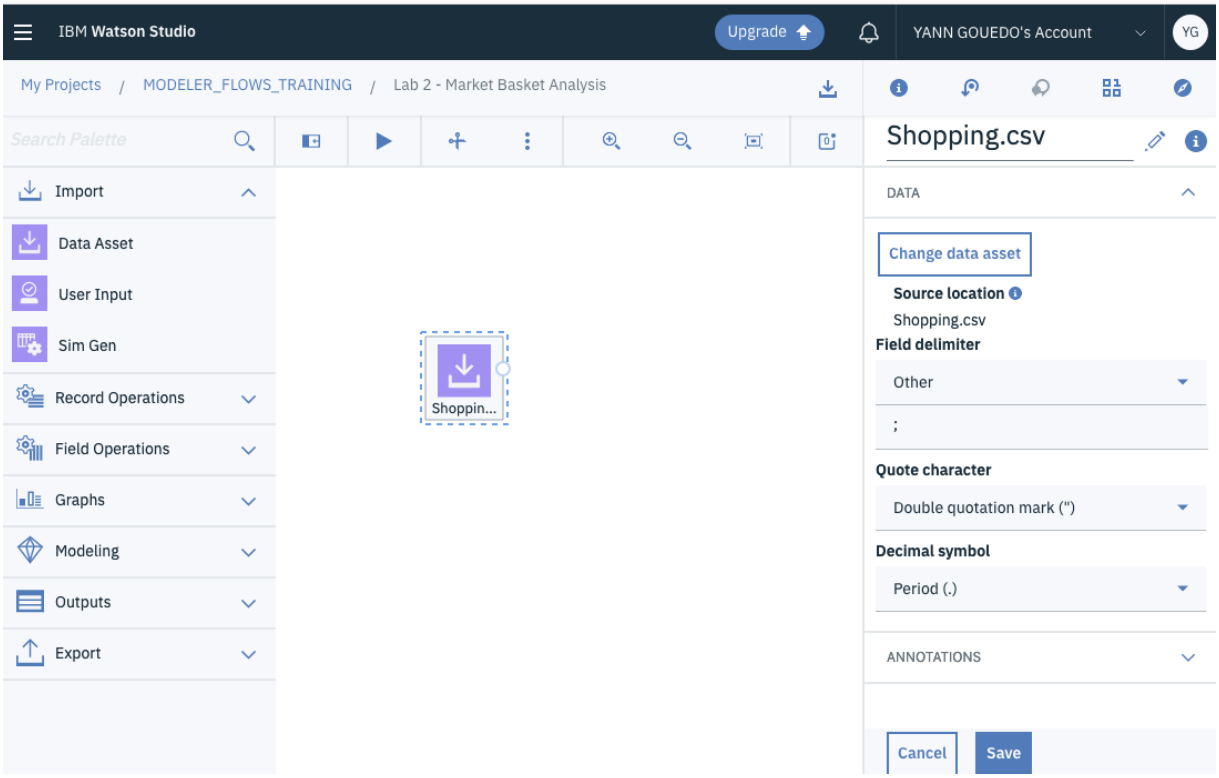
Industry context: retail
Business objective: to identify the affinity between products, building the Market Basket Analysis which is a technique which identifies the strength of association between pairs of products purchased together and identify patterns of co-occurrence. … It creates If-Then scenario rules, for example, if item A is purchased then item B is likely to be purchased.
Technical objective: to use association techniques through a visual programming approach.
Data: customer data: customer data, sales transactions - products bought

1 - **Create a new Modeler Flow**, named it Lab 2 - Market Basket Analysis.

2 - From the Import palette, **add a Data Asset node**.

3 - **Define the data asset** to use:

- Double-click on the Data Asset node on the canvas.
- Select *Shopping.csv* clicking on Change data asset / Data assets.
- Ensure to have the following parameters:



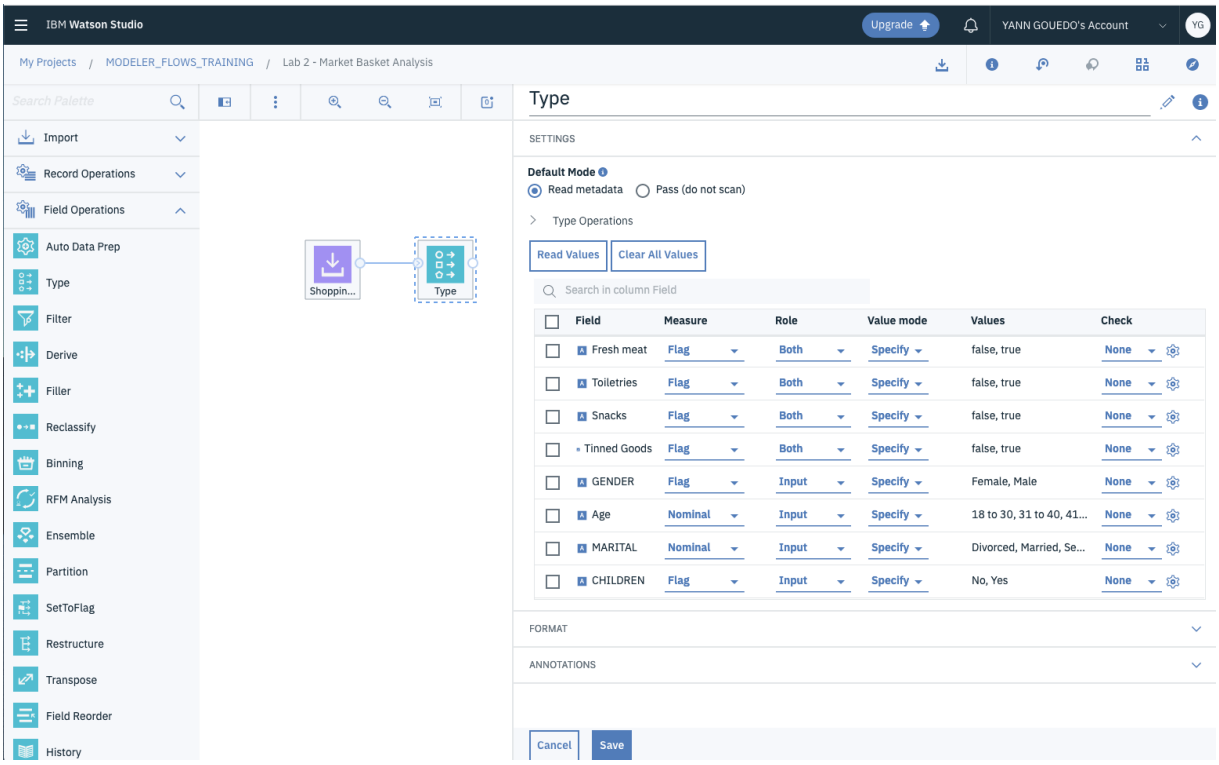
- Save the modification.

4 - Right click on the *Shopping.csv* file node and select Preview from the menu, **to preview the data**. This is an extraction of sales transactions data from a retail company's and the dataset includes products ID, and few customer information like gender, age, children, marital status.

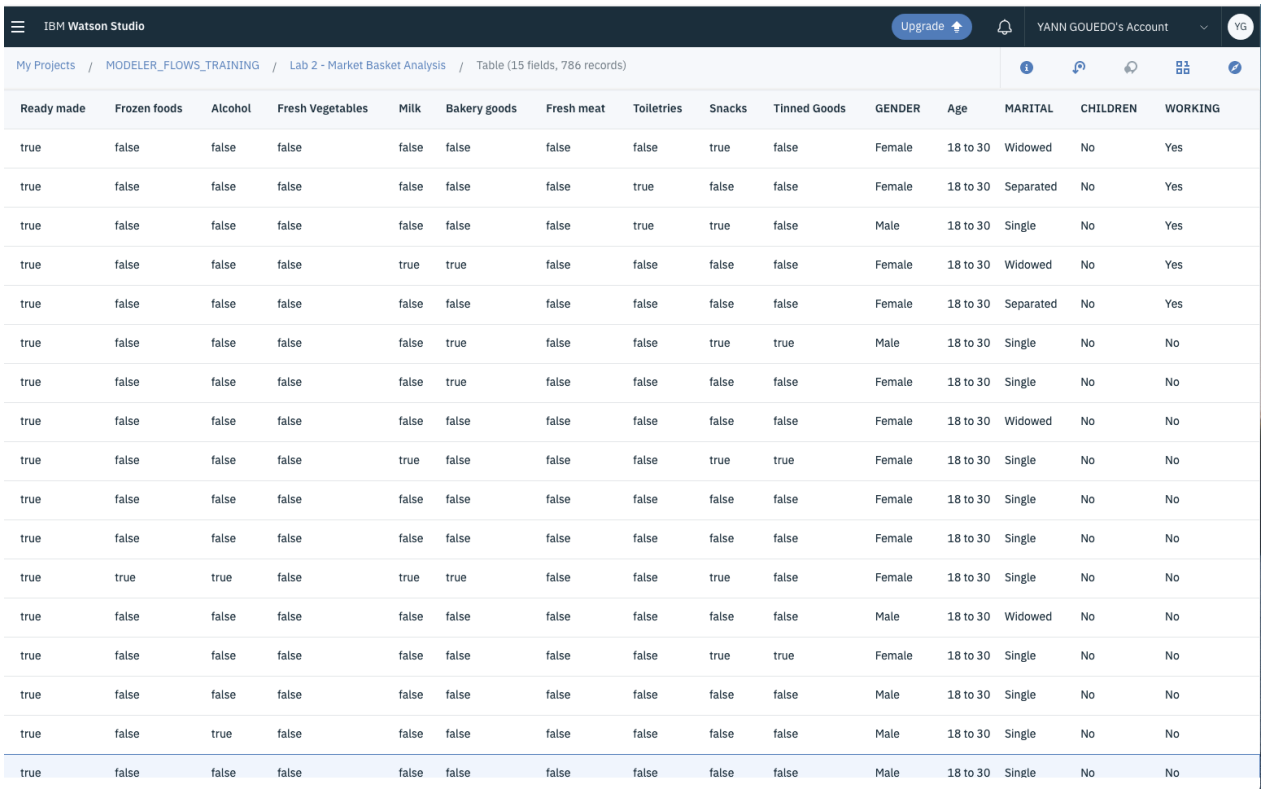
Ready made	Frozen foods	Alcohol	Fresh Vegetables	Milk	Bakery goods	Fresh meat	Toile
true	false	false	false	false	false	false	false
true	false	false	false	false	false	false	true
true	false	false	false	false	false	false	true
true	false	false	false	true	true	false	false
true	false	false	false	false	false	false	false
true	false	false	false	false	true	false	false

5 - **Define the roles of the products fields**:

- From the Field Operations palette, add a Type node to the canvas and connect it to the data source.
- Open the Type node, clicking twice, modify the Role of the 10 products fields to "Both" and their measure to "Flag", then save the modification. The goal is to understand the relationship between the items being purchased, which are inputs and the targets.
- Click on the Read Values button to scan the data as well as to display and update the range of values.



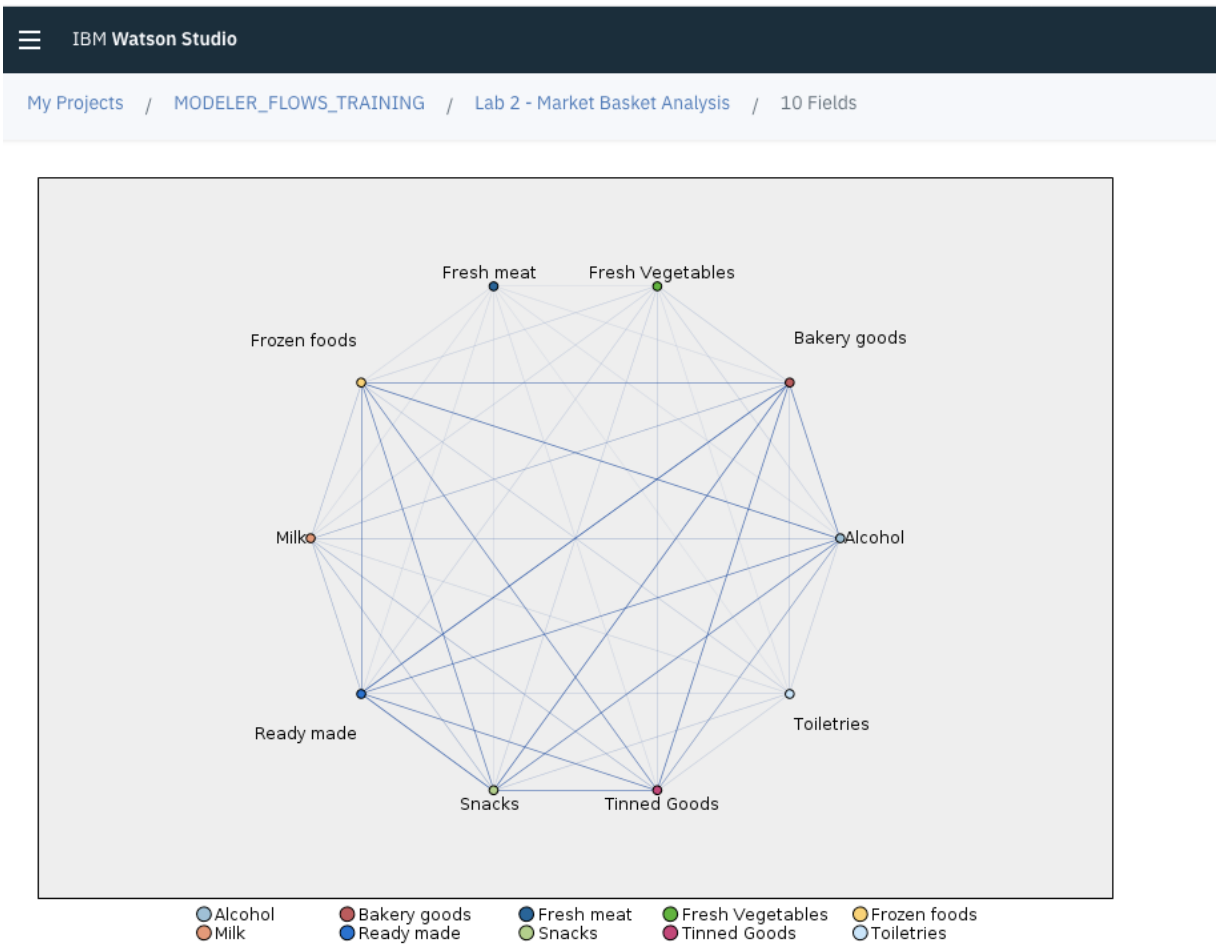
6 - From the Outputs palette, **add a Table node** to the canvas and connect it to the data source.
Display the data. Note that the data are sales transactions with some demographics about the buyer. The first 10 fields are binary where 1 indicates the product was purchased and 0 indicates it was not purchased.



Ready made	Frozen foods	Alcohol	Fresh Vegetables	Milk	Bakery goods	Fresh meat	Toiletries	Snacks	Tinned Goods	GENDER	Age	MARITAL	CHILDREN	WORKING
true	false	false	false	false	false	false	false	true	false	Female	18 to 30	Widowed	No	Yes
true	false	false	false	false	false	false	true	false	false	Female	18 to 30	Separated	No	Yes
true	false	false	false	false	false	false	true	true	false	Male	18 to 30	Single	No	Yes
true	false	false	false	true	true	false	false	false	false	Female	18 to 30	Widowed	No	Yes
true	false	false	false	false	false	false	false	false	false	Female	18 to 30	Separated	No	Yes
true	false	false	false	false	true	false	false	true	true	Male	18 to 30	Single	No	No
true	false	false	false	false	true	false	false	false	false	Female	18 to 30	Single	No	No
true	false	false	false	false	false	false	false	false	false	Female	18 to 30	Widowed	No	No
true	false	false	false	true	false	false	false	true	true	Female	18 to 30	Single	No	No
true	false	false	false	false	false	false	false	false	false	Female	18 to 30	Single	No	No
true	false	false	false	false	false	false	false	false	false	Female	18 to 30	Single	No	No
true	true	true	false	true	true	false	false	true	false	Female	18 to 30	Single	No	No
true	false	false	false	false	false	false	false	false	false	Male	18 to 30	Widowed	No	No
true	false	false	false	false	false	false	false	true	true	Female	18 to 30	Single	No	No
true	false	false	false	false	false	false	false	false	false	Male	18 to 30	Single	No	No
true	false	true	false	false	false	false	false	false	false	Male	18 to 30	Single	No	No
true	false	false	false	false	false	false	false	false	false	Male	18 to 30	Single	No	No

7 - **Usage of the Web node to analyze the link between the products bought**:

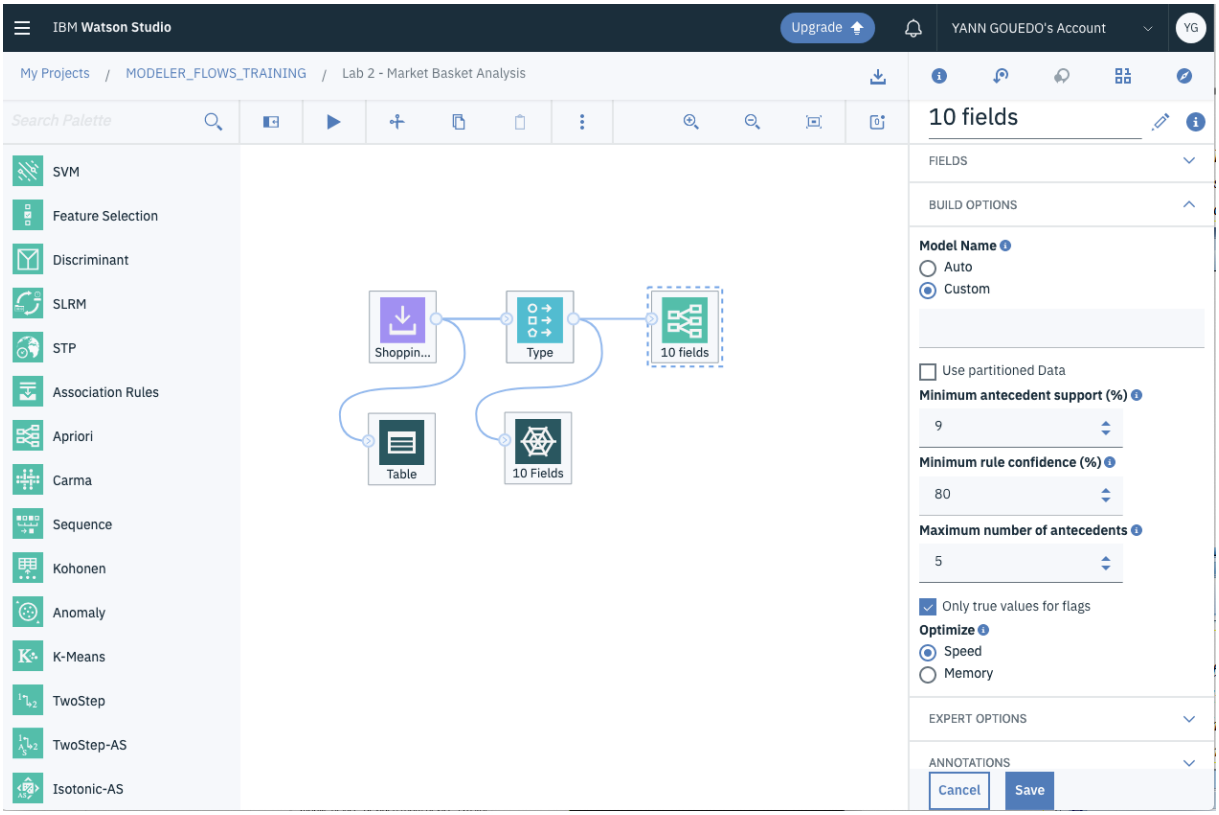
- From the Graphs palette, add a Web node to the canvas and connect it to the Type node.
- Change the Web node properties as the following:
    - Add the 10 first flag fields from *Ready Made* to *Tinned Goods* and click OK
    - Tick the box *Show true flag only*
    - Save it
    - Run the Web node. The output shows 10 items around the outside points of the graph. The lines connecting the points show the number of times the two items occurred in the same transaction.



8 - After reviewing the visualization, it is time to **build a model to understand these relationships and to make predictions as to which item is the next likely purchase**.

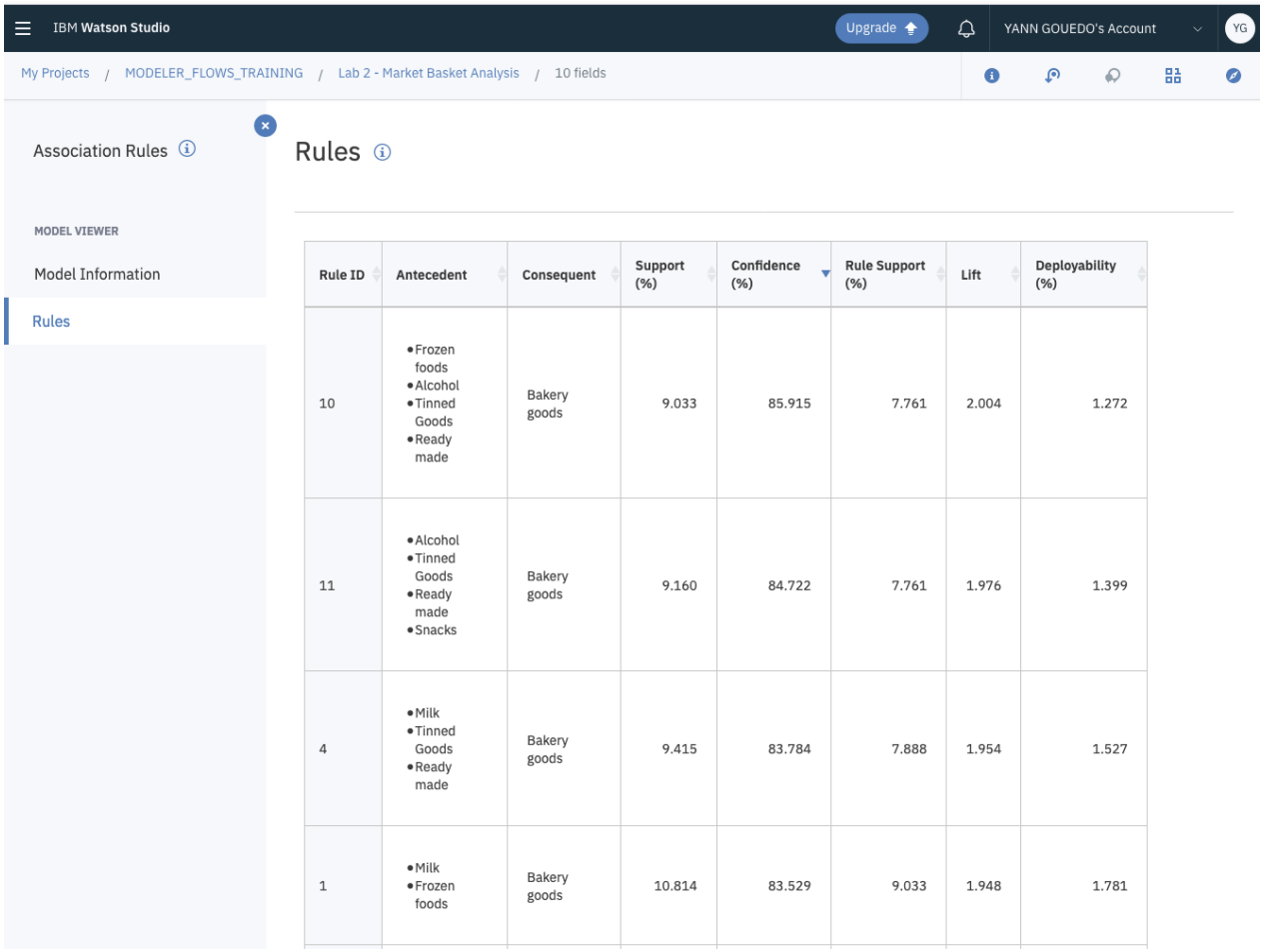
- From the Modeling palette, add an Apriori node to the Type node.
- Edit the Apriori node, useful to adjust the settings that control how rules are built.
- The antecedents are the “If” parts of the rule, e.g., If Milk and Frozen Foods, then Bakery Goods. Change the Minimum antecedent support (%) to 9.0.
  This means that only rules whose antecedents occur in at least 9% of the transactions will

be displayed.



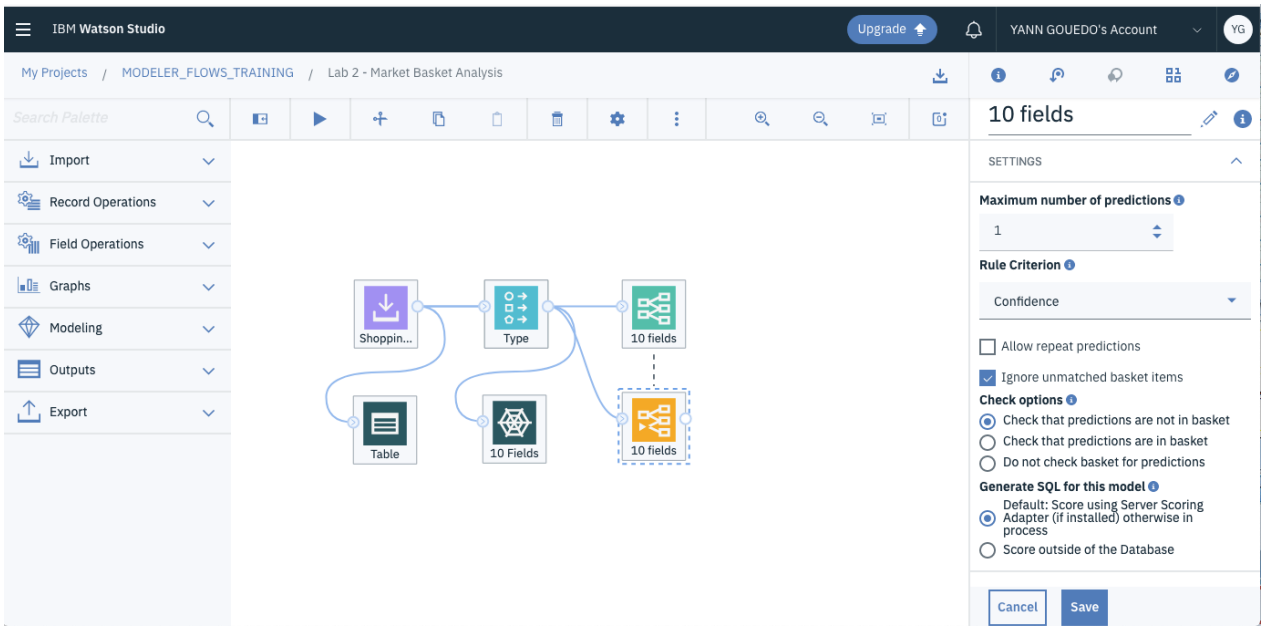
- Save the node
- Click the Run button when finished.

9 - **Open the generated model to see the rules** that have been found.
The way to interpret this is that the antecedent occurred in the support % of the transactions.
Of those, the confidence % also had the consequent.

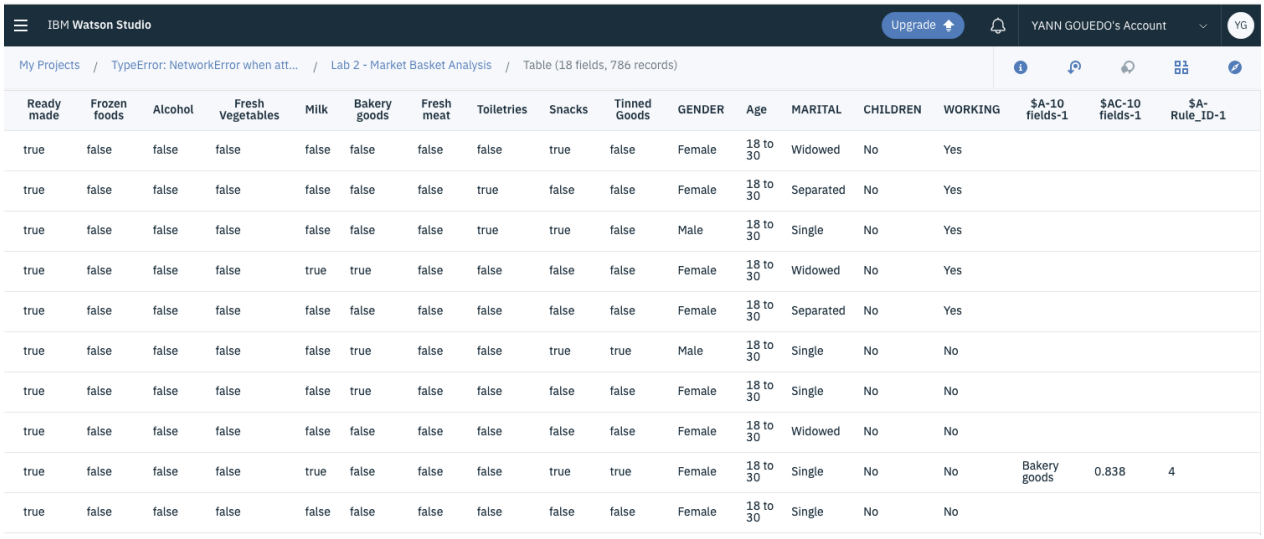


Rule ID	Antecedent	Consequent	Support (%)	Confidence (%)	Rule Support (%)	Lift	Deployability (%)
10	• Frozen foods • Alcohol • Tinned Goods • Ready made	Bakery goods	9.033	85.915	7.761	2.004	1.272
11	• Alcohol • Tinned Goods • Ready made • Snacks	Bakery goods	9.160	84.722	7.761	1.976	1.399
4	• Milk • Tinned Goods • Ready made	Bakery goods	9.415	83.784	7.888	1.954	1.527
1	• Milk • Frozen foods	Bakery goods	10.814	83.529	9.033	1.948	1.781

10 - Double click on the generated model, then **change the maximum number of predictions to 1**, then save it.
This will now make one prediction for each transaction for the based upon the rules .



11 - From the generated model, **Add a Table model**, then run it.
Three new columns were added to the right of the dataset.
The first of the last columns is the predicted recommendation, the second to the last columns is the confidence % based on other observed cases, and the last column shows the RuleID from the
generated model that corresponds to that prediction.



Ready made	Frozen foods	Alcohol	Fresh Vegetables	Milk	Bakery goods	Fresh meat	Toiletries	Snacks	Tinned Goods	GENDER	Age	MARITAL	CHILDREN	WORKING	$A-10 fields-1	$AC-10 fields-1	$A-Rule_ID-1
true	false	false	false	false	false	false	false	true	false	Female	18 to 30	Widowed	No	Yes			
true	false	false	false	false	false	false	true	false	false	Female	18 to 30	Separated	No	Yes			
true	false	false	false	false	false	false	true	true	false	Male	18 to 30	Single	No	Yes			
true	false	false	false	true	true	false	false	false	false	Female	18 to 30	Widowed	No	Yes			
true	false	false	false	false	false	false	false	false	false	Female	18 to 30	Separated	No	Yes			
true	false	false	false	false	true	false	false	true	true	Male	18 to 30	Single	No	No			
true	false	false	false	false	true	false	false	false	false	Female	18 to 30	Single	No	No			
true	false	false	false	false	false	false	false	false	false	Female	18 to 30	Widowed	No	No			
true	false	false	false	true	false	false	false	true	true	Female	18 to 30	Single	No	No	Bakery goods	0.838	4
true	false	false	false	false	false	false	false	false	false	Female	18 to 30	Single	No	No			

Lab 5 – Segmentation Modeling

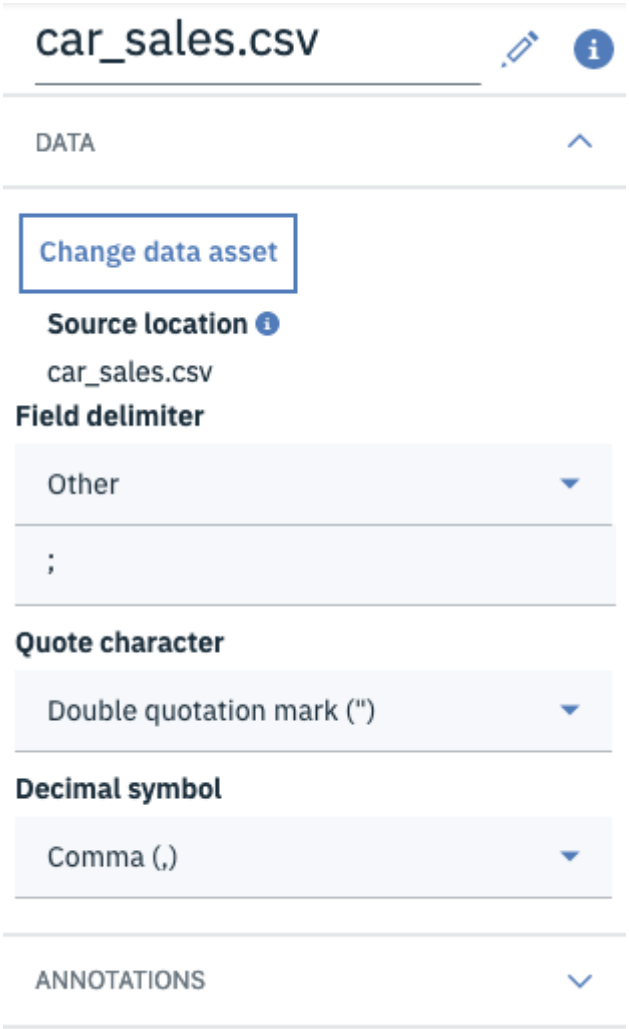
Industry context: automotive/retail
Business objective: to identify type of car purchases in order to define specific actions and decisions.
Technical objective: to use machine learning segmentation capabilities through a visual programming approach.
Data: vehicle information, transactions

1 - **Create a new Modeler Flow**, named it Lab 3 - Segmentation Modeling.

2 - From Import palette, **add an Data Asset node**.

3 - **Define the data asset** to use:

- Double-click on the Data Asset node on the canvas.
- Select *car_sales.csv* clicking on Change data asset / Data assets.
- Ensure to have the following parameters:

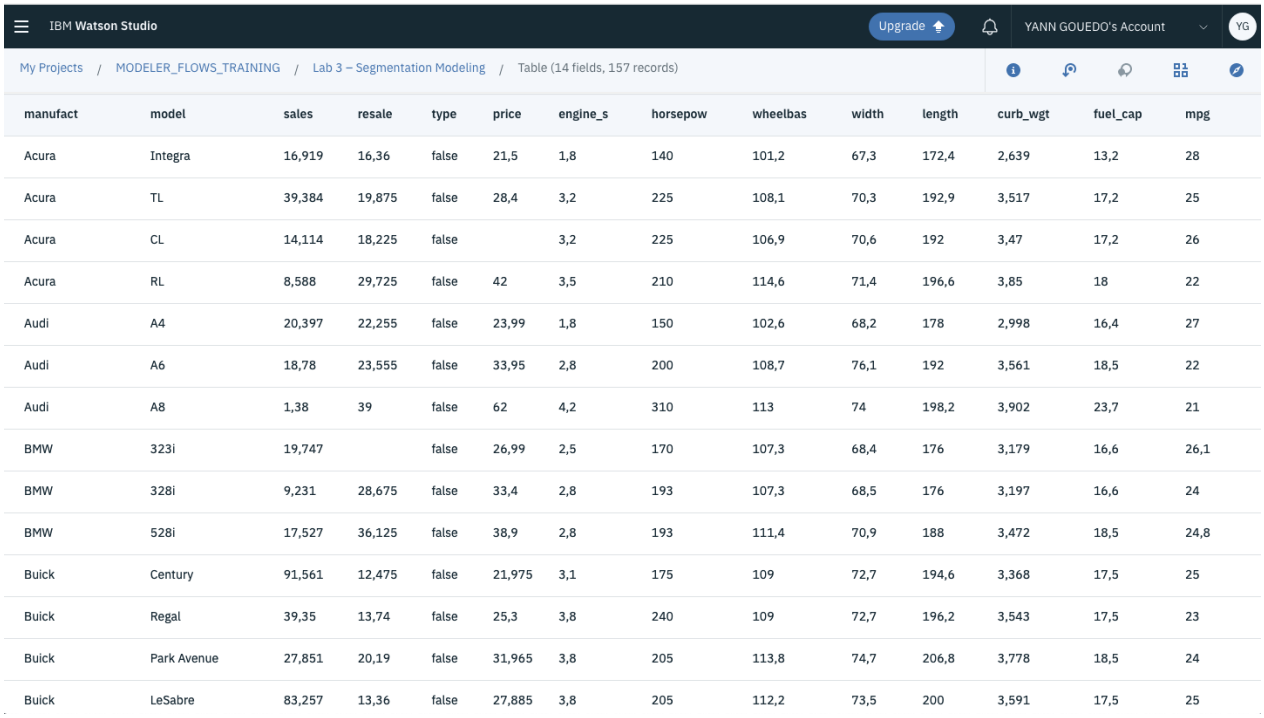


- Save the modification.

4 - Right click on the *car_sales.csv* file node and select Preview from the menu, **to preview the data**.

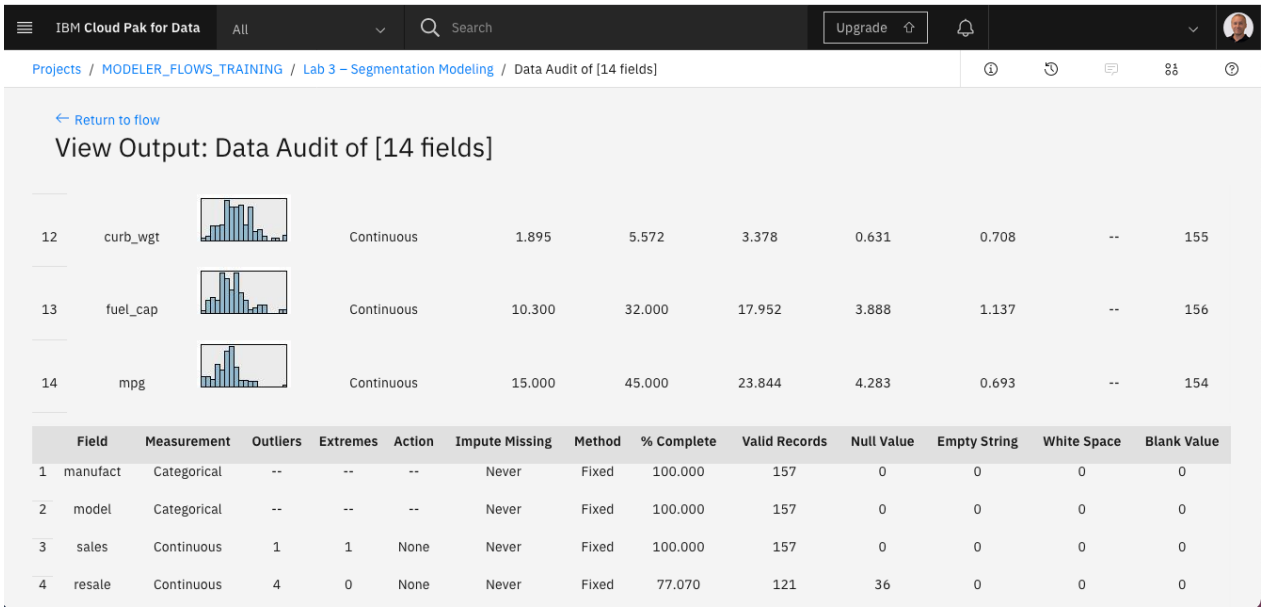
manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width	length
Acura	Integra	16,919	16,36	false	21,5	1,8	140	101,2	67,3	172,4
Acura	TL	39,384	19,875	false	28,4	3,2	225	108,1	70,3	192,9
Acura	CL	14,114	18,225	false		3,2	225	106,9	70,6	192
Acura	RL	8,588	29,725	false	42	3,5	210	114,6	71,4	196,6
Audi	A4	20,397	22,255	false	23,99	1,8	150	102,6	68,2	178
Audi	A6	18,78	23,555	false	33,95	2,8	200	108,7	76,1	192
Audi	A8	1,38	39	false	62	4,2	310	113	74	198,2
BMW	323i	19,747		false	26,99	2,5	170	107,3	68,4	176

5 - From the Outputs palette, **add a Table node** to the canvas and connect it to the data source. Display the data. Note that the data are about the various makes and models of motor vehicles.



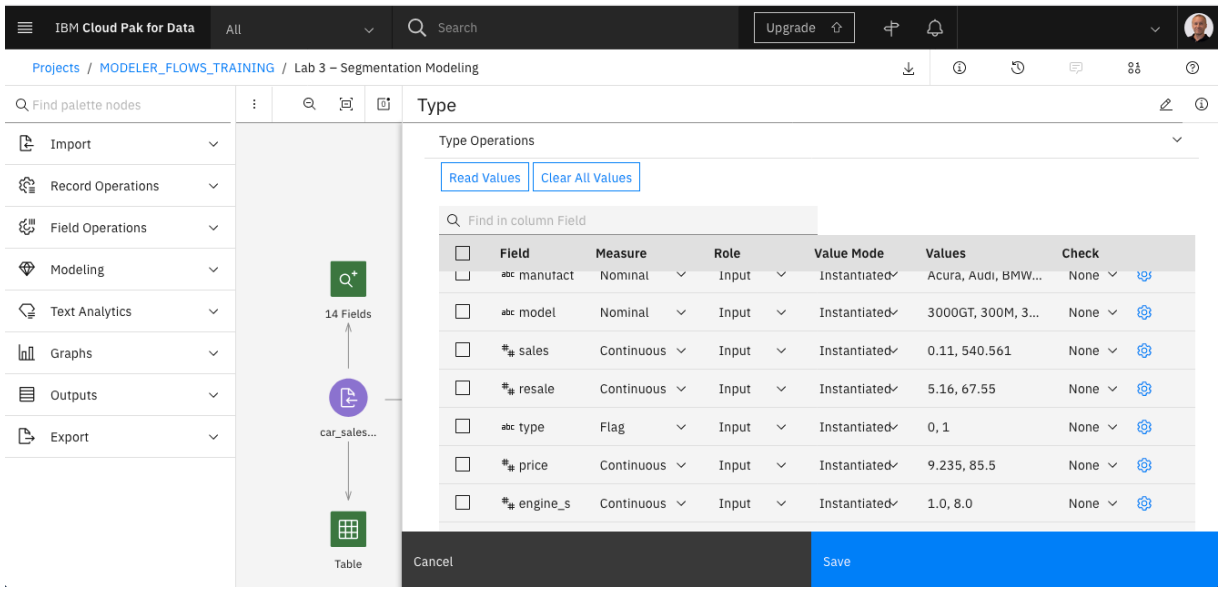
manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width	length	curb_wgt	fuel_cap	mpg
Acura	Integra	16,919	16,36	false	21,5	1,8	140	101,2	67,3	172,4	2,639	13,2	28
Acura	TL	39,384	19,875	false	28,4	3,2	225	108,1	70,3	192,9	3,517	17,2	25
Acura	CL	14,114	18,225	false		3,2	225	106,9	70,6	192	3,47	17,2	26
Acura	RL	8,588	29,725	false	42	3,5	210	114,6	71,4	196,6	3,85	18	22
Audi	A4	20,397	22,255	false	23,99	1,8	150	102,6	68,2	178	2,998	16,4	27
Audi	A6	18,78	23,555	false	33,95	2,8	200	108,7	76,1	192	3,561	18,5	22
Audi	A8	1,38	39	false	62	4,2	310	113	74	198,2	3,902	23,7	21
BMW	323i	19,747		false	26,99	2,5	170	107,3	68,4	176	3,179	16,6	26,1
BMW	328i	9,231	28,675	false	33,4	2,8	193	107,3	68,5	176	3,197	16,6	24
BMW	528i	17,527	36,125	false	38,9	2,8	193	111,4	70,9	188	3,472	18,5	24,8
Buick	Century	91,561	12,475	false	21,975	3,1	175	109	72,7	194,6	3,368	17,5	25
Buick	Regal	39,35	13,74	false	25,3	3,8	240	109	72,7	196,2	3,543	17,5	23
Buick	Park Avenue	27,851	20,19	false	31,965	3,8	205	113,8	74,7	206,8	3,778	18,5	24
Buick	LeSabre	83,257	13,36	false	27,885	3,8	205	112,2	73,5	200	3,591	17,5	25

6 - From the Outputs palette, **add a Data Audit node** to the canvas and connect it to the data source, then run the Data Audit node, named 14 Fields, and view the distributions of the data fields and the summary statistics.



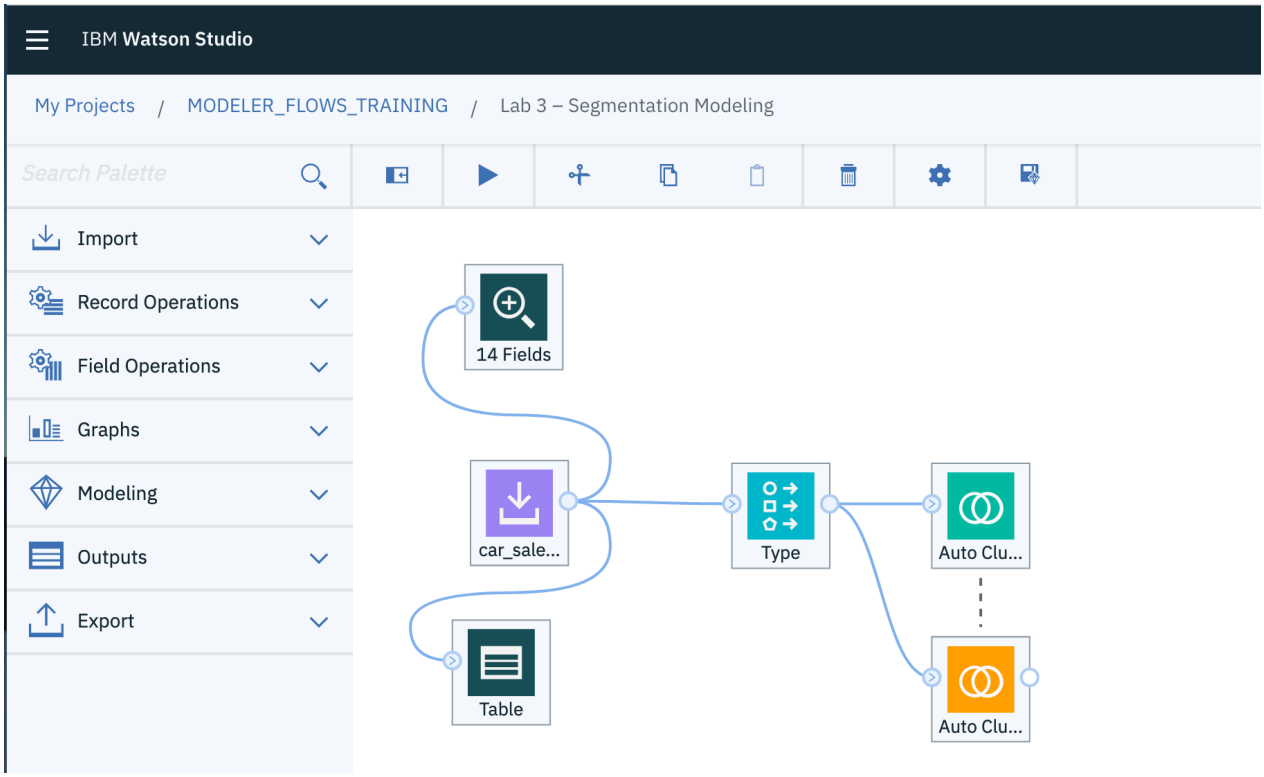
7 **Define the roles of the fields**:

- From the Field Operations palette, add a Type node to the canvas and connect it to the data source.
- Open the Type node, clicking twice, confirm that the roles of the fields is set to Input.
- Click the Read Values button to scan the data as well as to display and update the range of values.



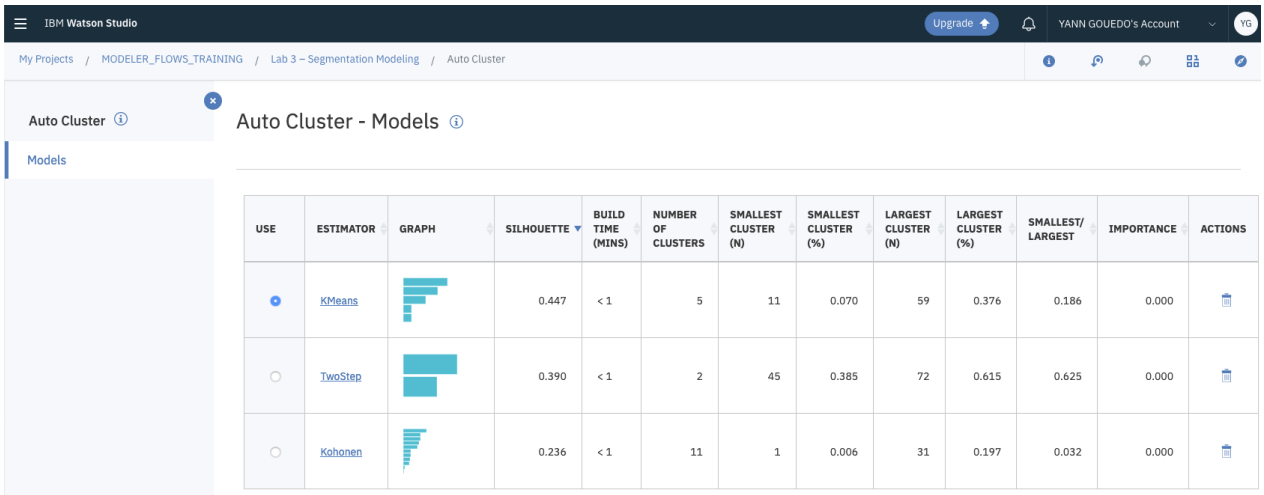
These fields will be used to cluster or group the vehicles based on their similarities.

8 - From the Modeling palette, **add the Auto Cluster** node to the Type node and run it.



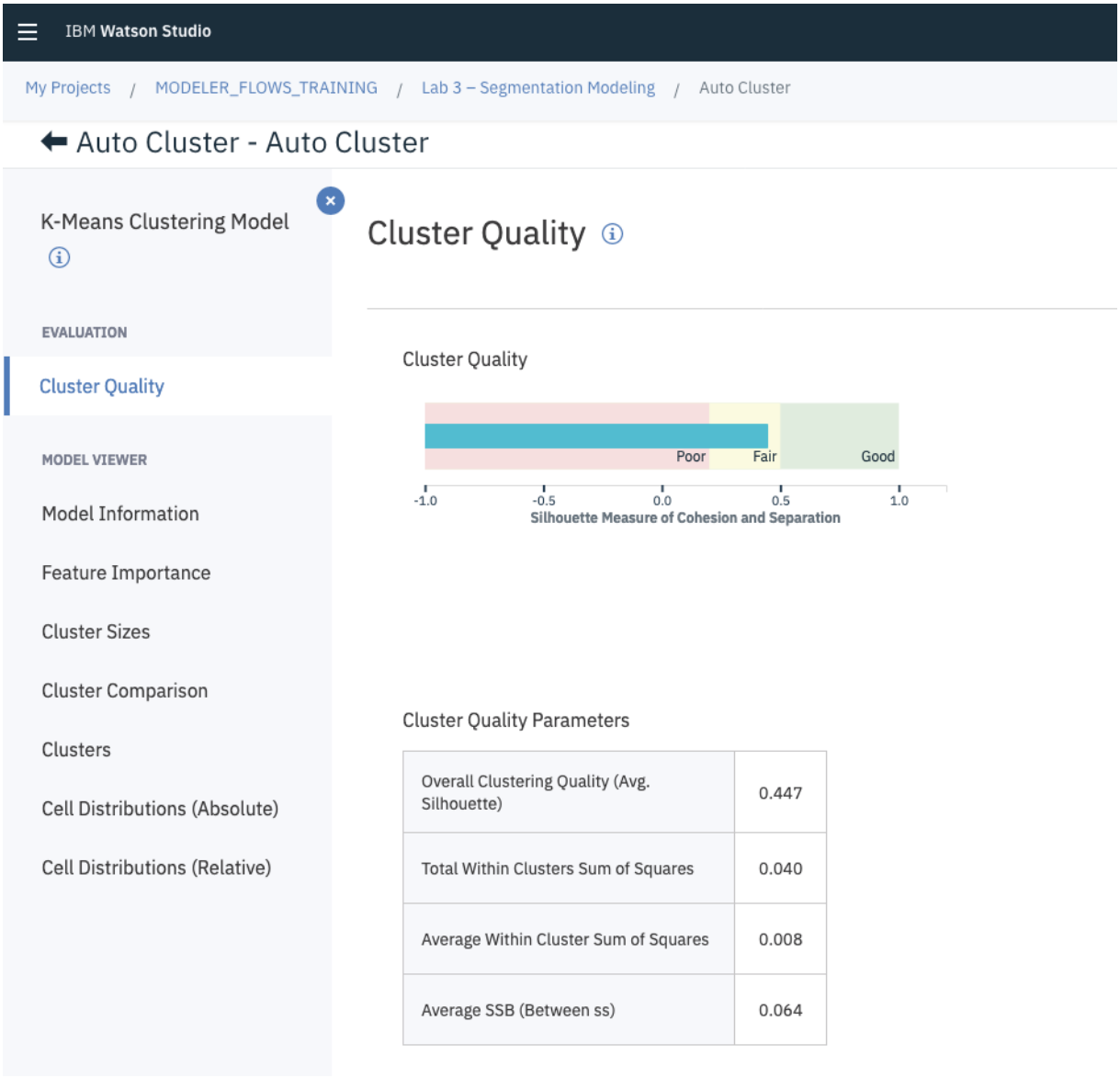
9 - From the Auto-Cluster model node, use the context menu to **view the results of the clustering calculation**.

This dialog box displays a graph illustrating the number of clusters for each algorithm as well as the method and a number of summary statistics used to evaluate the clustering method. The higher is the Silhouette, the higher is the segmentation accuracy.

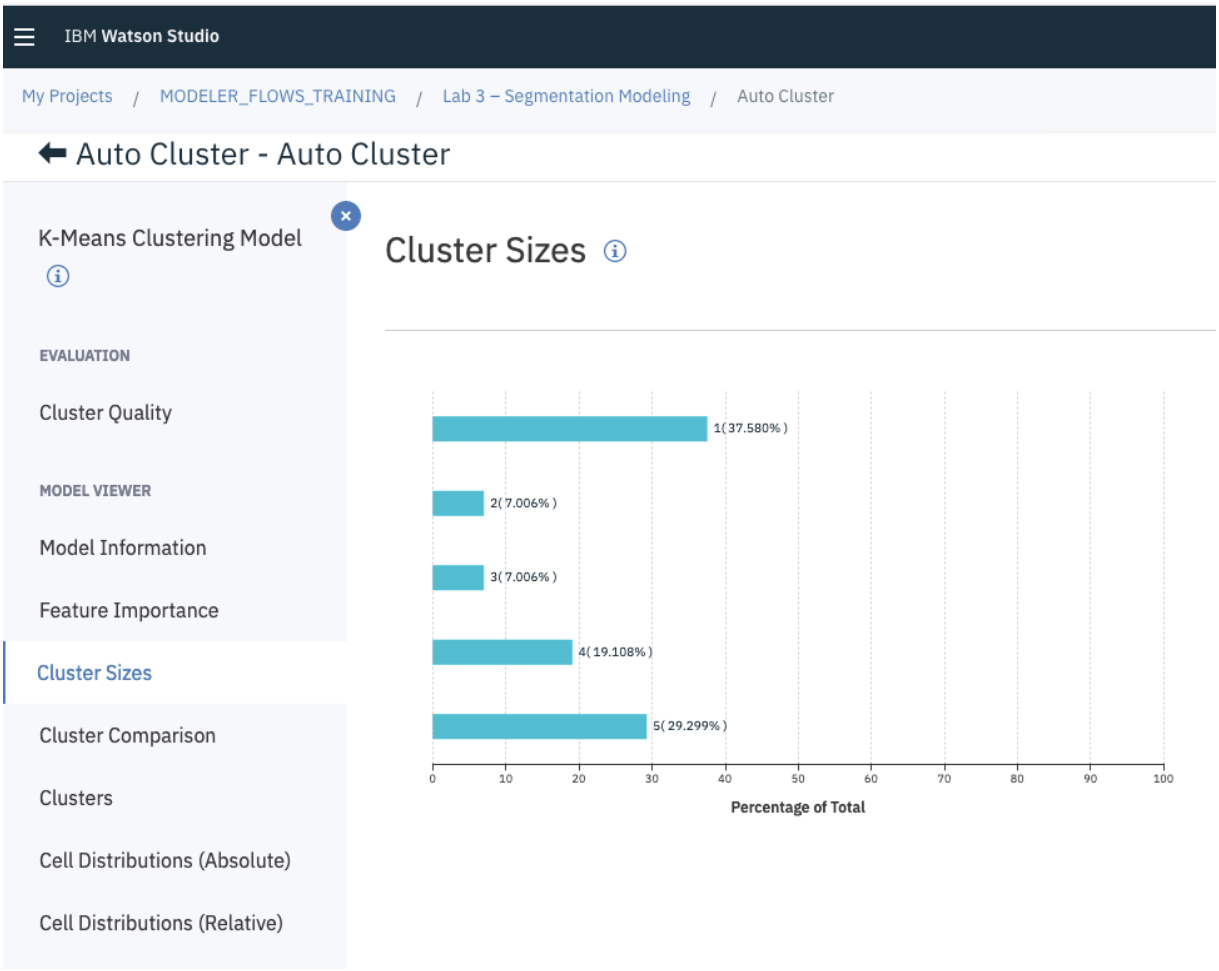


10 - Click the **K-means model** to see more details. The output is divided into several windows:

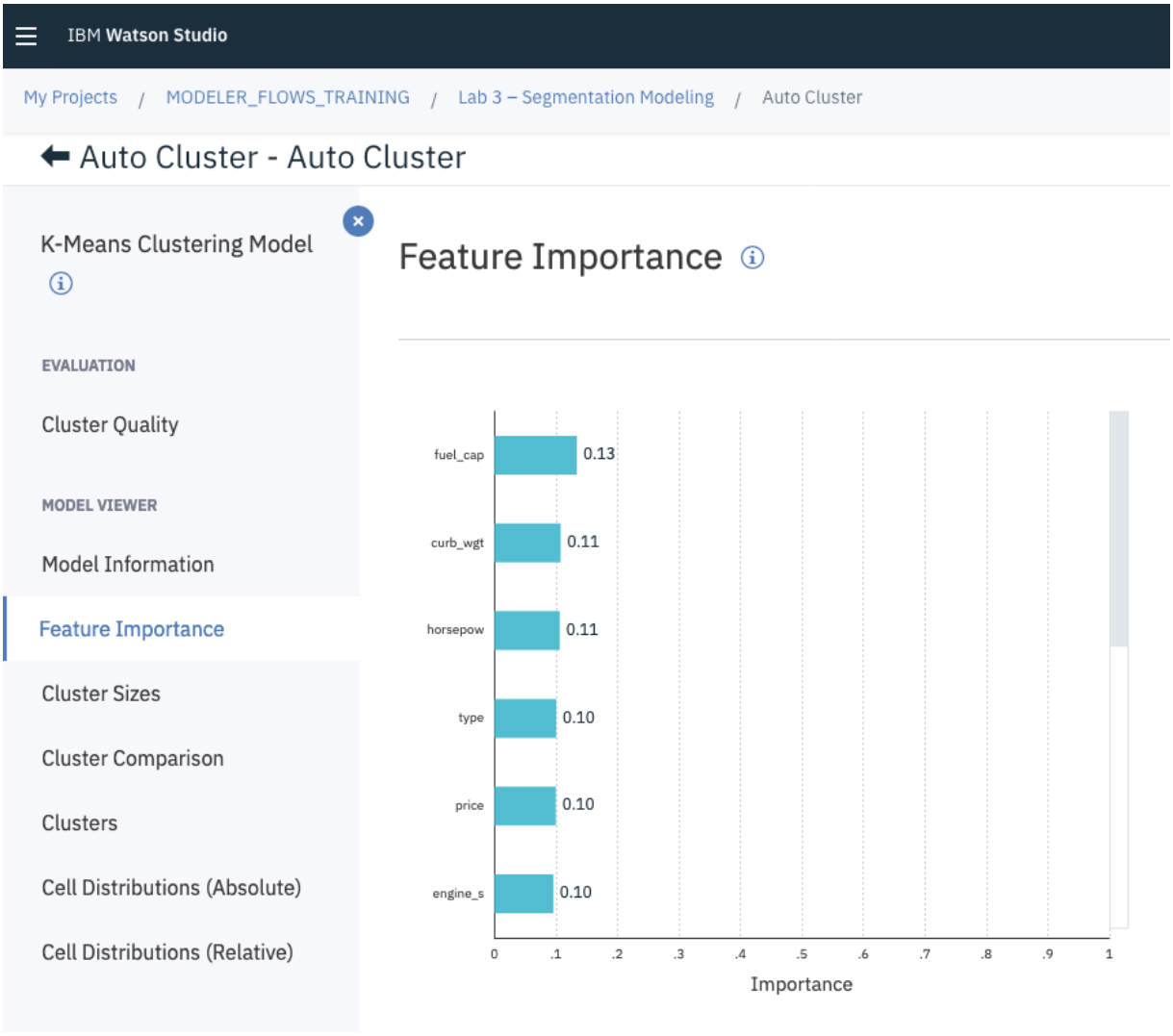
- Cluster Quality



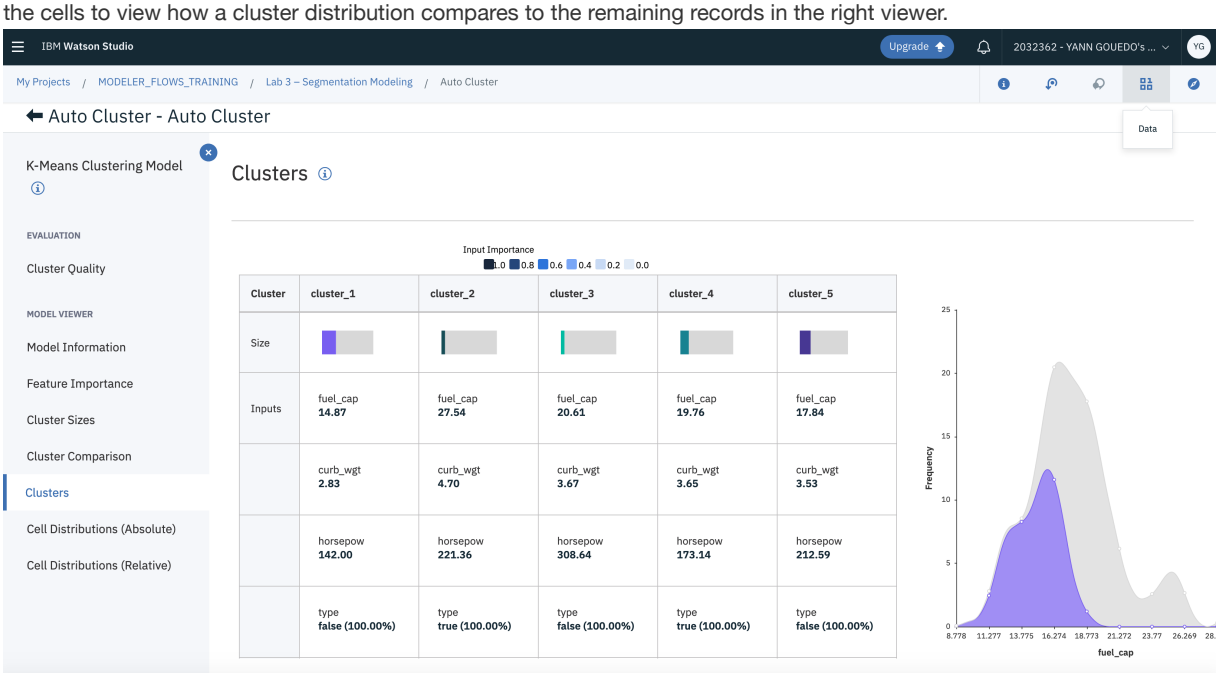
- Cluster Sizes



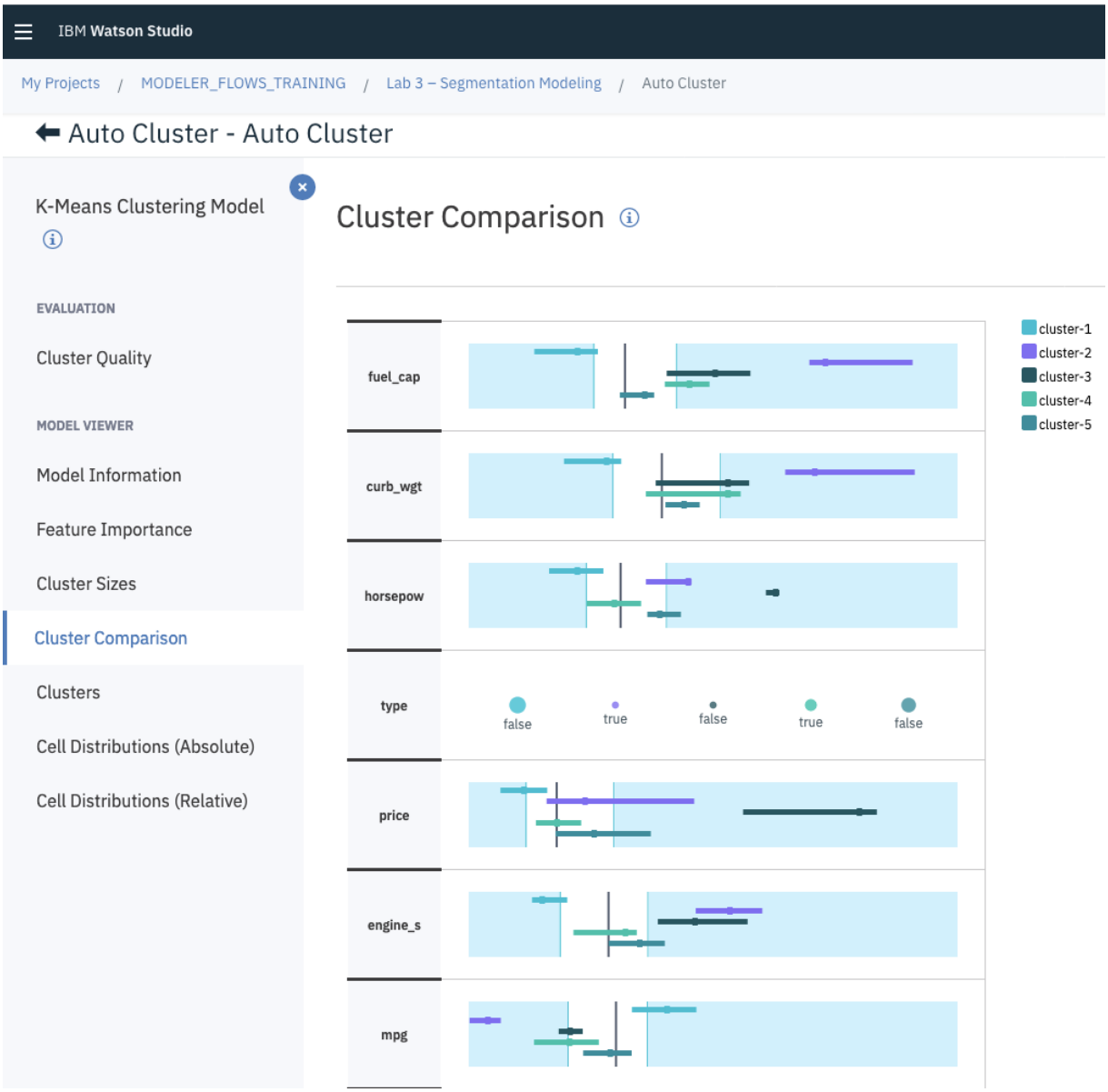
- Feature Importance: it displays a graph with the fields ranked in order of importance for the cluster creation.



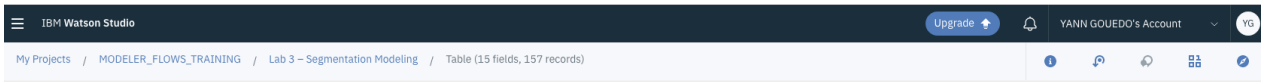
- Cluster: it displays a grid with the clusters listed and the fields listed in descending order based on importance. Click one of the cells to view how a cluster distribution compares to the remaining records in the right viewer.



- Cluster Comparison:



11 - **Add a Table** node to the generated model to see the cluster membership variable added to the dataset.



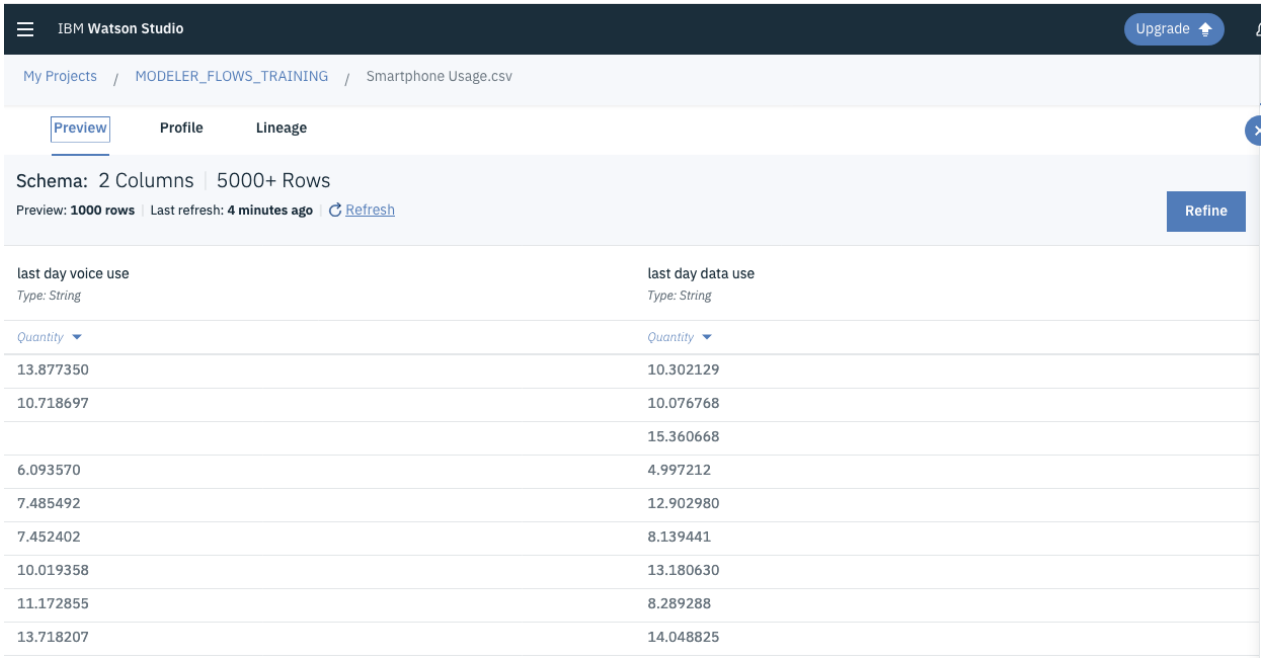
My Projects / MODELER_FLOWS_TRAINING / Lab 3 – Segmentation Modeling / Table (15 fields, 157 records)

manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width	length	curb_wgt	fuel_cap	mpg	$XC-autocluster
Acura	Integra	16.919	16.360	false	21.500	1.800	140	101.200	67.300	172.400	2.639	13.200	28.000	cluster-1
Acura	TL	39.384	19.875	false	28.400	3.200	225	108.100	70.300	192.900	3.517	17.200	25.000	cluster-5
Acura	CL	14.114	18.225	false		3.200	225	106.900	70.600	192.000	3.470	17.200	26.000	cluster-5
Acura	RL	8.588	29.725	false	42.000	3.500	210	114.600	71.400	196.600	3.850	18.000	22.000	cluster-5
Audi	A4	20.397	22.255	false	23.990	1.800	150	102.600	68.200	178.000	2.998	16.400	27.000	cluster-1
Audi	A6	18.780	23.555	false	33.950	2.800	200	108.700	76.100	192.000	3.561	18.500	22.000	cluster-5
Audi	A8	1.380	39.000	false	62.000	4.200	310	113.000	74.000	198.200	3.902	23.700	21.000	cluster-3
BMW	323i	19.747		false	26.990	2.500	170	107.300	68.400	176.000	3.179	16.600	26.100	cluster-1
BMW	328i	9.231	28.675	false	33.400	2.800	193	107.300	68.500	176.000	3.197	16.600	24.000	cluster-1
BMW	528i	17.527	36.125	false	38.900	2.800	193	111.400	70.900	188.000	3.472	18.500	24.800	cluster-5
Buick	Century	91.561	12.475	false	21.975	3.100	175	109.000	72.700	194.600	3.368	17.500	25.000	cluster-5
Buick	Regal	39.350	13.740	false	25.300	3.800	240	109.000	72.700	196.200	3.543	17.500	23.000	cluster-5
Buick	Park Avenue	27.851	20.190	false	31.965	3.800	205	113.800	74.700	206.800	3.778	18.500	24.000	cluster-5
Buick	LeSabre	83.257	13.360	false	27.885	3.800	205	112.200	73.500	200.000	3.591	17.500	25.000	cluster-5
Cadillac	DeVille	63.729	22.525	false	39.895	4.600	275	115.300	74.500	207.200	3.978	18.500	22.000	cluster-5

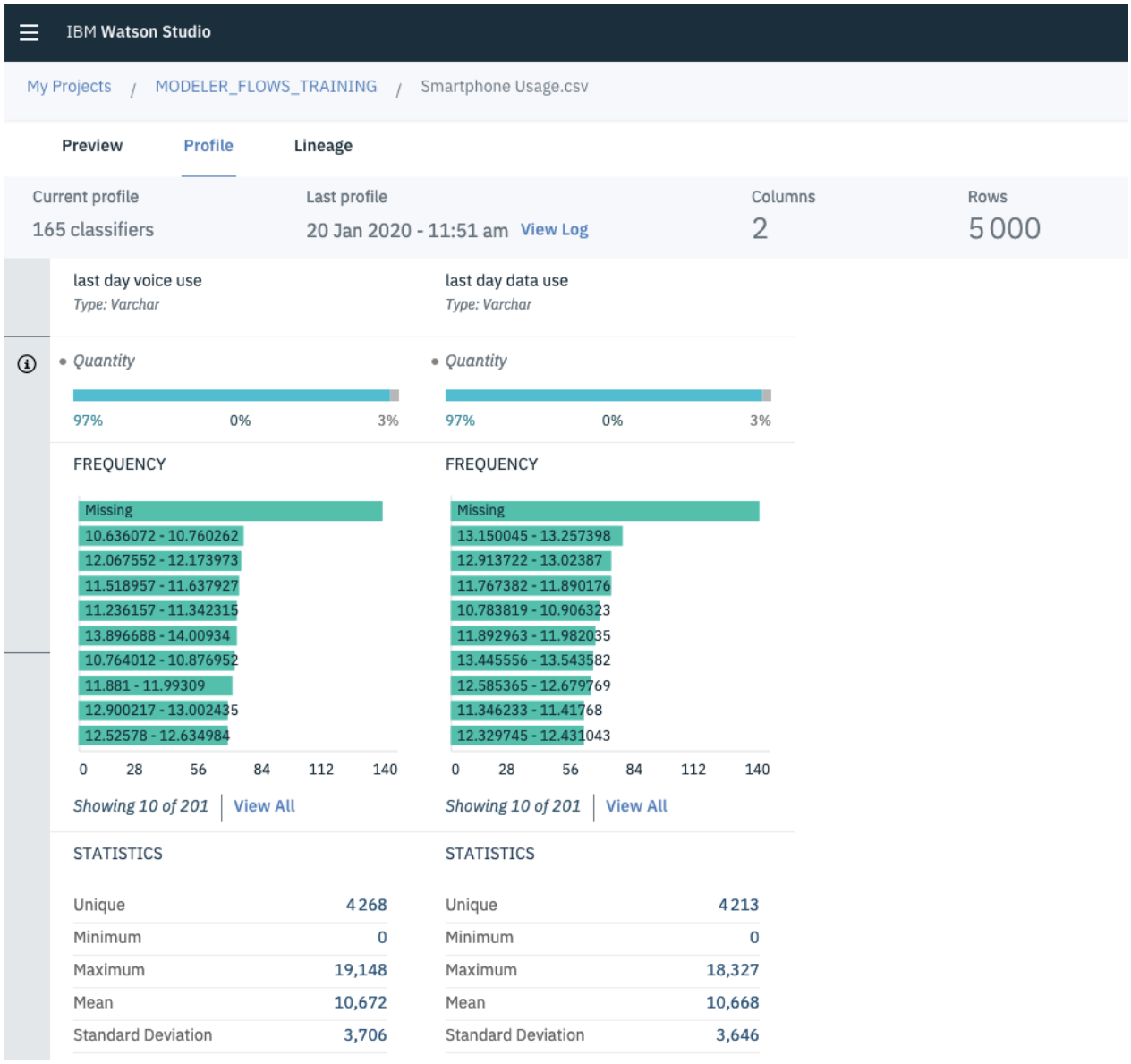
Lab 6 – Anomaly Detection

Industry context: telecommunication
Business objective: to identify abnormal behaviors regarding the consumption of data and voice data.
Technical objective: to use anomaly detection capabilities through a visual programming approach.
Data: consumption of data and voice data

1 - From the Data Assets section, click on the *Smartphone Usage.csv* **to view the data, and the profile**. It is an extraction of the consumption (voice usage and data) from customers who have a contract with a telecommunications company's.



last day voice use *Type: String*	last day data use *Type: String*
Quantity	*Quantity*
13.877350	10.302129
10.718697	10.076768
	15.360668
6.093570	4.997212
7.485492	12.902980
7.452402	8.139441
10.019358	13.180630
11.172855	8.289288
13.718207	14.048825



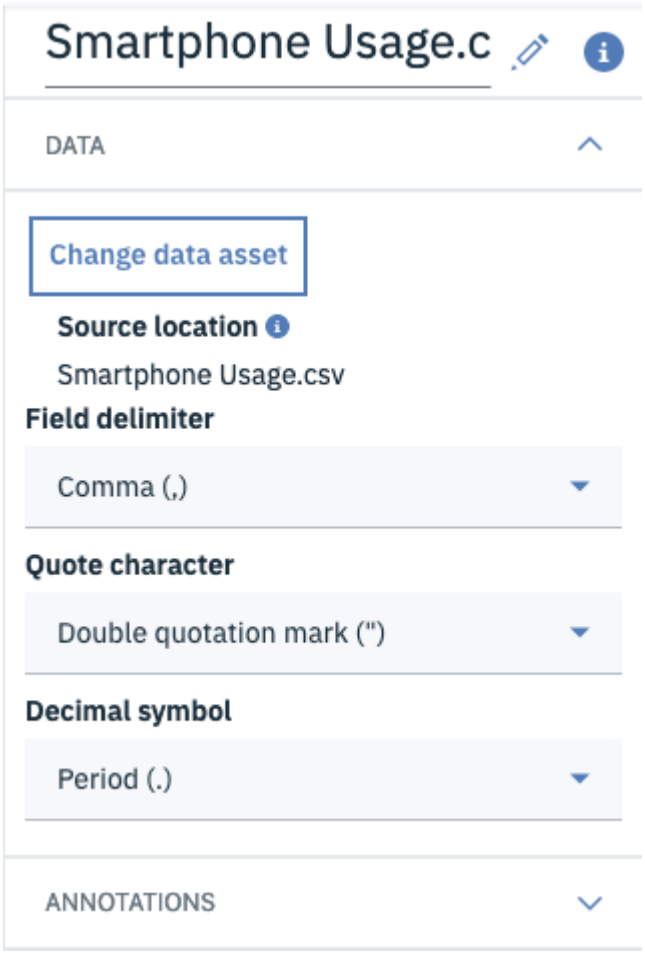
2 - **Create a new Modeler Flow**, named it Lab 5 - Anomaly Detection.

3 - From Import palette, **add an Data Asset node**.

4 - **Define the data asset** to use:

- Double-click on the Data Asset node on the canvas.
- Select *Smartphone Usage.csv* clicking on Change data asset / Data assets.

- Ensure to have the following parameters:



- Save the modification.

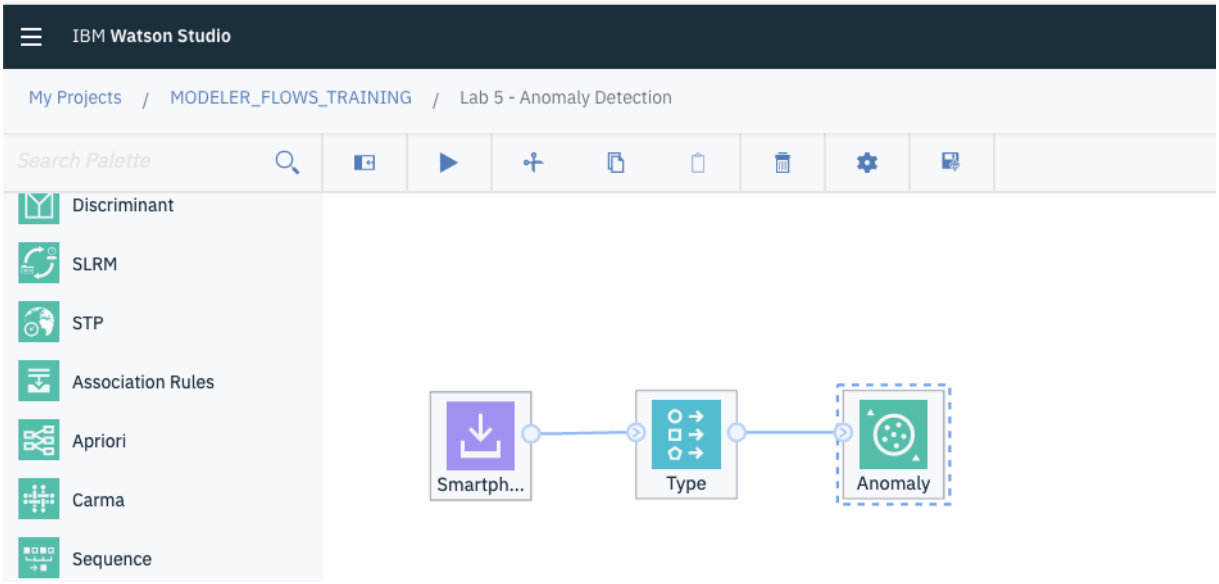
5 - Right click on the *Smartphone Usage.csv* file node and select Preview from the menu, **to preview the data**.

6 - **Read the values of the fields**:

- From the Field Operations palette, add a Type node to the canvas and connect it to the data source.
- Open the Type node, clicking twice, and click the Read Values button to scan the data as well as to display and update the range of values.
- Save it.

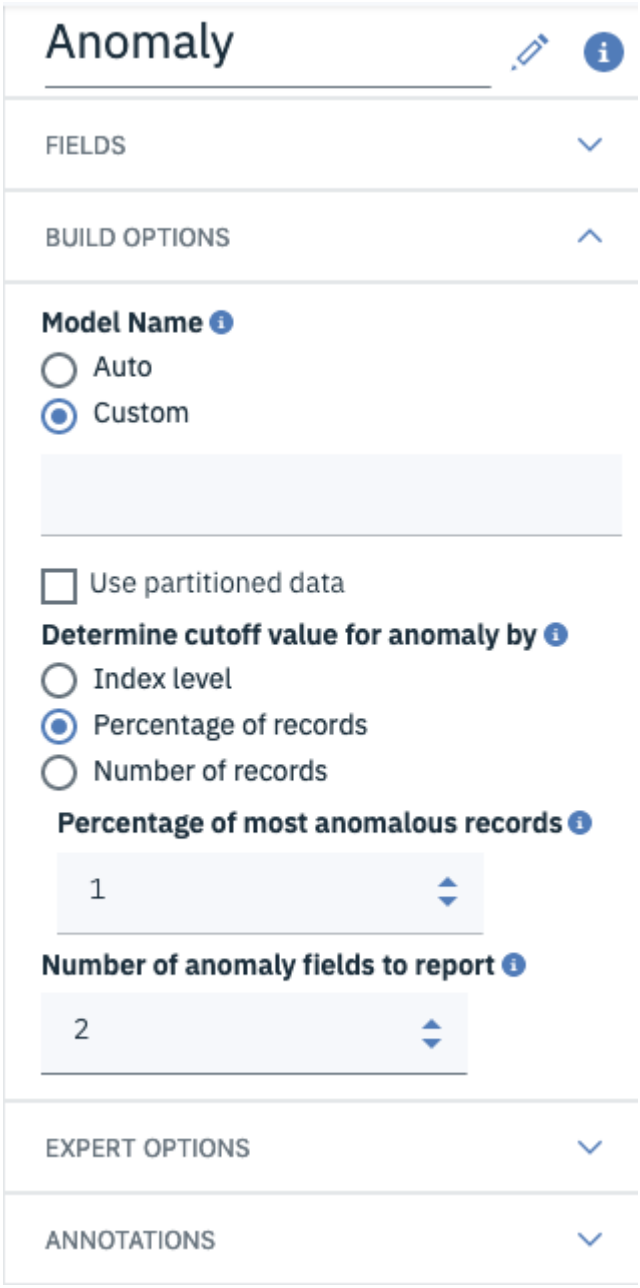
7 - Use of the **anomaly detection** algorithm. Anomaly detection (also known as outlier detection) is the search for items or events which do not conform to an expected pattern.The patterns thus detected are called anomalies and often translate to critical and actionable information in several application domains, particularly useful in applications, such as fraud detection, where new patterns may constantly be emerging.

- From the Modeling palette, **add an Anomaly node** to the Type node.

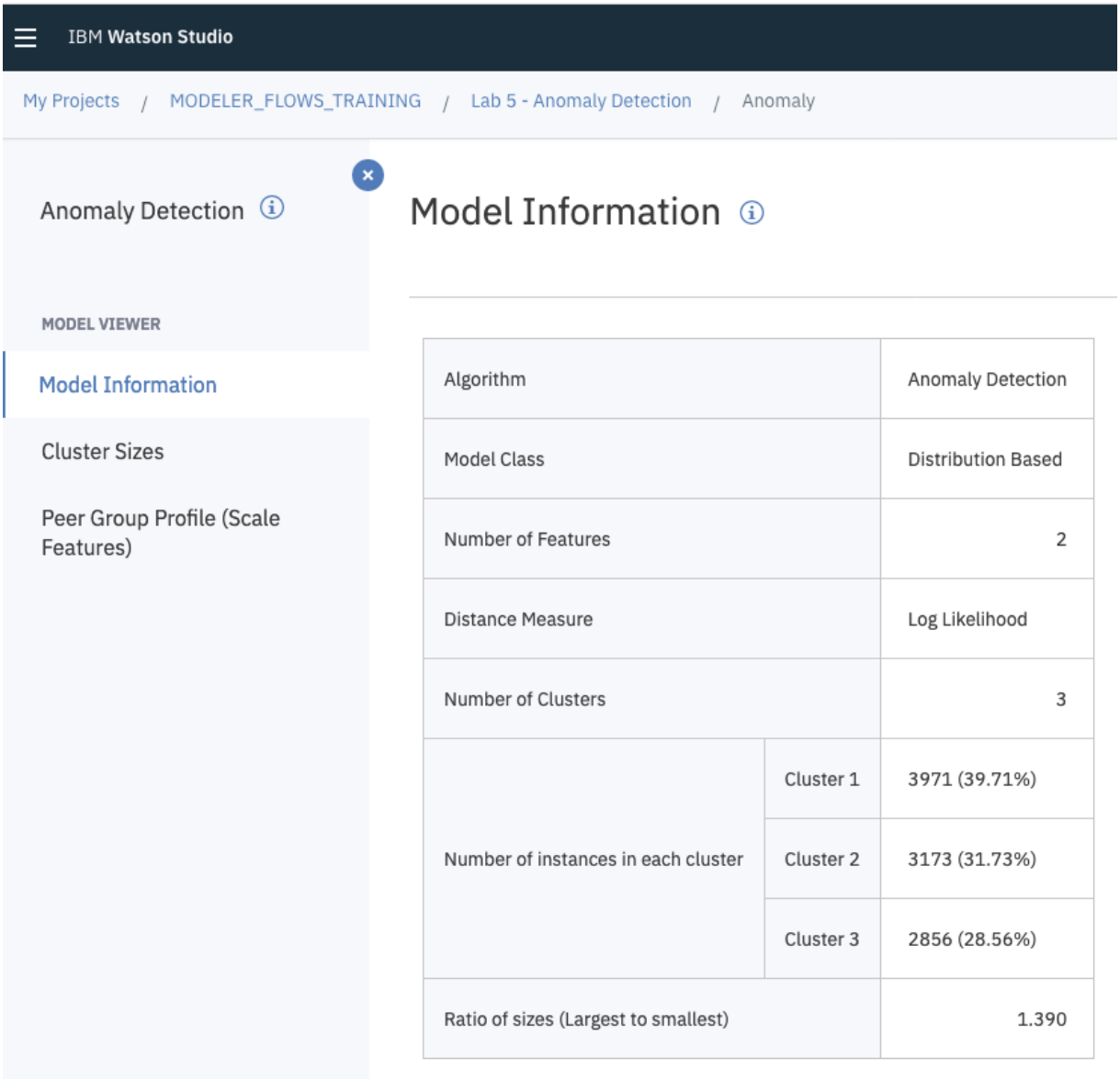


- Open the Anomaly detection node, clicking twice, then modify the number of anomaly fields to report to 2 (because of the *Smartphone Usage.csv* file has only 2 fields) within the Build Options tab. The calculation will highlight 1% of the most

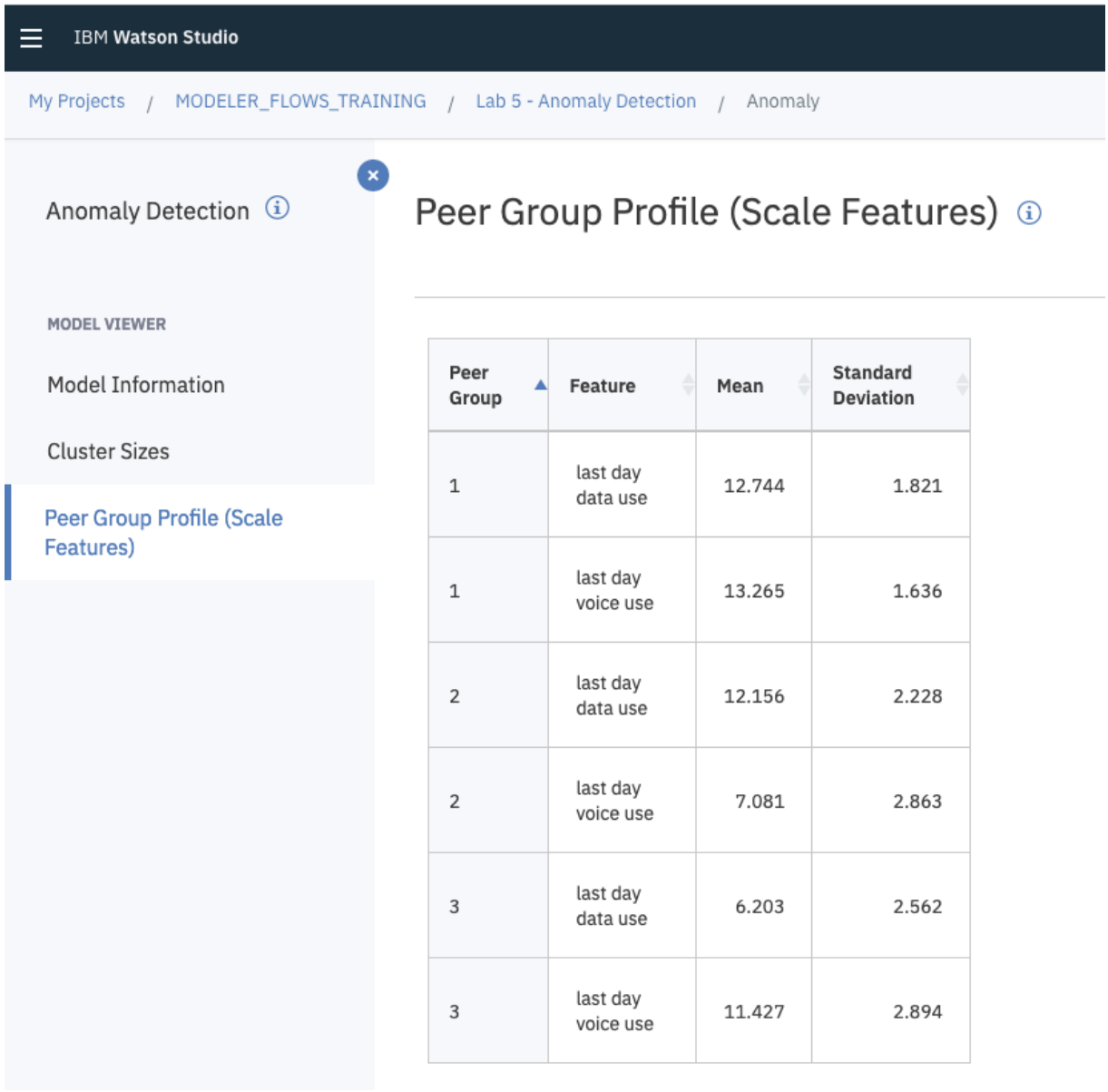
anomalous records.



- Save it.
- Run the Anomaly detection node from the context menu. The model (named a pattern) is generated (yellow nugget node), and added automatically to the canvas and connected to the Type node.
- Explore the anomaly model, using the View Model from the context menu.



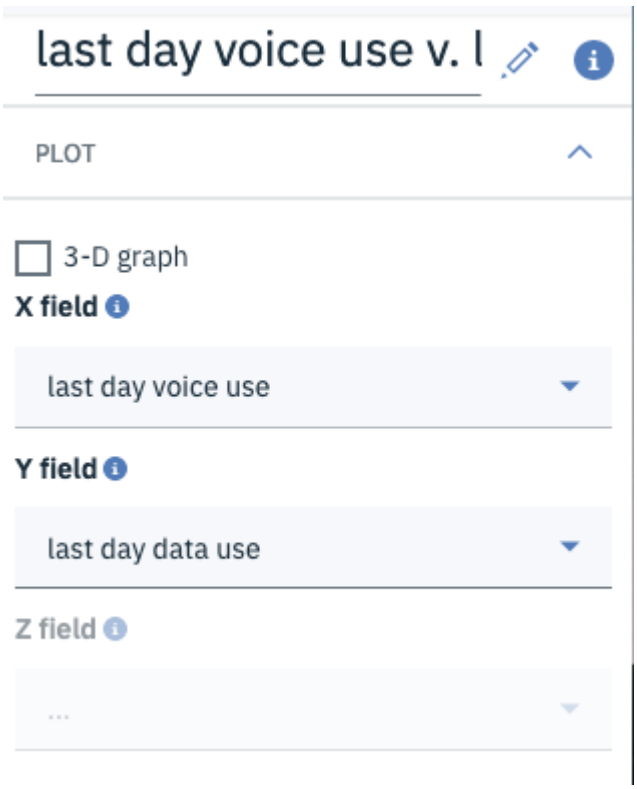
Algorithm		Anomaly Detection
Model Class		Distribution Based
Number of Features		2
Distance Measure		Log Likelihood
Number of Clusters		3
Number of instances in each cluster	Cluster 1	3971 (39.71%)
	Cluster 2	3173 (31.73%)
	Cluster 3	2856 (28.56%)
Ratio of sizes (Largest to smallest)		1.390



Peer Group	Feature	Mean	Standard Deviation
1	last day data use	12.744	1.821
1	last day voice use	13.265	1.636
2	last day data use	12.156	2.228
2	last day voice use	7.081	2.863
3	last day data use	6.203	2.562
3	last day voice use	11.427	2.894

8 - Display of Last day voice use and Last day data use.

- From the Graphs palette, select a Plot node to Display the last day voice use and last day data use, with a group colored by the anomaly index.


last day voice use v. l
PLOT
3-D graph
X field
last day voice use
Y field
last day data use
Z field
...

last day voice use v. l

PLOT

OVERLAY

Color

$O-AnomalyIndex

Size

...

Shape

...

Panel

...

Transparency

...

Overlay type

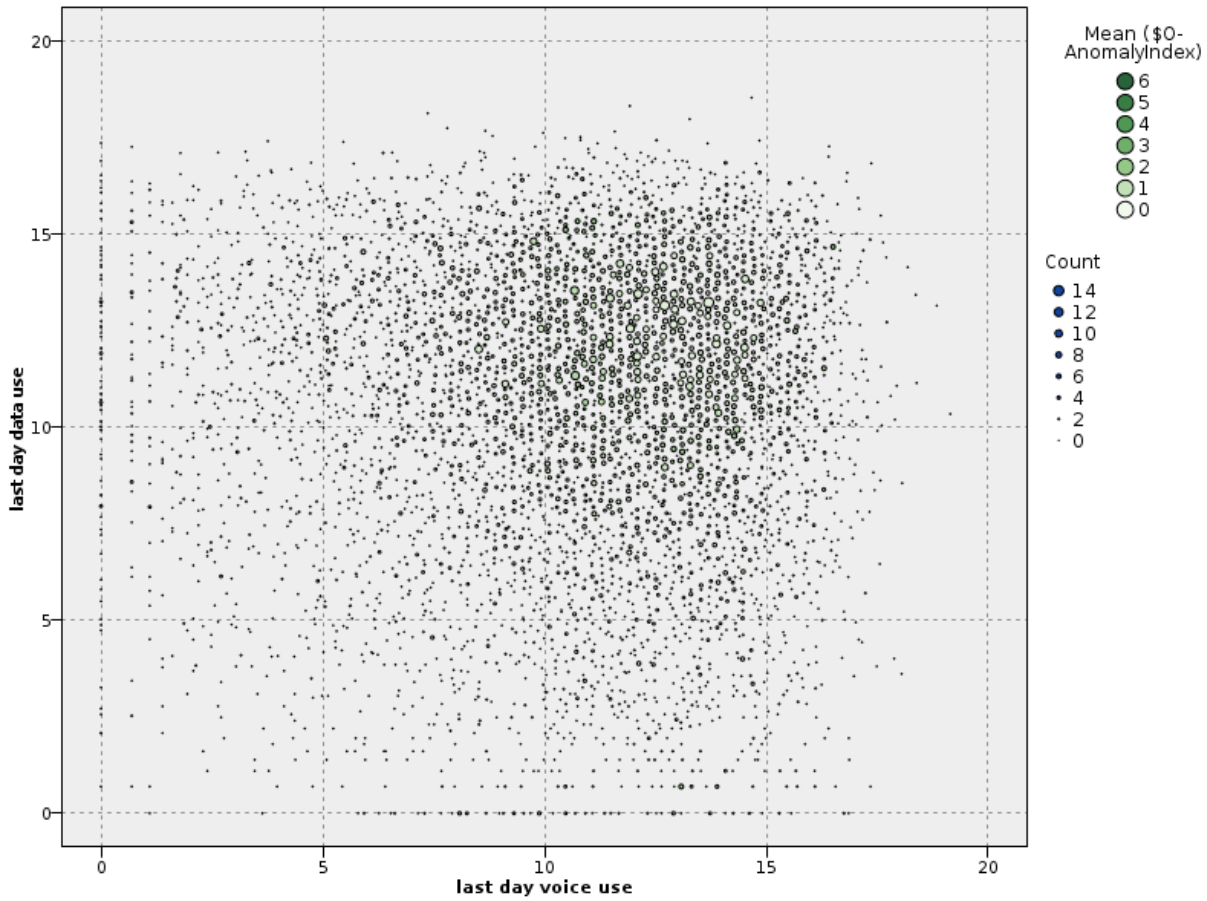
- (•) None
- ( ) Smoother
- ( ) Function

OPTIONS

APPEARANCE

ANNOTATIONS

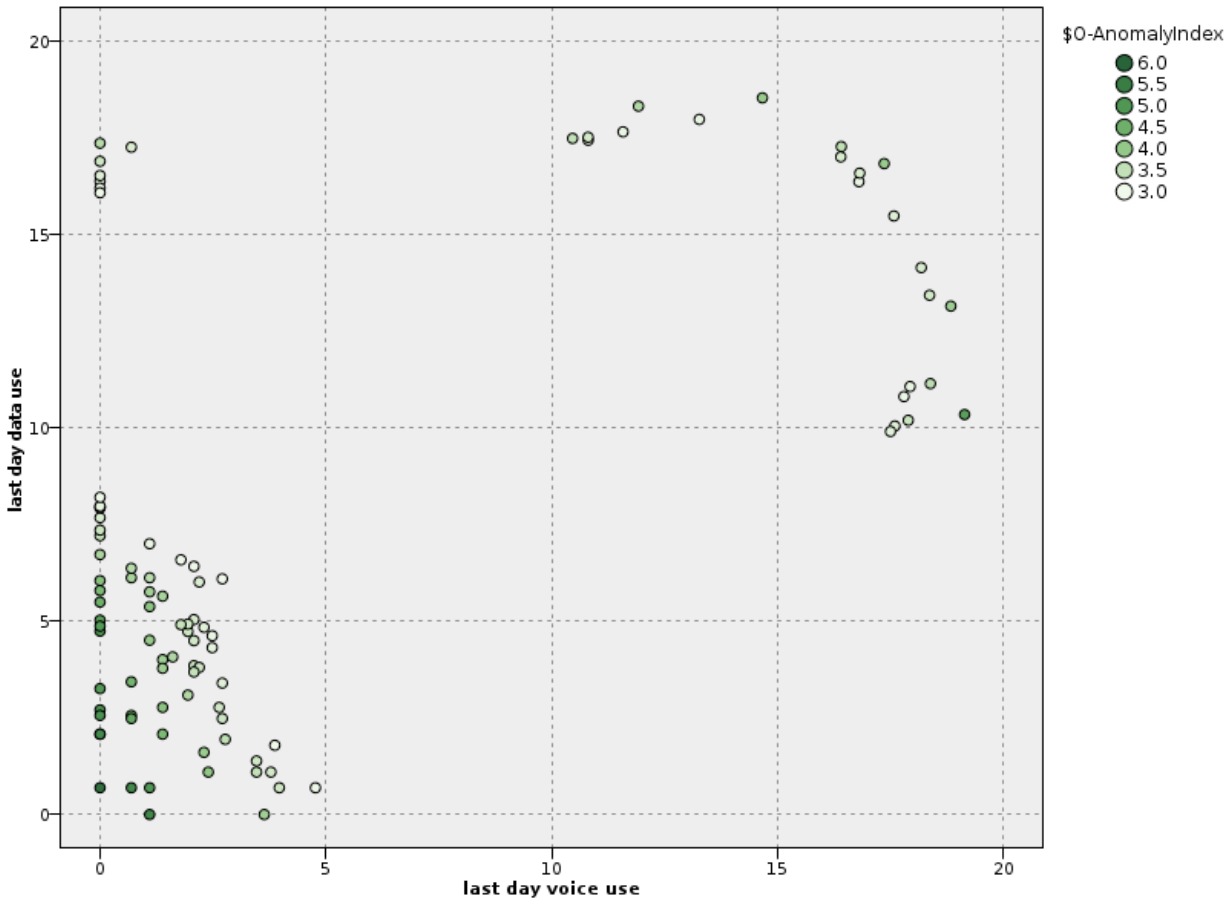
- Run the plot node from the context menu



9 - From the Record Operations palette, **add a Select node** and connect it to the Anomaly model node. Edit the Select node, then use the Expression Builder **to define the following selection**: '*$O-Anomaly*' = "*F*". Discard the selected records Click OK, then save it.

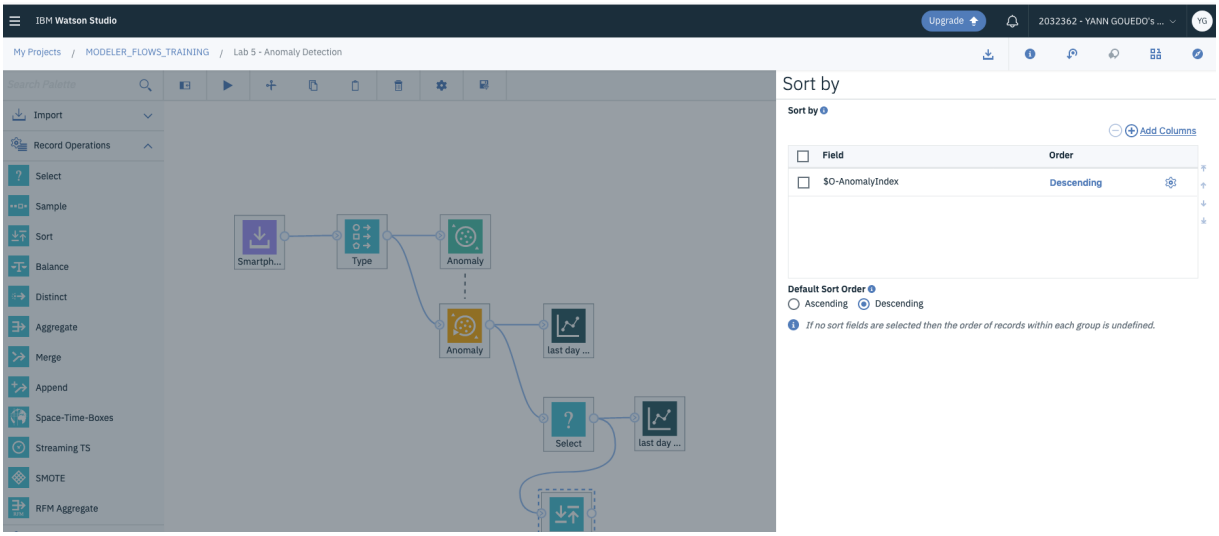
10 - **Copy the previous Plot node**, from the Edit/Copy context menu.

11 - Paste it into the flow and connect it to the Select node. Run the node **to plot the 1% of most anomalous records**.

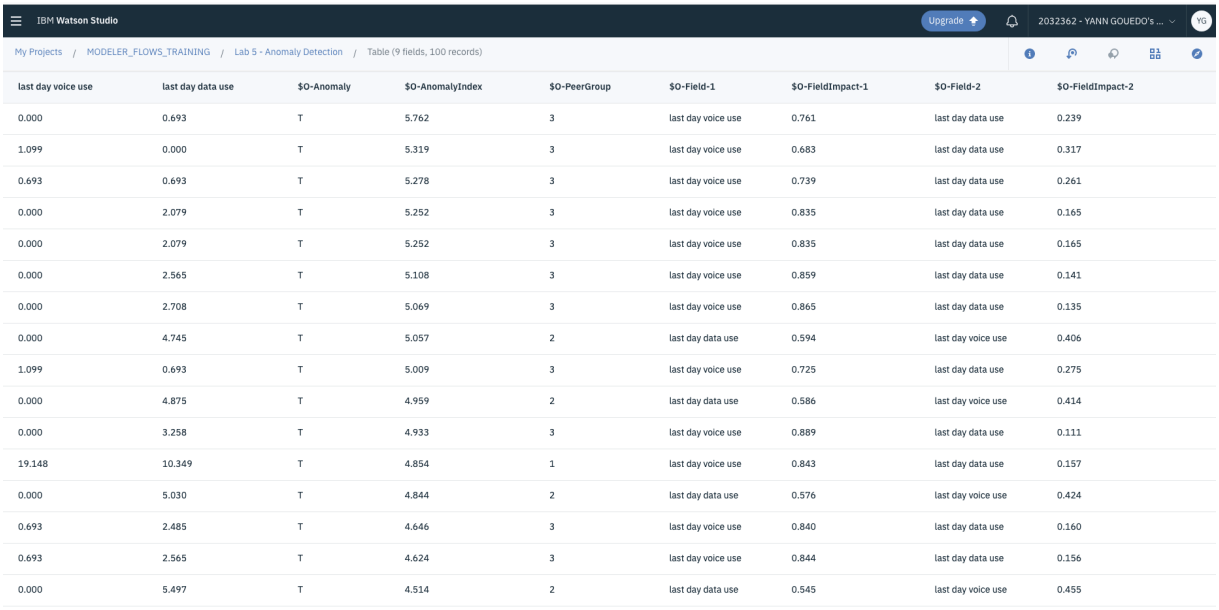


12 - **Display of the most anomalous records**

- Sort the anomalous records in descending, using the Sort node from the Record Operations palette. The node has to be connected to the Select node.



- Connect a Table node to the Sort node and run it. It displays the descending list of the most anomalous records.



last day voice use	last day data use	$O-Anomaly	$O-AnomalyIndex	$O-PeerGroup	$O-Field-1	$O-FieldImpact-1	$O-Field-2	$O-FieldImpact-2
0.000	0.693	T	5.762	3	last day voice use	0.761	last day data use	0.239
1.099	0.000	T	5.319	3	last day voice use	0.683	last day data use	0.317
0.693	0.693	T	5.278	3	last day voice use	0.739	last day data use	0.261
0.000	2.079	T	5.252	3	last day voice use	0.835	last day data use	0.165
0.000	2.079	T	5.252	3	last day voice use	0.835	last day data use	0.165
0.000	2.565	T	5.108	3	last day voice use	0.859	last day data use	0.141
0.000	2.708	T	5.069	3	last day voice use	0.865	last day data use	0.135
0.000	4.745	T	5.057	2	last day data use	0.594	last day voice use	0.406
1.099	0.693	T	5.009	3	last day voice use	0.725	last day data use	0.275
0.000	4.875	T	4.959	2	last day data use	0.586	last day voice use	0.414
0.000	3.258	T	4.933	3	last day voice use	0.889	last day data use	0.111
19.148	10.349	T	4.854	1	last day voice use	0.843	last day data use	0.157
0.000	5.030	T	4.844	2	last day data use	0.576	last day voice use	0.424
0.693	2.485	T	4.646	3	last day voice use	0.840	last day data use	0.160
0.693	2.565	T	4.624	3	last day voice use	0.844	last day data use	0.156
0.000	5.497	T	4.514	2	last day data use	0.545	last day voice use	0.455