Watson Studio 2.1: Data Science on Unstructured Data in Watson Studio on CP4D

Session ID: EDA23T030BH

Sidney Phoon, NA Team, Data Science and IA
Steven Reeves, NA Team, Data Science and IA

Table of contents

Contents

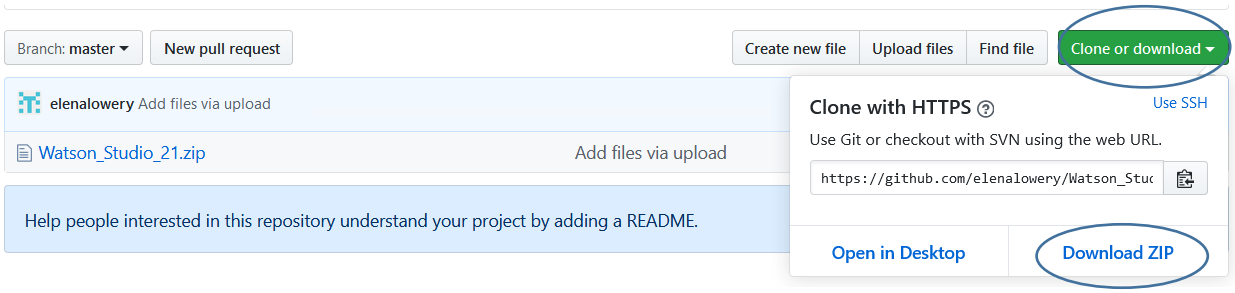
Overview

In this lab you will complete two text analytics exercises in **Watson Studio 2.1**:

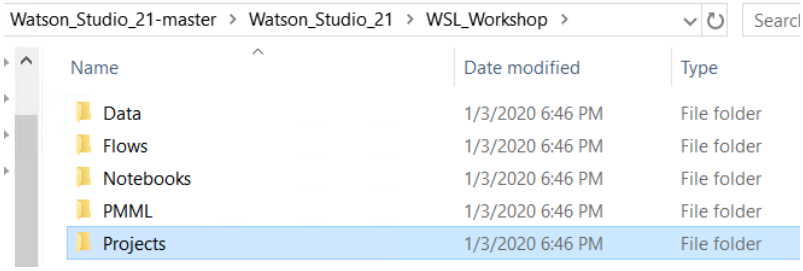
- Text analytics with open source
- Text analytics in SPSS Modeler in Watson Studio.

Required software, access, and files

- To complete this lab, you will need access to a **Cloud Pak for Data** (CP4D) cluster with **Watson Studio**.
- You will also need to download and unzip this GitHub repository: https://github.com/elenalowery/Watson_Studio_21



- Unzip the files until you get to this directory structure:

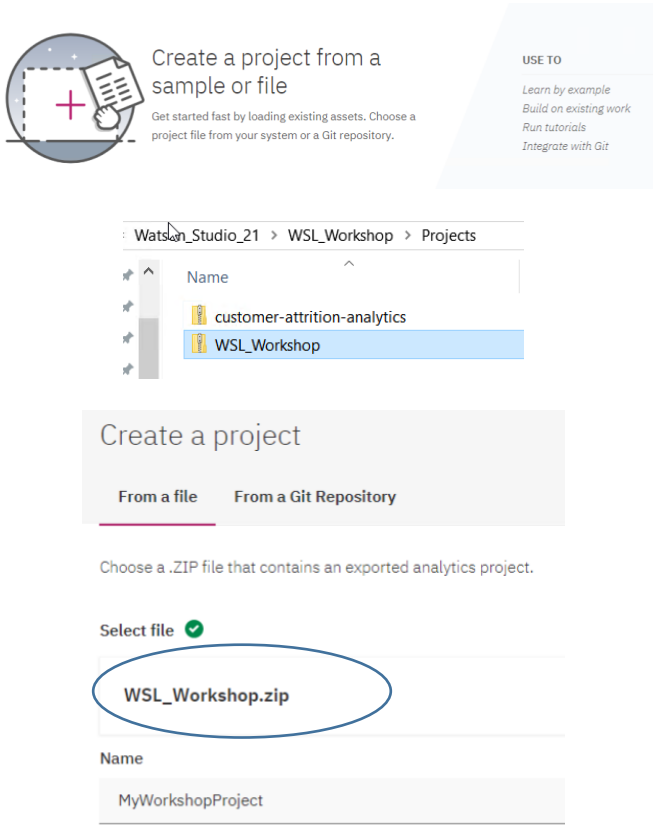


In the lab we will refer to this folder as the *git repo* folder.

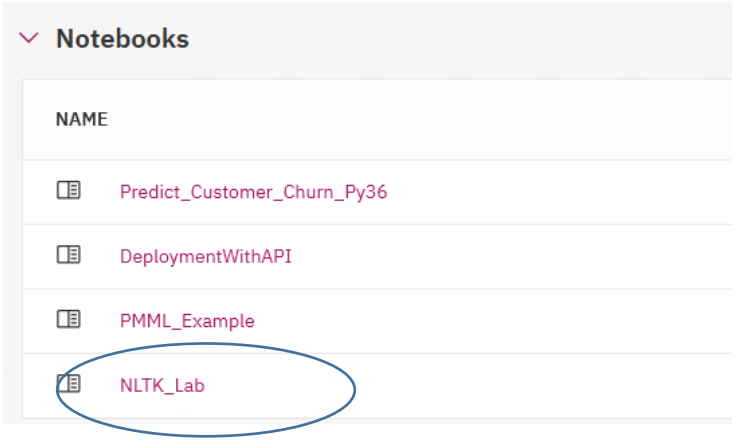
Part 1: Text Analytics in Open Source

1. In Watson Studio create a new project from file – *Watson_Studio_21.zip*, located in the *git repo\Projects* folder. You can use any value for the project name.

   ***Note**: If you already created this project in another lab, you can skip this step.*



2. Navigate to the **Assets** view and open the *NLTK* lab in *Edit* mode.



3. Review and run the notebook.

Part 2: Text Analytics in SPSS Modeler

In this section you will learn how **SPSS Modeler** can be used to convert unstructured data into a format that can be used for building and scoring classification models.

We will use a component of SPSS Modeler, Text Analytics, to complete this task. Text analytics in Modeler can be used to perform various text analytics tasks, not just unstructured to structured data conversion. You can learn more about SPSS Text Analytics here:
https://www.ibm.com/support/producthub/icpdata/docs/content/SSQNUZ_current/wsd/nodes/_nodes_TA.html

Use Case

Goal: Identify who is likely to respond to a marketing offer.
Approach:

- Use a data extract from a CRM
- Extract concepts from open ended comments in a customer survey
- Define which fields to use
- Choose the modeling technique
- Automatically generate a model to identify who is likely to respond
- Review results

Why?

- Target those likely to respond to offers to increase revenue, cut costs
- Using unstructured data improves modeling accuracy and provides more insight

Customer Reference

A telecommunications provider in the United States uses predictive modeling of customer data to increase revenue by billions and reduce its customer churn rate to less than 1 percent, lower than any of its competitors.

Business challenge

Reactive and reflective marketing strategy is giving way to predictive modeling. One wireless communications company in the United States knew customer churn was hitting its bottom line and began looking for a solution that would enable it to deepen its customer focus by proactively targeting customers more prone to churn.

IBM's implemented solution:

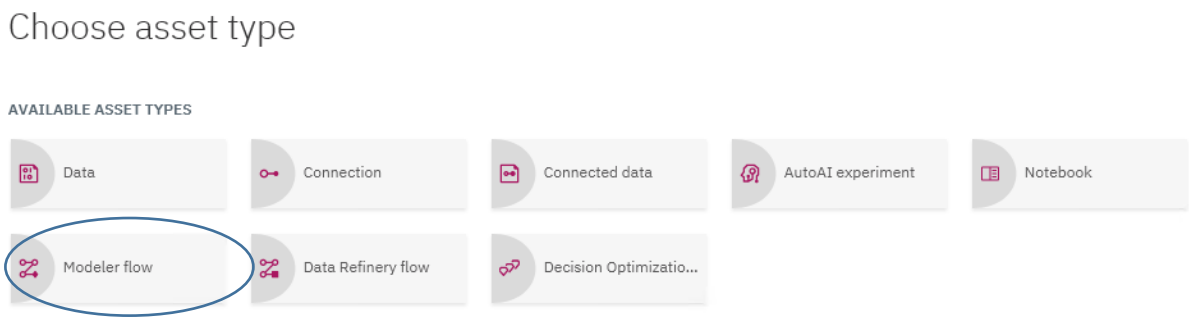
Reduced customer churn by two-thirds to 0.94 percent, the lowest churn of any wireless provider in the country.
Grew company revenue by 7 billion in one year, a 10 percent increase.
Increased modeling accuracy for data by as much as 12 percent.
Enabled the company to evaluate more than 450 variables for predicting customer defection within 90 days.

Customer quote: *Enhanced predictive modeling not only helps us retain valued customers, but also helps us do it in a way that best preserves and enhances company profitability and alignment with our business goals.*

Create a Text Analytics Modeler Workflow

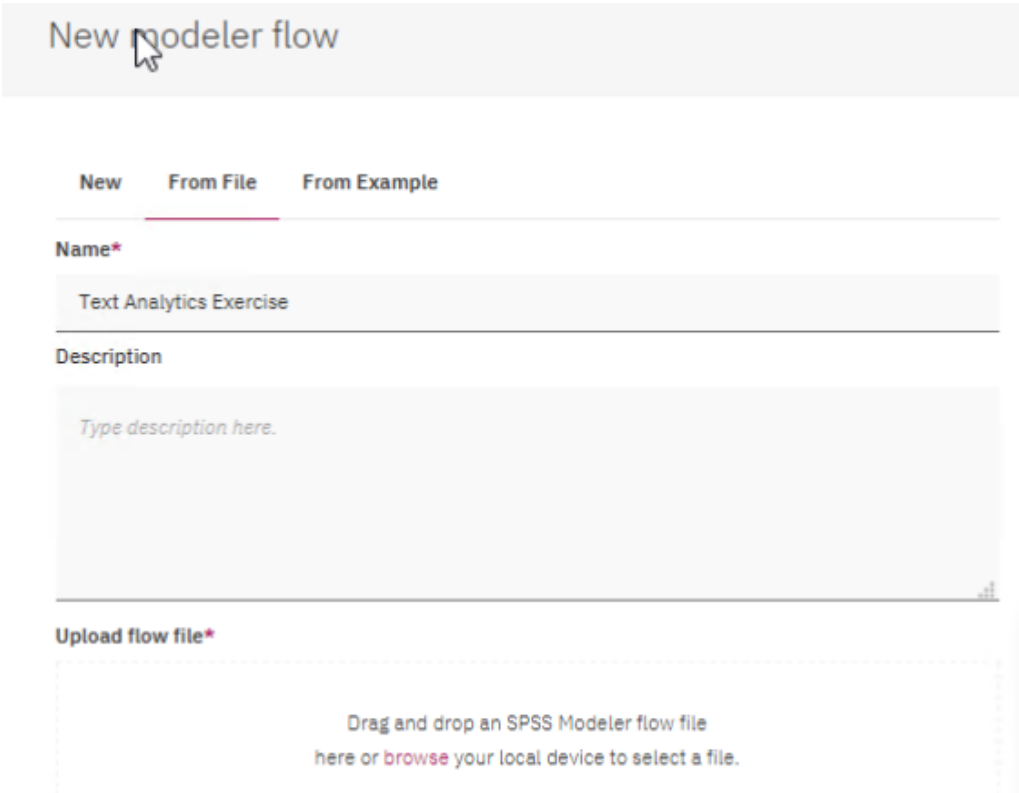
1. In **Watson Studio** navigate to your *WSL_Workshop* project.

2. Click **Add to Project -> Modeler flow**

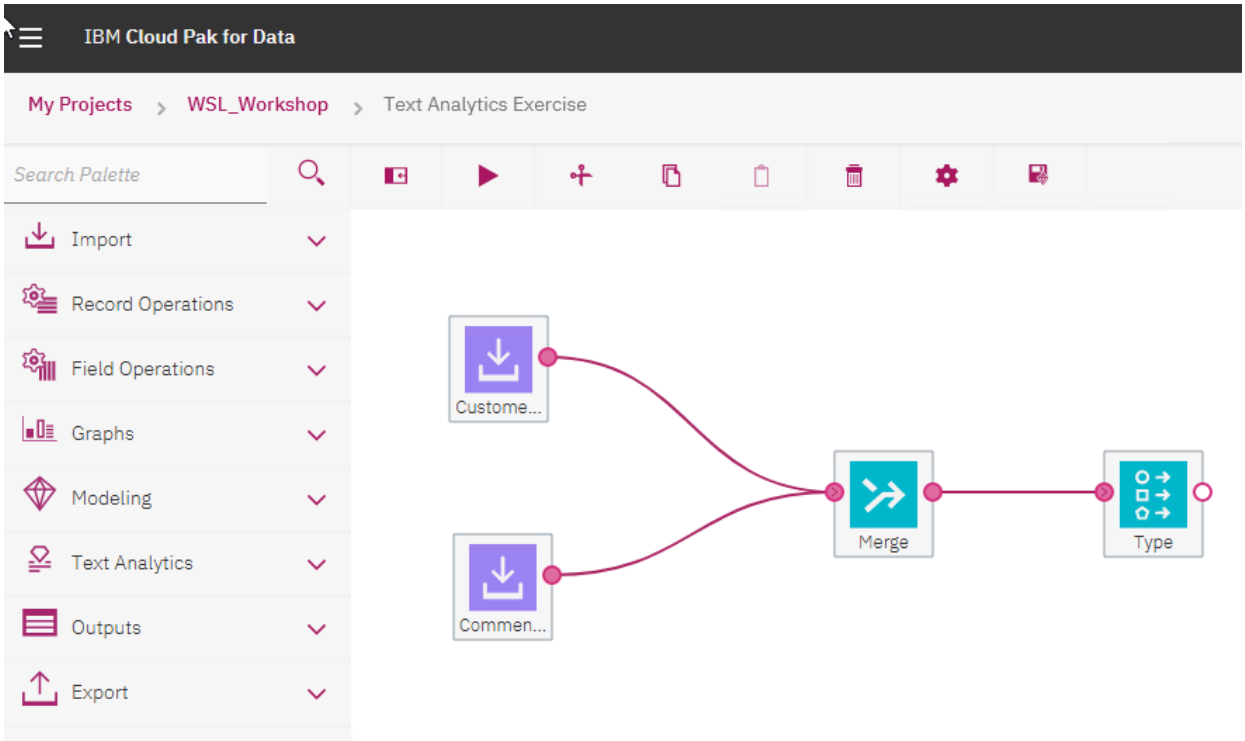


3. Select **From File**, browse to *git repo/Flows* folder, and select *Text Analytics Exercise.str.* You can provide any name for the flow.

   Click **Create**.

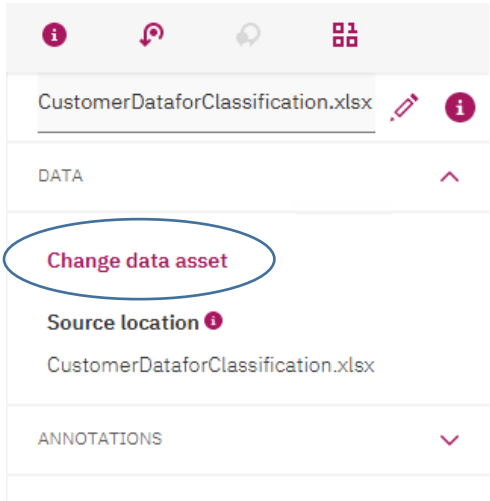


We have a partially constructed Modeler flow. Let's configure the data sources that are already loaded in the project.

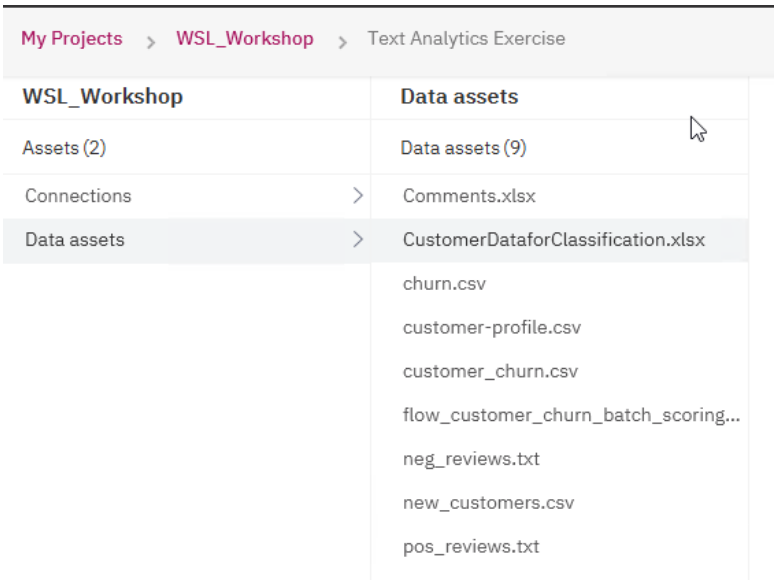


4. Double click on the *Customer* node (purple node).

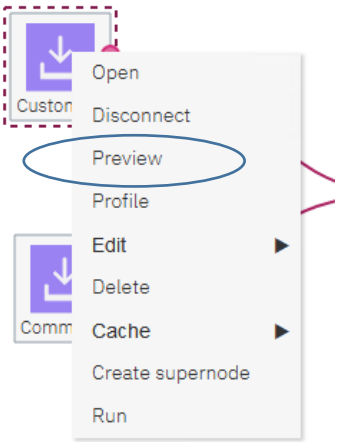
5. In the **Settings** view click **Change data asset**.



6. Select *CustomerDataforClassicication.xlsx*. Click **OK**, then **Save** in the **Properties** view.

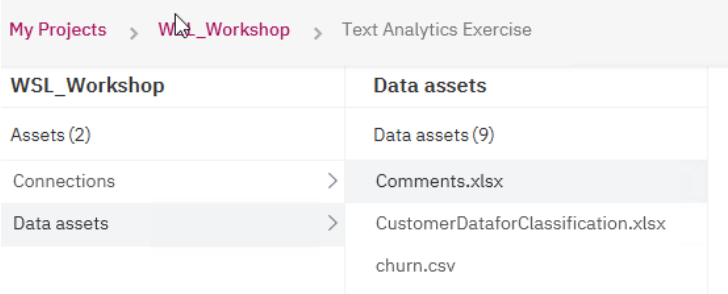


7. Right click on *Customer* node and select **Preview** to verify that you can access data.



ID	Sex	Region	Children	Est_Income	Car_Owner	Status	Paymethod	LocalBilltype	Customer_Segments
1.000	F	2.000	1.000	38000.000	N	S	CC	Budget	High Income Families
6.000	M	3.000	2.000	29616.000	N	M	CH	FreeLocal	Low Value and No Kids
8.000	M	1.000	0.000	19732.800	N	M	CC	FreeLocal	High Income Families

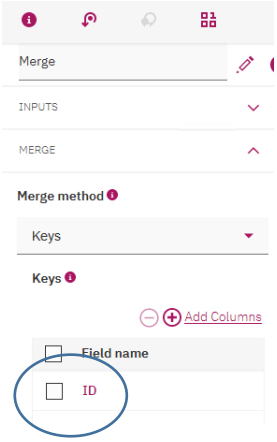
8. Repeat the same steps to configure the *Comments* input node. Select the *Comments.xlsx* file as the data asset.



Next, we'll review the stream.

9. Double click on the **Merge** node and select the **Merge** tab.

   Notice that the two data sources on the canvas are being joined by a common key, *ID*. Close the **Merge** node using the **Cancel** button.



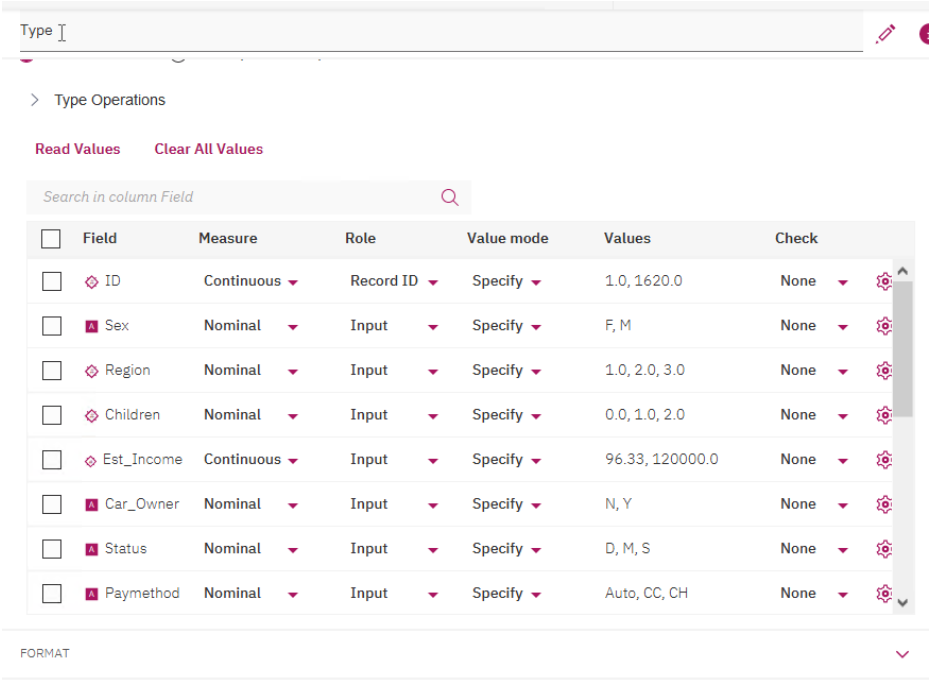
10. Right mouse click on the **Merge** node and select **Preview**.

    Scroll to the bottom and then right. Customer comments have been added to each customer record (joined by *ID*).

ustomer_Loyality_Code	Number_Of_Transactions_Current_Year	Response	Age_BIN	Comments
000	12.000	Responded	1.000	little, light
000	15.000	Did Not Respond	3.000	Battery life. Portability. Accessories. Style.

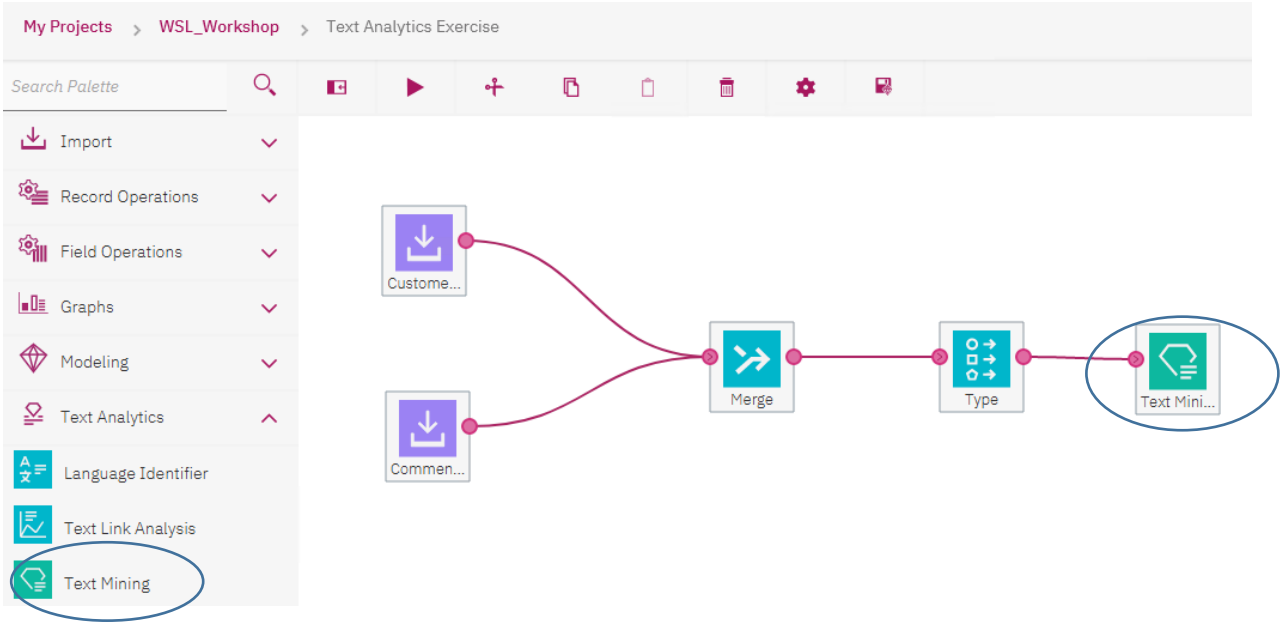
11. Double click on the **Type**.

The **Type** node was added to the canvas to instantiate the data types and define the measurement level and role of each field in the analysis.



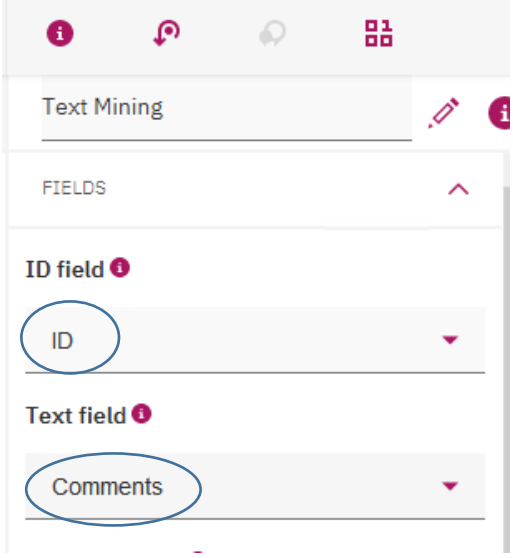
Click **Cancel** to close the **Type** node.

12. From the **Text Analytics** palette, add the **Text Mining** node to the canvas, and connect it to the **Type** node.

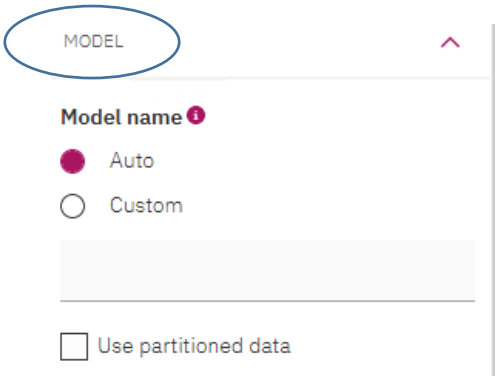


13. Double click on the **Text mining** node.

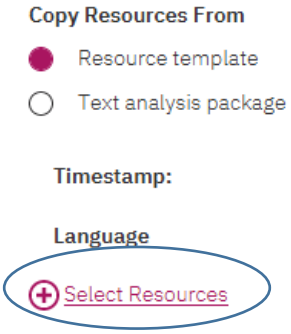
Select *ID* in the **ID field** dropdown and *Comments* in the **Text field** dropdown.



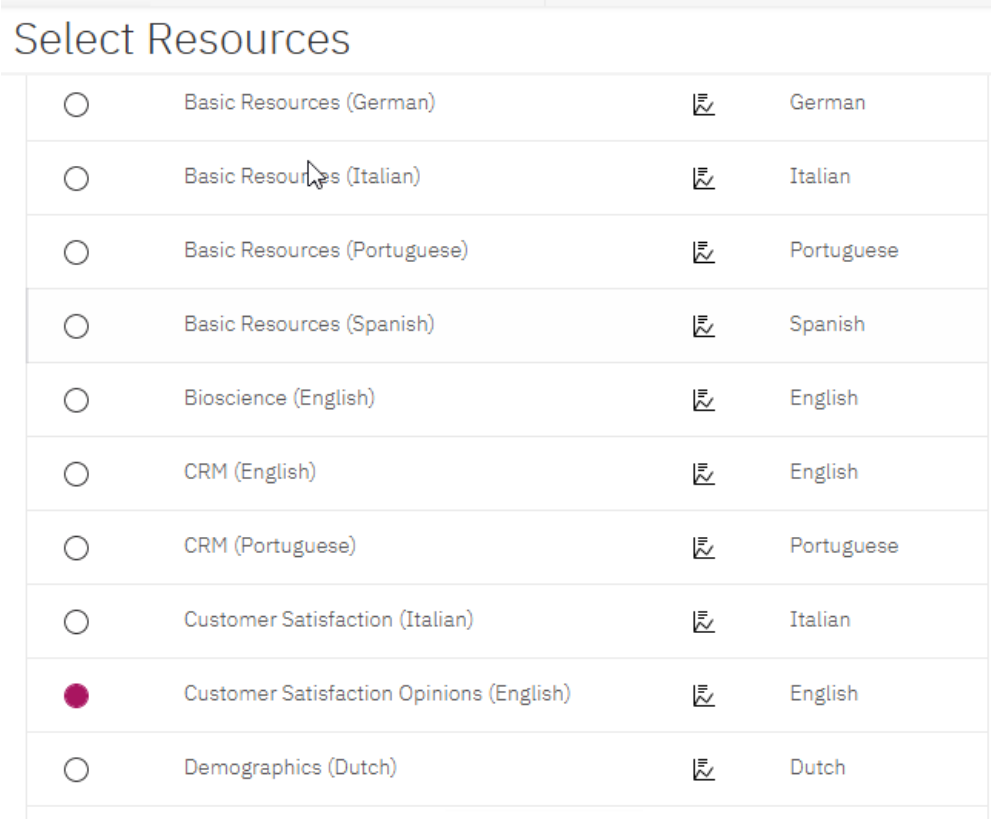
Select the **Model** tab.



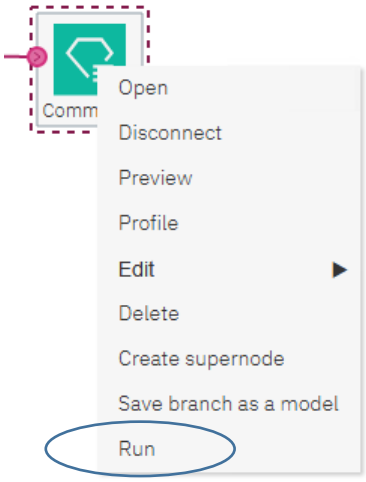
Scroll down to the **Copy Resources From** section. Click on **+Select Resources**.



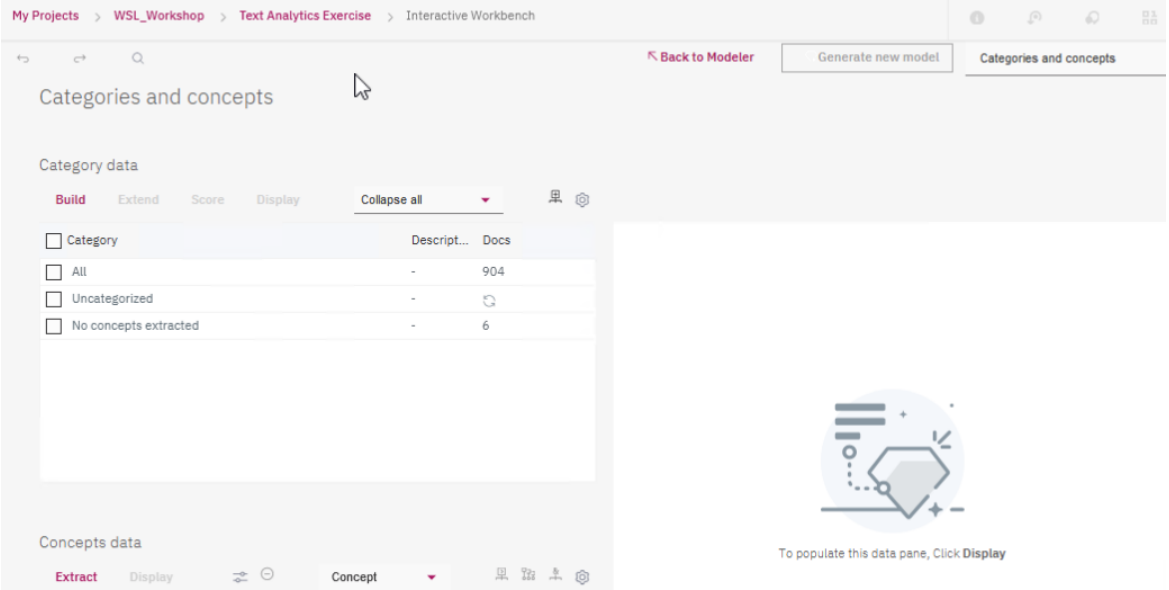
Select the *Customer Satisfaction Opinions (English)*, resources template._This will load pre-built resources into the text mining process. Click **OK** and then **Save**.



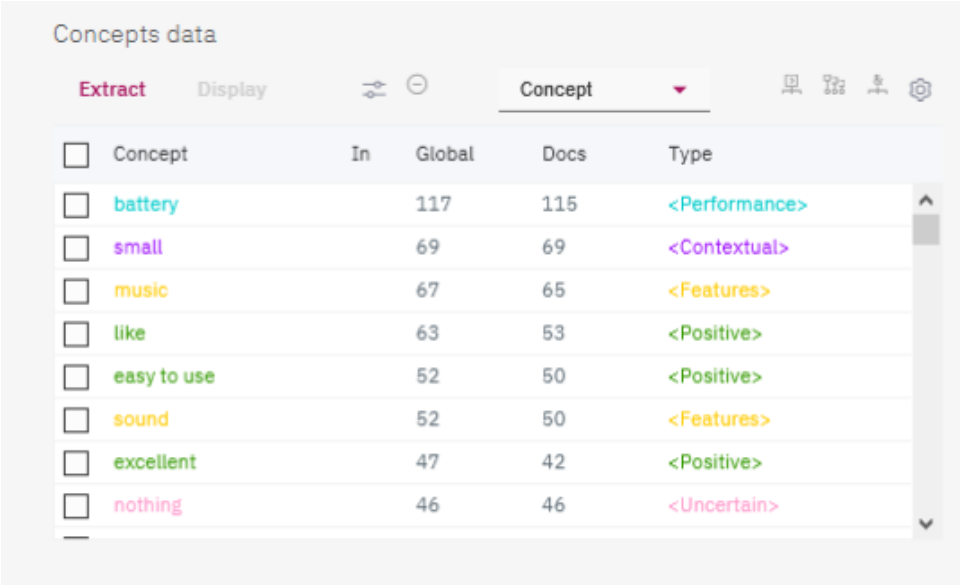
14. Double click on the **Text mining** node now labeled *Comments* and select **Run**.



Once the libraries and resources are loaded and the extraction process is complete, the **Interactive Workbench** is displayed.



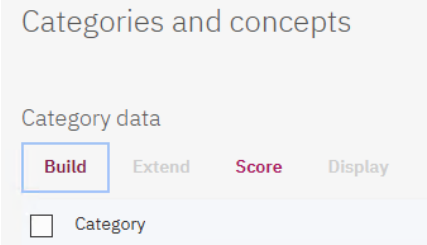
The list of extracted concepts is displayed in the lower left panel of the interface. These are not just words, phrases, or character strings which were matched to some search criteria. They are ***concepts*** and **types**, extracted and tagged, using Natural Language Processing (NLP), through reference to a comprehensive collection of libraries, provided with Text Analytics canvas for Watson Studio Desktop & Cloud Pak.



The concepts and types will be used as the basis for building the categorization model. While in practice, Text Mining is an iterative and interactive effort, for this workshop, we will run the text analysis engine without making any changes to the defaults.

Important Note: at this time, the Template Editor, which is available in SPSS Text Analytics is not available in Watson Studio. This means that you do not currently have the capability to make changes and updates to the existing libraries.

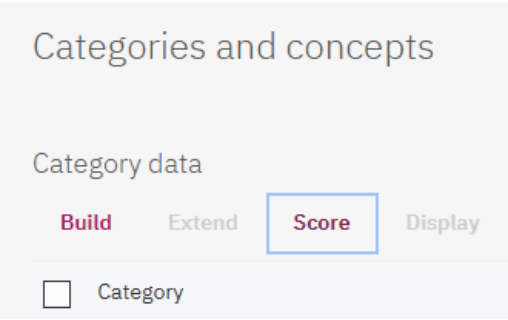
15. To build categories (a *Taxonomic Classification*), click on the **Build** button in the upper left quadrant of the screen.



The tool uses a *Semantic Network Algorithm* to create a multi-level Taxonomy based upon the *concepts* and *types* extracted during the process.

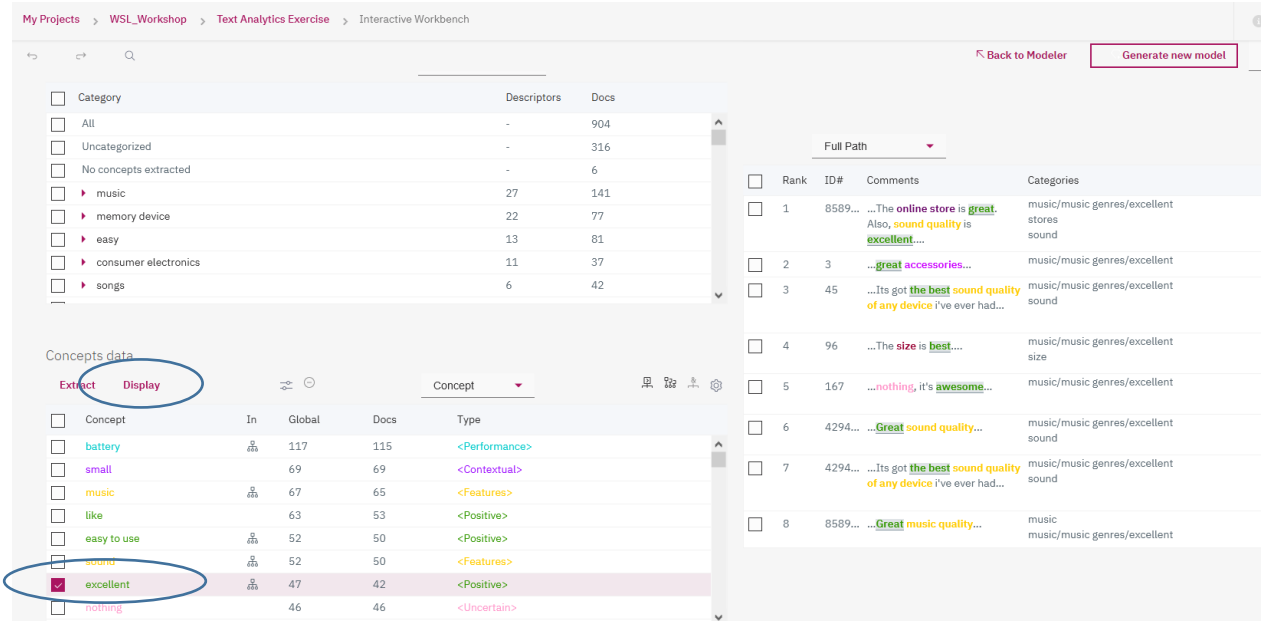
After the taxonomy is created, the process of scoring the records needs to occur. Scoring is the process of comparing the business rules from the taxonomic leaf(s), to the extracted concepts and types. Once compared, if the business rule fires, a classification designation is added to the row of text that is processed.

To score the data, select the **Score** button in the top left-hand corner of the screen.



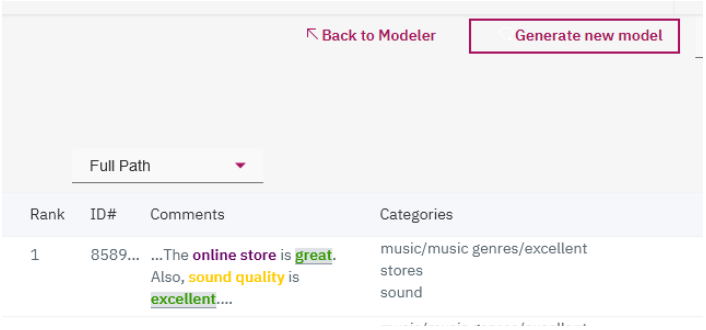
Note the new numbers in the **Docs**, column after scoring.

To display the concept data in the righthand pane, select one of the **Concept** check boxes and then select the **Display** button.

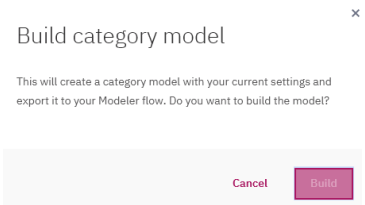


Once the build and scoring processes have completed, the user reviews the results, and works with the linguistic resources and category definitions to ultimately arrive at a set of categories, which are both useful and meaningful to the analysis. However, for purposes of this workshop, we will proceed with the categories as they are now.

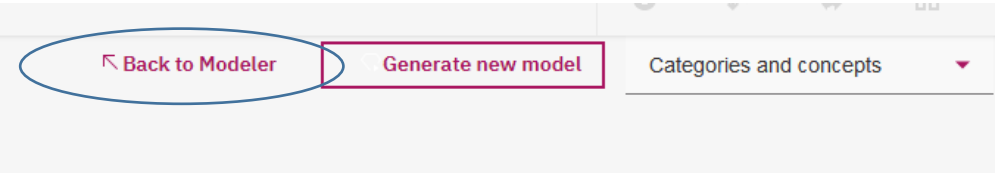
16. Click on the **Generate New Model** button at the top right-hand side of the page.



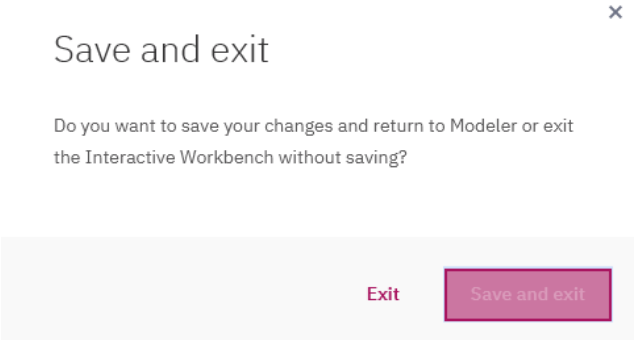
Then select **Build** for the build category model.



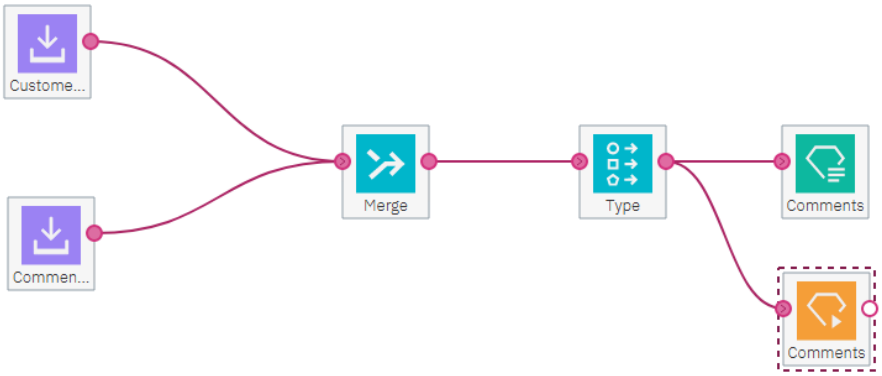
Once the model finishes building, click **Back to Modeler**.



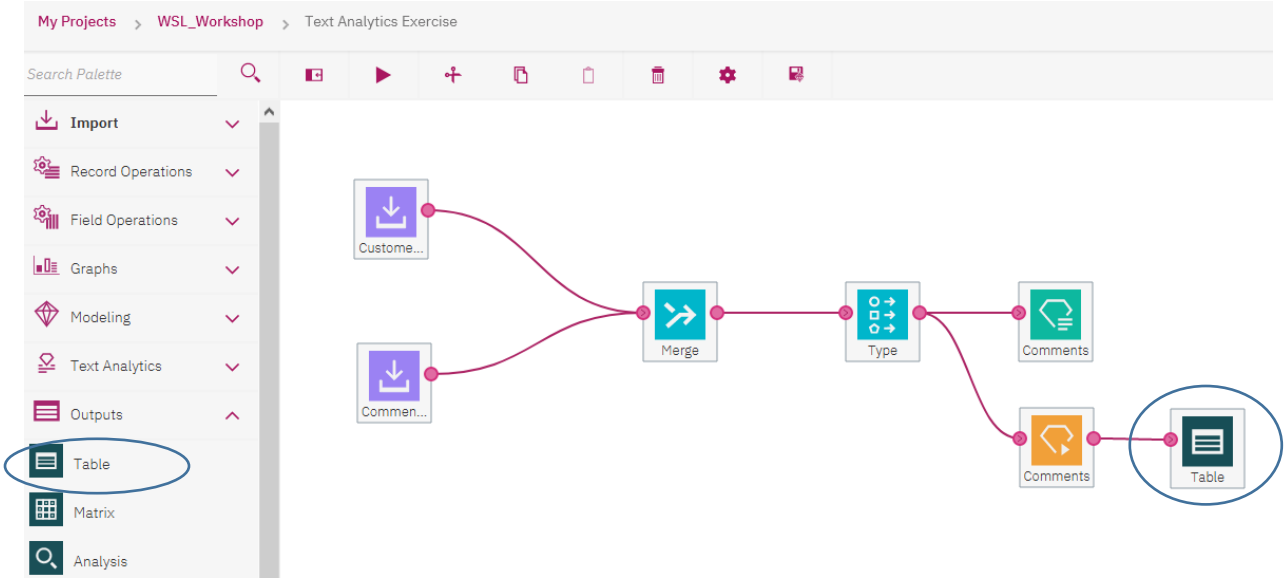
Cick **Save and Exit**.



We now have the *Category* model on the canvas (the orange nugget), that was automatically connected to the **Type** node.

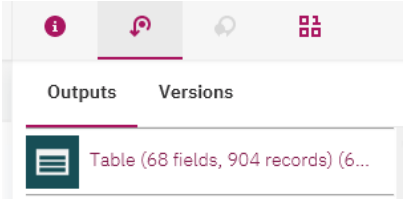


17. Add a **Table** node to the *Comments* model nugget. The **Table** node is located on the **Output** tab.



18. Right click on the **Table** node and select **Run**.

Results are accessed by clicking on the **Output** icon.



Double click on the **Table** in the **Output** view.

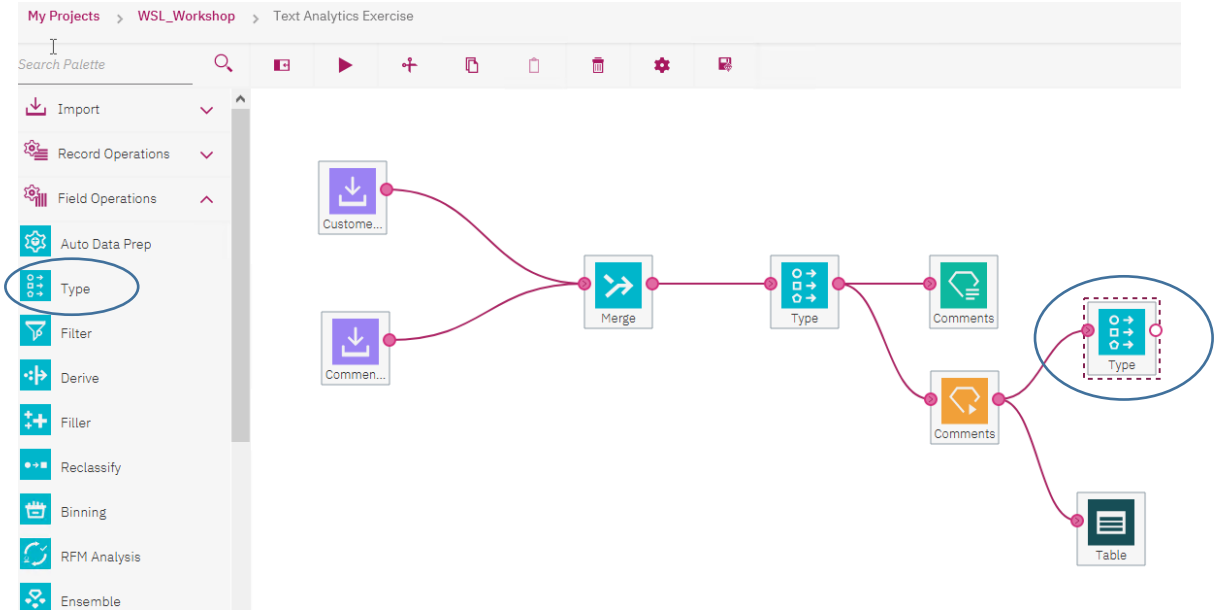
Notice that categories that were created by the text analytics model are now added to the original dataset. Value *F* means that the category was not present in the comment, and value *T* means that it was present.

Unstructured comments have been converted to structured data, and now it can be used for creating a classification model.

Category_and munitions	Category_and munitions/battery	Category_capacity	Category_car	Category_clothes items	Category_clothes items/button	Category_color	Category_colour
F	F	T	F	F	F	F	F
F	F	F	F	F	F	F	F
F	F	F	F	F	F	F	F
F	F	F	F	F	F	F	F

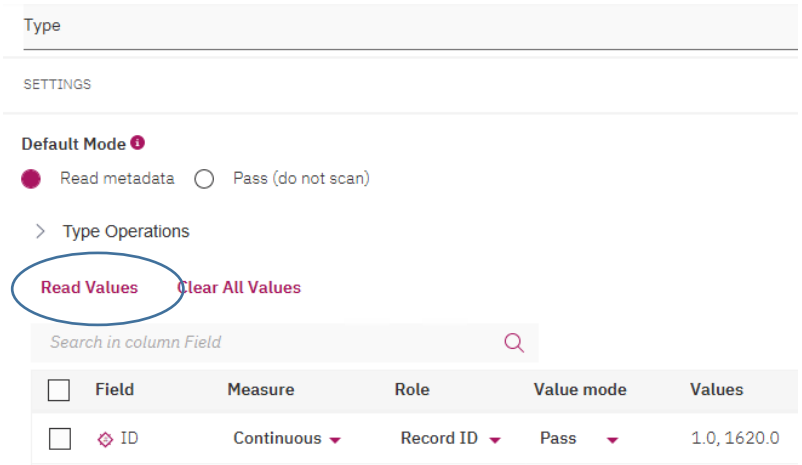
To use the newly created categories for modeling, we need to use a **Type** node**.**

19. Add a **Type** node to the stream (located on the **Field Operations** tab) and connect it to the *Comments* model nugget.



20. Double click on the **Type** node.

    Click **Read Values**.

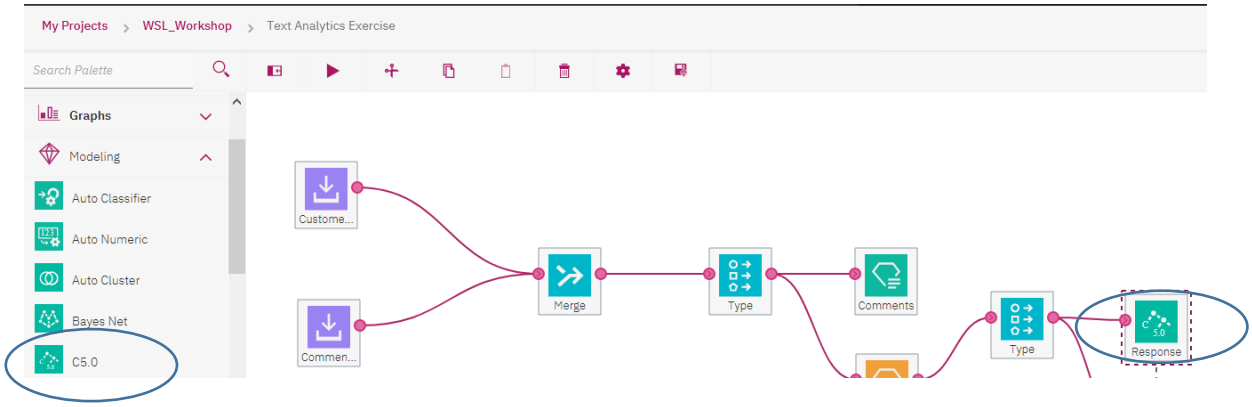


Notice that many fields generated by the text mining model can now be used as input fields for a predictive model.

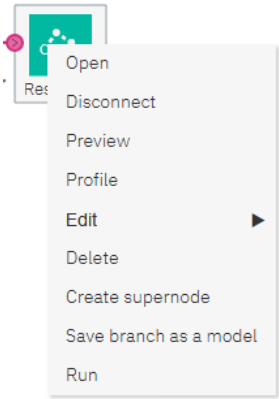
Click **Save** on the **Type** node to close it.

21. Add the **C5** node (from the **Modeling** tab) node to the last **Type** node.

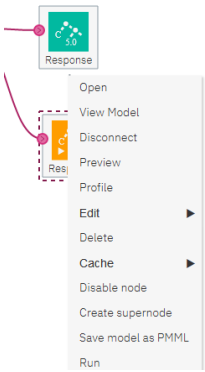
    C5 is one of the frequently used algorithms in Modeler. *C5* is a decision-tree based algorithm. It's especially useful if you want to understand how each field influences the prediction.



22. Right click on the **C5** node and select **Run**.

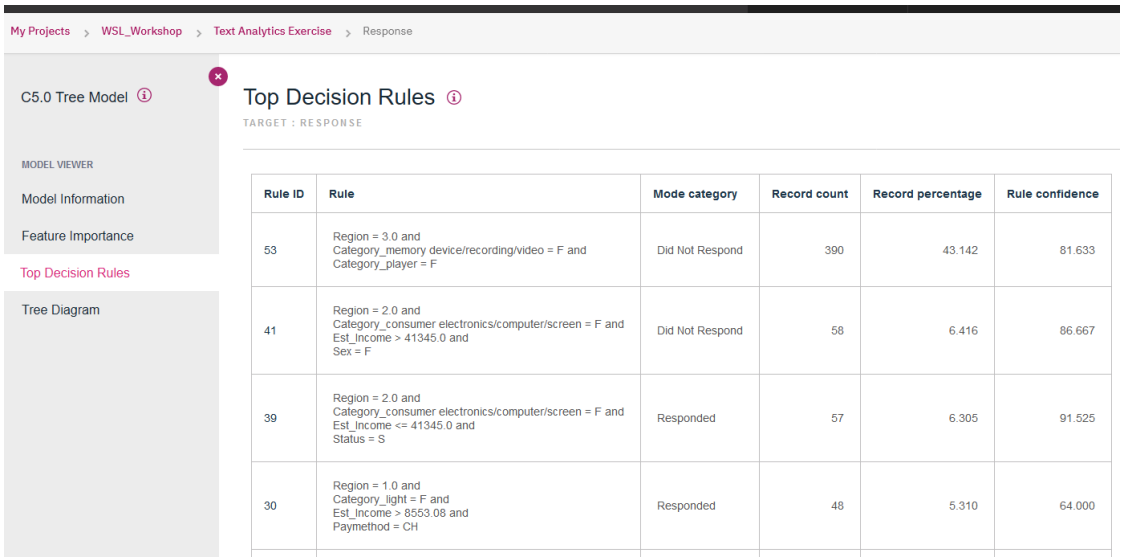


23. Right click on the generated *Response* model and select **View Model**.



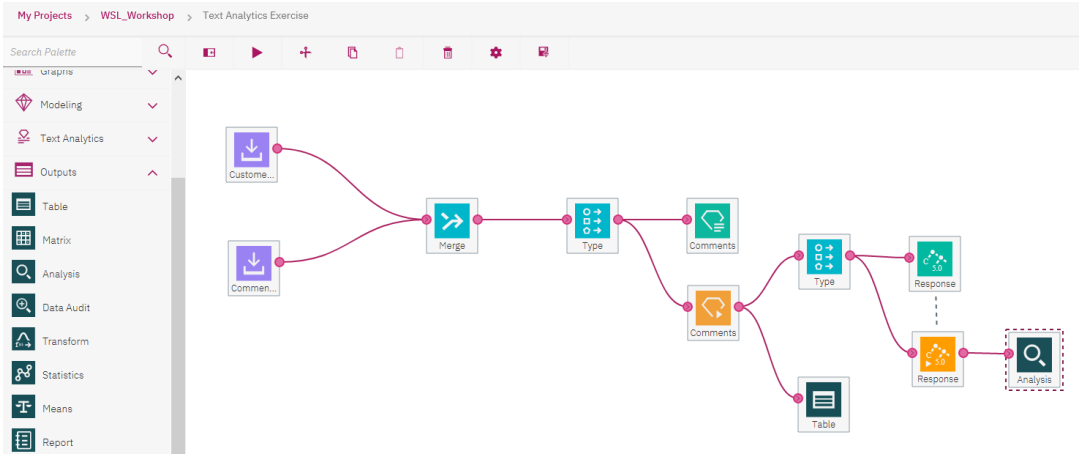
24. Click on **Top Decision Rules** tab.

Notice that the generated categories are used in several decision rules in the model.



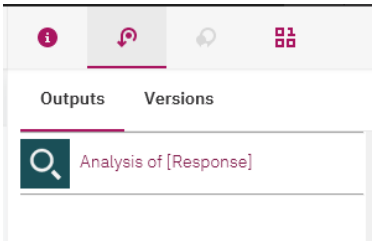
Rule ID	Rule	Mode category	Record count	Record percentage	Rule confidence
53	Region = 3.0 and Category_memory device/recording/video = F and Category_player = F	Did Not Respond	390	43.142	81.633
41	Region = 2.0 and Category_consumer electronics/computer/screen = F and Est_Income > 41345.0 and Sex = F	Did Not Respond	58	6.416	86.667
39	Region = 2.0 and Category_consumer electronics/computer/screen = F and Est_Income <= 41345.0 and Status = S	Responded	57	6.305	91.525
30	Region = 1.0 and Category_light = F and Est_Income > 8553.08 and Paymethod = CH	Responded	48	5.310	64.000

25. Add the **Analysis** node (from the **Output** tab) and connect it to the **C5** *Response* model.



26. Right click on the **Analysis** node and select **Run**.

27. Click on the **Output** icon, then double click on the **Analysis** output.



The results from the Analysis node show that the mode accuracy is 83.3%.

My Projects > WSL_Workshop > Text Analytics Exercise > Analysis of [Response]

Results for output field Response

Comparing $C-Response with Response

Correct	753	83.3%
Wrong	151	16.7%
Total	904	

In this lab we have shown the development of an SPSS Text Analytics model. If this model is deployed for scoring, *comments* can be passed in to SPSS flow deployment in unstructured format (the same format that was used during model building). This is possible because SPSS supports the deployment of the entire flow, and that means that during scoring unstructured data will first be converted to structured, and then passed in for scoring with the rest of the fields.

You have finished the **SPSS Text Analytics** lab.