PMML Model Prediction

Overview

Yellowfin offers a PMML transformation step that allows users to import data science models stored in a PMML file format into Yellowfin and then apply them on data to generate results, such as predictions. The step works within Yellowfin's data transformation module by creating a transformation flow.

In order to use this transformation step, you will first need to install a PMML plug-in in your BI platform. Once that has been installed, you will be able to see a new PMML step in the data transformation module's transformation list. This transformation step will be used to integrate and apply your PMML model. Click here to learn more about installing plugins.

> ⊘ It would be ideal to get a better understanding of the data transformation module before continuing with this guide.

Required Transformation Steps

The data transformation module creates transformation flows by using different "steps". For this procedure, it will require *at least* 2 types of steps:

Note: You can add more steps to your flow, but require at least the first two for the PMML model to generate results.

- An **input step** to extract data from a data source and load it into Yellowfin. This is the data that you want to use for your model, so ensure that you have access to the data source, and that the data is in the correct format. There are various ways of extracting data using the different types of input steps, all of which are supported for the PMML step. Refer to this section to understand how to use and configure these steps.
- A **PMML transformation step**. This step will generate the output that your PMML model is designed to produce (such as making predictions based on your data). This works by loading your model (saved in a PMML file) into the Data Transformation module and applying it to the data from the previous step (i.e. the one it is connected to). You will need to configure this step to match input fields and specify the output fields. (**Note:** This step only appears in the transformation step list once you have installed the required PMML plug-in into Yellowfin.)

> ⚠ An output step will be needed if you wish to store the result of the PMML model in a writable database.

The Data Transformation module lets users create different types of transformation flows and include the PMML step when required. For instance, you can apply the PMML model on data that has undergone other transformations, and even continue to use further transformations on the result it generates.

Supported Model Types

Yellowfin's PMML transformation step supports the following types of data science models:

- Association rules
- Cluster model
- General regression
- Naive Bayes
- K-Nearest neighbors
- Neural network
- Regression
- Rule set
- Scorecard
- Support Vector Machine
- Tree model
- Ensemble model

Supported PMML File Versions

The PMML transformation step fully supports PMML files of versions 3.0 and above. However, some earlier file versions might also work. In addition to that, even though PMML standard allows for multiple models within one PMML file, the PMML step will consider only one model in a PMML file. If there is more than one model described in a file then Yellowfin will use the first one.

Configuring the PMML Step

When the PMML transformation step is added to the flow, it will need to be configured. This involves uploading the PMML model, and once it has been properly parsed, configuring the input and output fields specified in the model. Note: It is assumed that the model being used, has specified input and output fields.

Input Configuration

The model's input refers to the fields of data that are required to conduct calculations. For example, in case of a regression model, the input is a list of independent variables required for the model. To properly execute the PMML step, we will need to map the data fields to fields required by the model.

Output Configuration

In turn, the model's output is a set of result fields that the model is designed to generate. In some cases the output can consist of only one field, such as in the case of a regression model where the output is usually one predicted field. But on the other hand, a clustering model can have either a specified number of clusters as an output, or fields with distances to centroids of each cluster. When configuring the output, the user can choose which of the available fields they wish to see in the result, but at least one output field should be selected. The datatype of the output field will be whatever has been specified in the model, however if a type isn't specifically defined, then the result will be considered as Text.

Taking an example, a model can be designed to predict the fare paid by passengers of the Titanic based on their travel class and age. In this case, age and class will be the configured inputs and the resulting field selected would be fare.
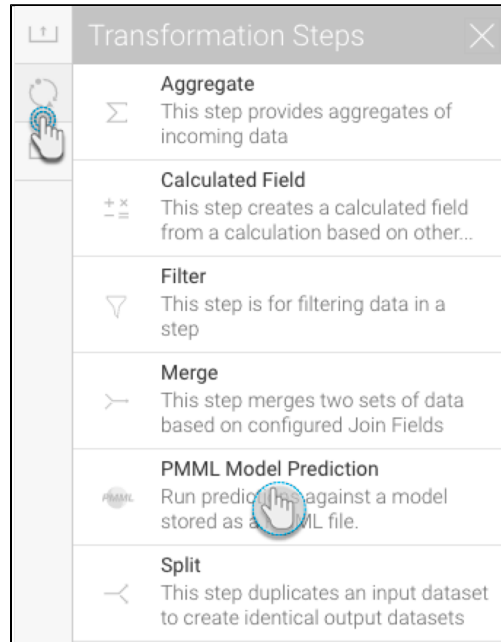
How to use a PMML Transformation Step

This guide shows you how to integrate your data science model using the PMML transformation step into Yellowfin.

1. Build a model in PMML and save it in .xml format. (This file will be referred to as the **PMML file** in this guide.)
2. Navigate to Yellowfin's Data Transformation module. (Create button > Transformation Flow)
3. Create a transformation flow beginning with an input step to extract data. (Click here if you want to learn how to create a basic flow, or here to learn about the different ways data can be extracted).
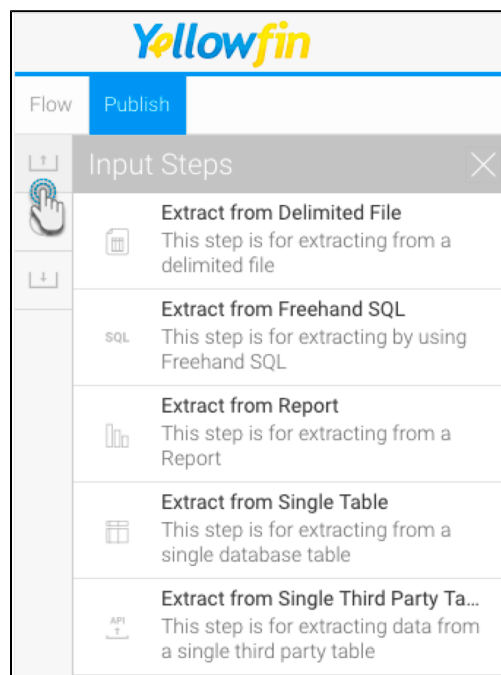

4. The extracted data will appear in the data preview panel. You can add more steps to further transform the data.
5. Once you are ready to use the PMML model, follow the procedure below.

6. Drag the PMML model prediction step from the transformation step list onto the canvas. (**Note:** If you don't see this step, ensure that you have installed the PMML plug-in.)

7. Using a connection point, create a connection from the last step to the PMML transformation step.
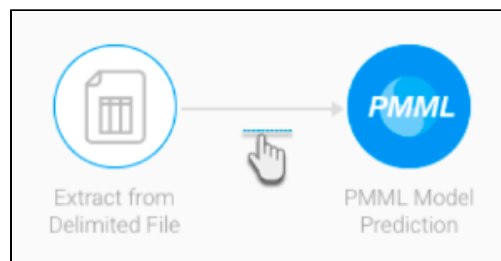
8. Next, you need to configure the PMML step. (Make sure it is selected to bring up its configuration screen.)

9. Upload your PMML file. There are two ways of doing this.
   a. Drag the file into the specified panel, or browse and select it. (This is the preferred method if the file is located on your local machine.)
   b. Or enable the *Load from Path* toggle and load the file by providing its path or URL. Then click on the Load File button.

   c. Provide the full URL to access the PMML file, for example: *http://dmg.org/pmml/pmml_examples/knime_pmml_examples /ElNinoLinearReg.xml*
   (**Note:** To be able to load a PMML file using a URL, you would need an internet connection, otherwise the file will not load.)

   Example of a full file path on a Windows machine is given below:
   C:\Users\admin\Desktop\audit_cluster.xml

   Example of a file path on a Linux machine is given below:
   /root/audit_cluster.xml

   Note: If the file cannot be parsed, or if the path provided is invalid, then the following error message will appear.

Note 2: When using the drag and drop method to upload a file, the file name might disappear when you first upload it, but will reappear once an action on the step is performed.

10. Once the file is successfully uploaded, detailed configuration options will appear to map the input fields and specify the output. The input and output fields that appear will be dependent on the model.

11. First, match the input columns required by your model with the appropriate columns in the extracted file. For example, our model requires input in the form of income, age and gender.
    Note: It is essential to make the correct mapping here in order to produce the correct result. The data fields may not have the same names as they do in the model. You should ensure that the selected fields match the datatype that the model is expecting. (If an incorrect mapping is made, the Errors tab will generate errors for each of the data values.)

Configure Model Input

Income	Income ▼
Age	Age ▼
Gender	Gender ▼

Note 2: All the input fields are mandatory for mapping, the step will not run unless all these are properly matched.

12. Next, select which model output result you would like to generate. The configure panel will display all the possible outputs that your model is designed to produce. You can choose one or more outputs (however at least one output field must be selected).
Note: The "Default Target" output field contains the default output of the model. If the PMML file does not explicitly specify an output field, then only the "Default Target" option will be available.

Configure Model Output

Fields to Extract
☑ predictedValue
☑ clusterAffinity_1
☐ clusterAffinity_2
☐ clusterAffinity_3
☐ Default Target

13. Once the input and output fields have been properly configured, click Save. (**Note:** If you click anywhere outside of the configure panel after configuring the model and before clicking Save, you will lose the configuration details.)
14. On doing so, columns(s) with the model's result will appear in the data preview panel. (**Note:** Since we selected two output fields, therefore our example shows two new result columns.)

> ⚠ The datatype of these output columns will depend on what has been specified in the model. However, if not specified, then the system will use Text as the default type.

	clusterAffi... TEXT	predictedV... TEXT	ID NUMERIC	Age NUMERIC	Employment TEXT	Education TEXT	Marital TEXT	Occupation TEXT	Income NUMERIC
1	2.08103...13...	3	1004641	38	Private	College	Unmarried	Service	
2	1.935939290450...	3	1010229	35	Private	Associate	Absent	Transport	
3	4.744540086232...	2	1024587	32	Private	HSgrad	Divorced	Clerical	1546
4	0.505548740580...	1	1038288	45	Private	Bachelor	Married	Repair	277

15. You can perform further transformations or save your result in a database.

Editing a PMML Transformation Step

You can edit the PMML step by either replacing the PMML file with one with a different model, or by changing the input or output configurations. However, this will affect your transformation flow.

In our example, we will display a very simple transformation flow, even though in most cases the flows will have a more complex structure.

1. As seen in the above example, the flow includes a PMML model that has been configured to map input values from a delimited file and generate outputs. It then aggregates the predicted value generated. The chosen model requires income, age, and gender as input values and has a few output fields.
2. You can edit this step by reconfiguring the input fields (for instance in case the input data gets changed). In such situations, the step will still be executed successfully if the new mapping is valid.
3. Or you can change the PMML model by loading another PMML file in the configure panel (ensure that the PMML transaction step is selected).
4. The step's input and output configuration fields will change to those of the new model.

5. Map the data to the new input fields (ensure that the datatypes match) and select the result to generate.
   Note: Make sure that the data coming from the previous step match the requirements of the new model. If not, you might need to change the data.
   Note 2: If the new model consists of any fields with the same name as the previous model, then those will be matched automatically.

6. Click Save.

7. If correct input fields are configured (that is the data types match), then the model will generate a new column with results, based on the selected output field (and remove the previously existing result columns).

	Predicted_... TEXT	ID NUMERIC	Age NUMERIC	Employment TEXT	Education TEXT	Marital TEXT	Occupation TEXT	Income NUMERIC	Gender TEXT	
1	-2614.2...14...	1004641	38	Private	College	Unmarried	Service	81838	Female	
2	-1020.38800635...	1010229	35	Private	Associate	Absent	Transport	72099	Male	
3	-1398.16790286...	1024587	32	Private	HSgrad	Divorced	Clerical	154676.74	Male	
4	-1972.85040695...	1038288	45	Private	Bachelor	Married	Repair	27743.82	Male	

8. However, now you will need to run the remaining steps in the flow again. The run logs will display that the remaining steps are not run (which in our example is an aggregate transformation step).

9. Click on the next step, ensure the configurations are accurate (if new output fields are generated, you will have to configure them again) and click Apply.
10. Repeat this for all the remaining steps in the flow.