Governed MLOps Workshop

# Training AI Models with Watson Studio

Document version: March 2023

DISCLAIMER

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.

The development, release, and timing of any future features or functionality described for our products remains at our sole discretion I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results like those stated here.

Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. **This document is distributed "as is" without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity.** IBM products and services are warranted per the terms and conditions of the agreements under which they are provided.

IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."
**Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.**

Performance data contained herein was generally obtained in controlled, isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.

It is the customer's responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products about this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. **IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.**

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

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

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

OpenShift is a trademark of Red Hat, Inc.

UNIX is a registered trademark of The Open Group in the United States and other countries.

# Table of Contents

## Training AI Models – Introduction

In this module, we focus on illustrating how to leverage the data science capabilities in Cloud Pak for Data to develop AI models. You will assume the role of the datascientist user (aka model developer) who typically trains and evaluates AI models and leverage the following data sets.

- CUSTOMER_PERSONAL_INFORMATION: This data captures personal information of the customers such as gender, marital status, income, age, and similar data.
- CUSTOMER_TRANSACTION_DATA: This data captures the transaction data for the customers.
- CUSTOMER_CHURN_LABELS: This data captures information about whether a customer did or did not churn.

The rest of this module is written with the assumption that the relevant data assets have been identified, virtualized, cleansed, quality-verified, published to the enterprise catalog and are ready to be consumed (as explained in earlier modules). However, if you haven't gone through the previous modules around data governance and data virtualization, you can continue this module and work with the relevant data sets from box (screenshots may not match exactly but the general flow still applies):

- customer_personal_info_simplified.csv
- customer_data_transactions.csv
- customer_churn_labels.csv

## AI Model Lifecycle Tracking

As explained earlier, AI Governance implies the ability to track an AI model through its lifecycle starting with a business use case of what the model should deliver all through model retirement when it is no longer needed. It is becoming more and more important to track AI models through all stages from initial exploration and development to operationalization in production where they deliver business value.

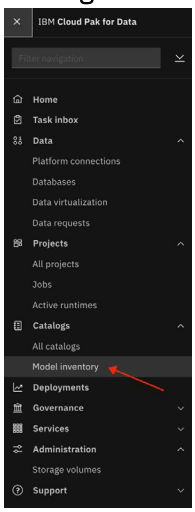
In the previous module, you've seen how to initiate a model entry and associated model in OpenPages and how that automatically integrates with the data science tooling to represent and track the model synchronously throughout its lifecycle. Cloud Pak for Data tooling specifically outlines the following stages for tracking AI models:

- ⇨ Develop: This stage captures the details and relevant information about the AI model as it is being developed by data scientists.
- ⇨ Test: This stage captures the details and relevant information about AI model when it moves from development to deployment, typically by MLOps engineers. At this point, the data science lead would have reviewed the performance of the model and approved its deployment.
- ⇨ Validate: This stage captures the details and relevant information about AI model validation, typically by an independent data science team to make sure the model meets the desired specifications in terms of quality, performance, fairness, and explainability.
- ⇨ Operate: This stage captures the details and relevant information about the AI model after it is approved and released to production where it can deliver business value to the organization.

To proceed with tracking AI models, you need to validate that a Model use case is already defined in the platforms Model inventory.

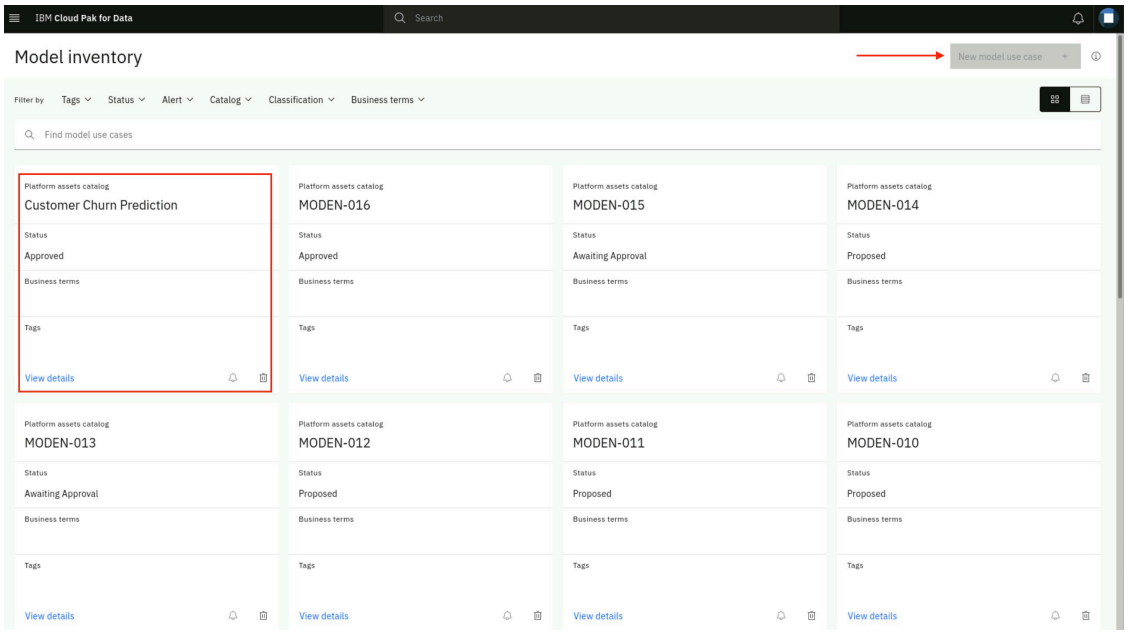
1- Log into Cloud Pak for Data as *datascientist* user.

2- Select Model inventory by clicking on the Navigation menu (top left hamburger icon) and selecting Catalogs → Model inventory (annotated with red arrow).

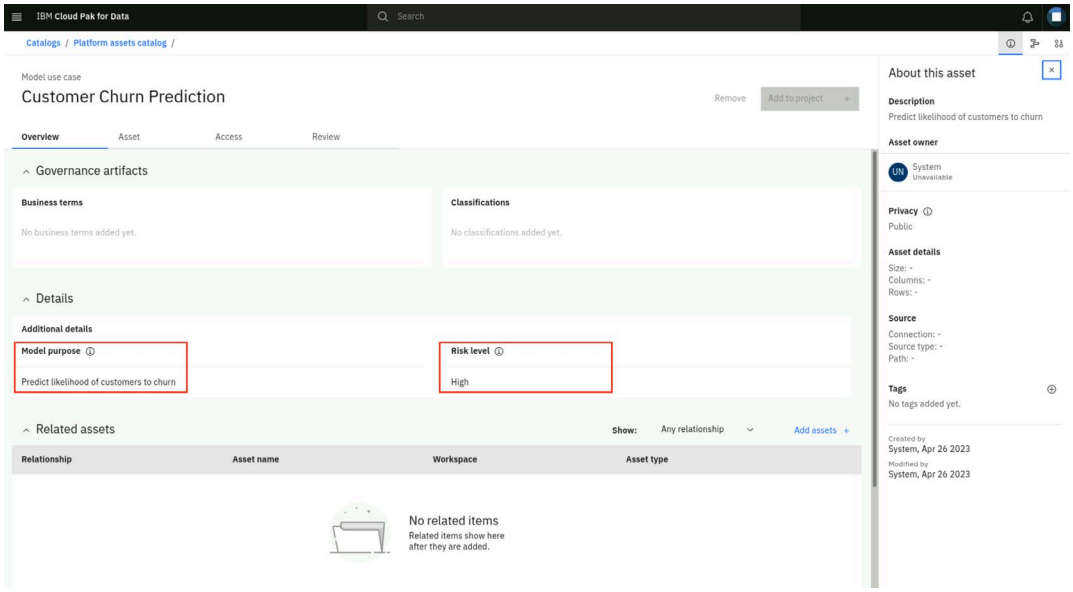



3- On the Model inventory page, review the set of existing model use cases. Specifically, note the Customer Churn Prediction model use case (annotated with red rectangle) which was automatically synchronized from the model entry you created in OpenPages in the previous module. The other model use cases have been pre-created for you and we'll review those in a later part of this workshop.

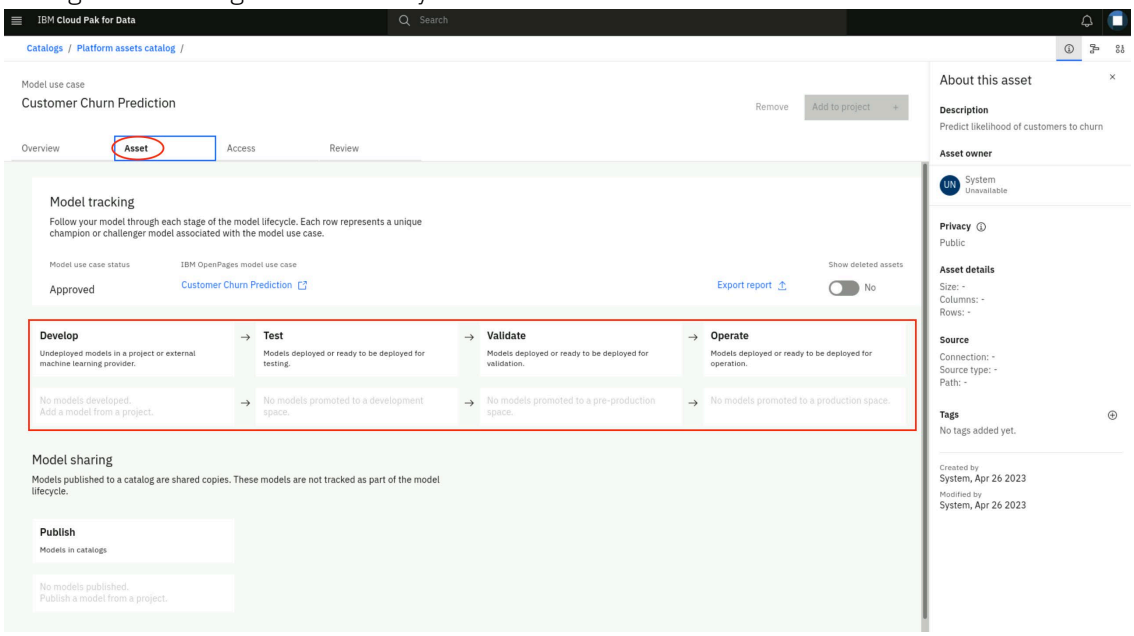
Note that the **New model use case** button (annotated with red arrow) is deactivated (grayed out) and this is because this environment has been configured to integrate with OpenPages. If for other scenarios, you would like the ability to create a Model Inventory independent of OpenPages, you can disable synchronization by clicking the Manage tab and disabling OpenPages integration (you need to be admin to do so; if you are logged in as datascientist user, you will not see the Manage tab on Model Inventory page).




4- Click View details link on the Customer Churn Prediction tile and take a couple of minutes to review the information on the Overview tab of the model use case such as **Model purpose** and **Risk level**. Note that these details were synchronized and you won't be able to edit them in this view. Any edits need to be made in OpenPages.



5- Click the **Asset** tab (annotated with red oval) and note the various stages, **Develop, Test, Validate, and Operate** (annotated with red rectangle) which will be auto-populated as the AI model(s) progress through various stages of their lifecycle.



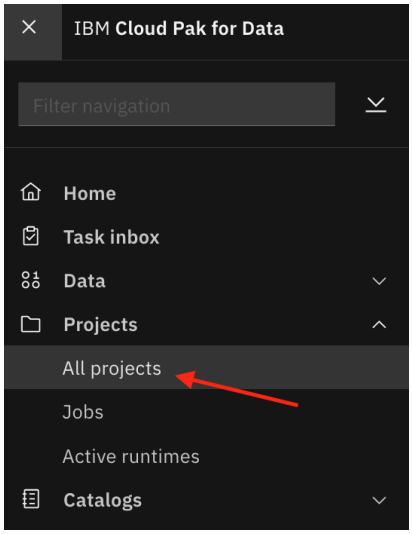

Now that you've validated the model use case is properly setup in the Model inventory, you can proceed to developing your AI models which you'll track through this model use case.

## Self-Serve Data Access

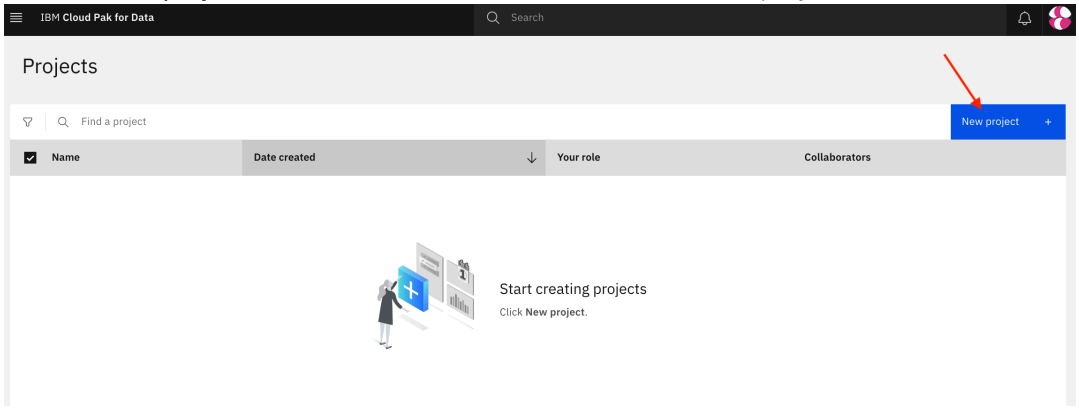
Several studies and analyst reports indicate that data scientists spend a significant portion of their time searching for data that may be relevant for the projects they are working on. One key advantage of Cloud Pak for Data is that it delivers a self-serve data access capability. This is achieved by enriching data assets across hybrid data sources with relevant business terms and data classes which resonate with data scientists and business users. Once enriched, the data is available via a semantic search interface to return most relevant data based on end user queries.

Once data scientists search for and evaluate different data assets, they can add relevant assets to their project which they would then use for data preparation and training AI models. In IBM Cloud Pak for data, a project is how you organize your resources to achieve a particular goal. A project allows for high-level isolation, enabling users to package their project assets independently for different use cases or departments. Your project resources can include data, collaborators, scripts, and analytic assets like notebooks and models.

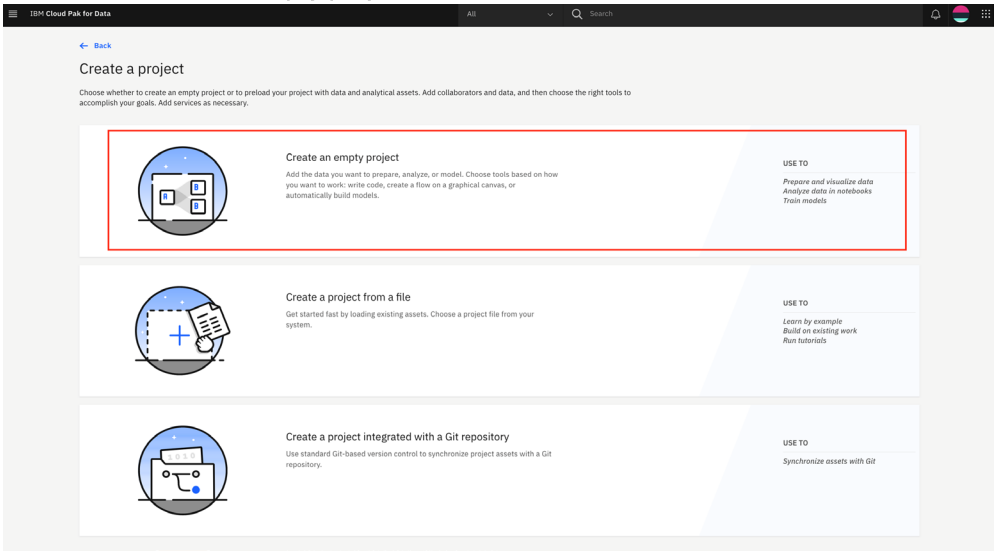
1- If logged out, log back into Cloud Pak for Data as *datascientist* user.
2- Select All projects by clicking on the Navigation menu (top left hamburger icon) and selecting **Projects → All projects** (annotated with red arrow).



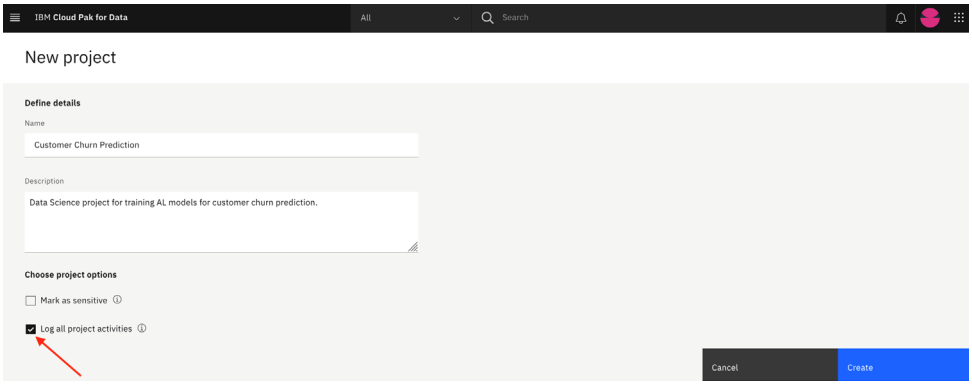

3- Click on **New project** (annotated with red arrow) to create a new project.



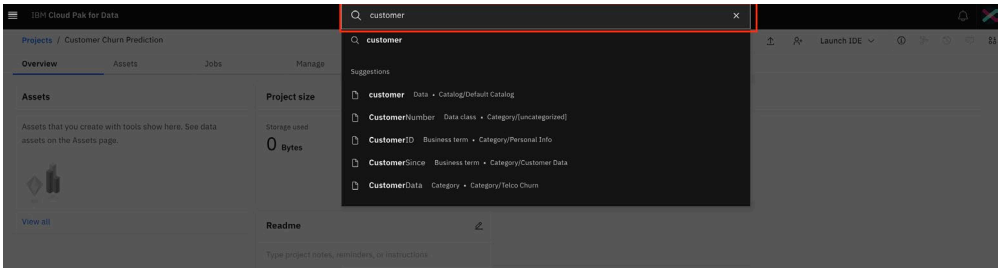
4- Select the **Create an empty project** option (annotated with red rectangle).




5- On the **New project** page, provide a Name and Description (optional) for the project. Select the check box next to Log all project activities (annotated with red arrow; this is optional but helps track project activities). Click **Create**.

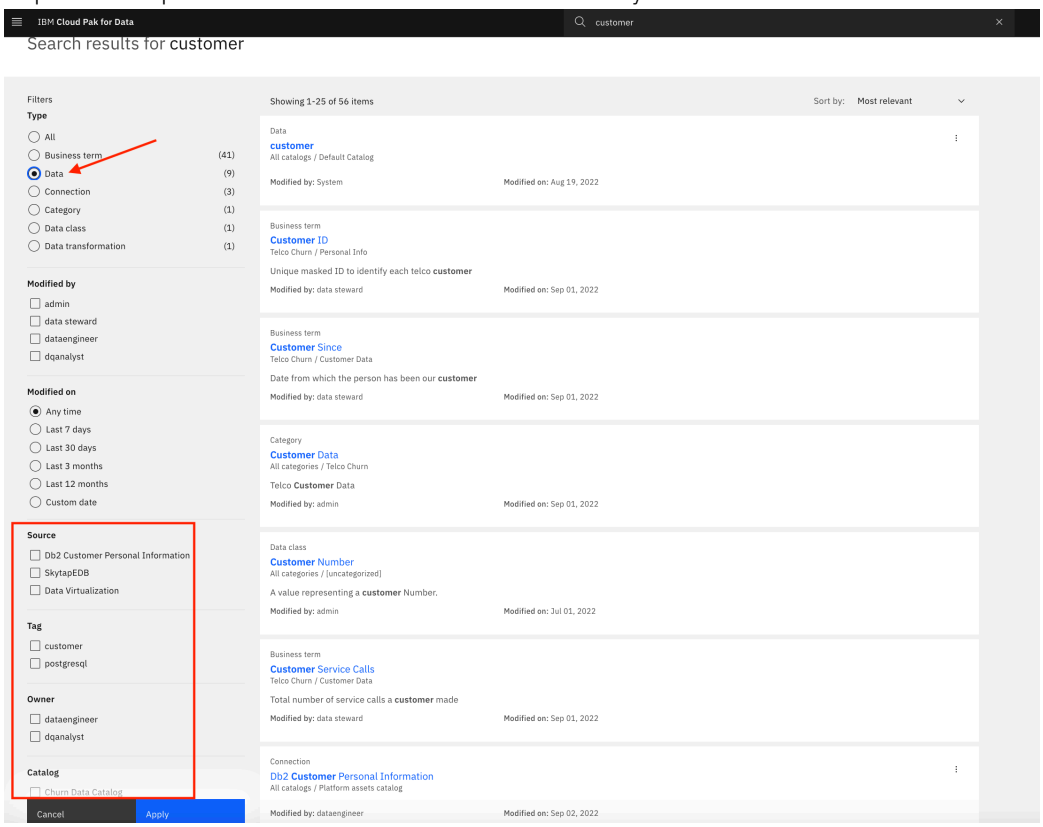



6- Now that the project is created, the datascientist user will shop for relevant data assets that can help them with training this AI model for customer churn prediction. To do so, data scientists can leverage global semantic search capability of Cloud Pak for Data to find data assets related to customers and churn. Type “customer” in the global search bar (annotated with red rectangle) and hit **Enter**.

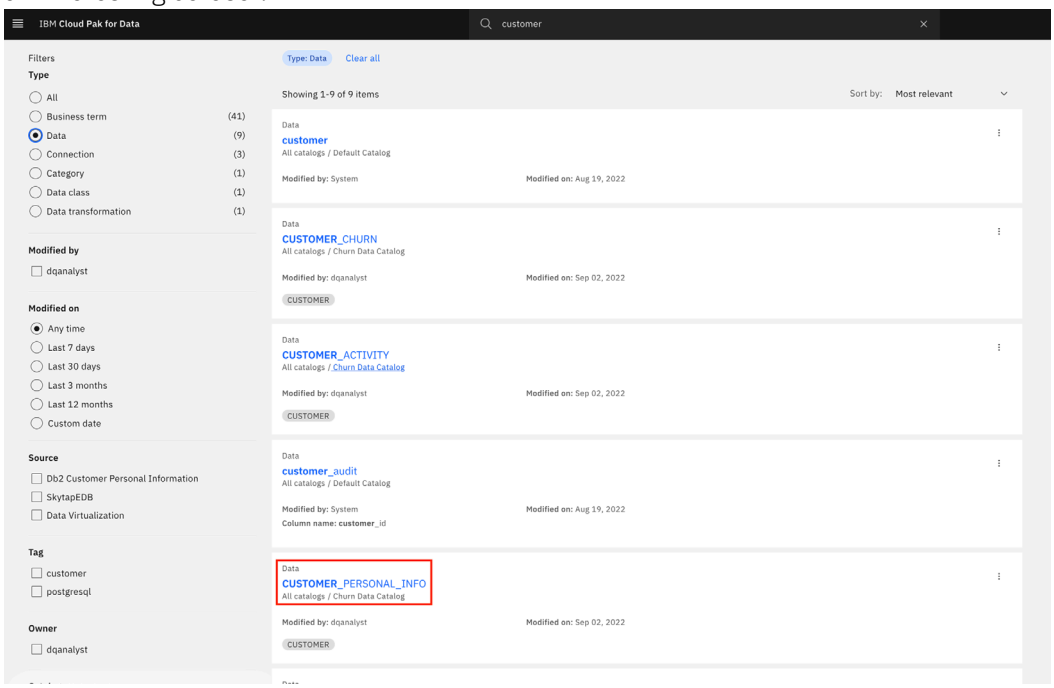



7- Explore the returned search results and note the different types of assets returned including business terms, data assets, data classes, categories, and others. Under the **Type** filter, select **Data** (annotated with red arrow) to filter the results and return only data assets related to the customer term. Click **Apply** button (blue button bottom left view under Filters section).

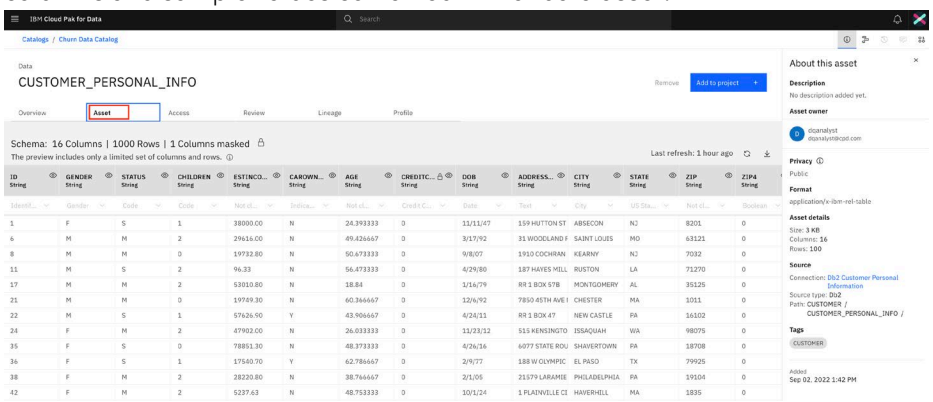
Note that when Data is selected as the Type, additional facets get populated like Source, Owner, Tag, and Catalog (annotated with red rectangle) to allow the user to filter the returned results. For example, consider the scenario where the data scientist is informed that a certain data source system includes relevant assets. Using that information, the data scientist can quickly apply those filters which helps identifying relevant assets more quickly.
*Optional* Explore the various filters and see how they affect the returned results.



8- Click on the CUSTOMER_PERSONAL_INFO data asset (annotated with red rectangle) as it seems like an interesting dataset.

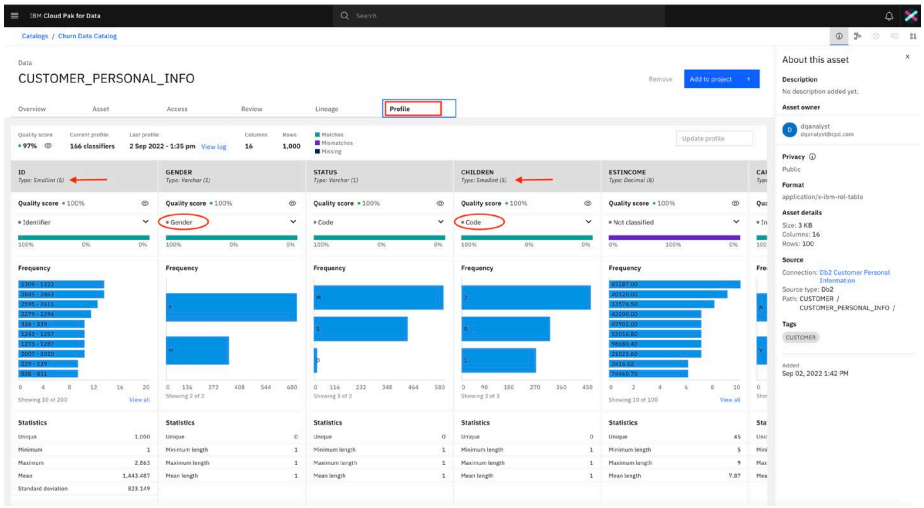



9- On the data set view, click the **Asset** tab (annotated with red rectangle) to take a quick look at the columns and sample values contained in that data asset.

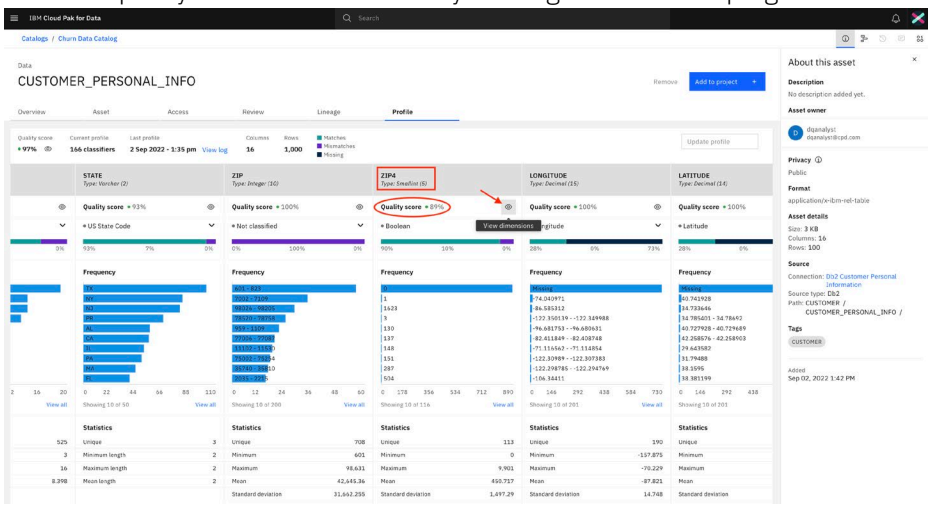



10- Next, click on the **Profile** tab (annotated with red rectangle) to get a profile view of the data which includes data types (annotated with red arrows), assigned data classes (annotated with red ovals),

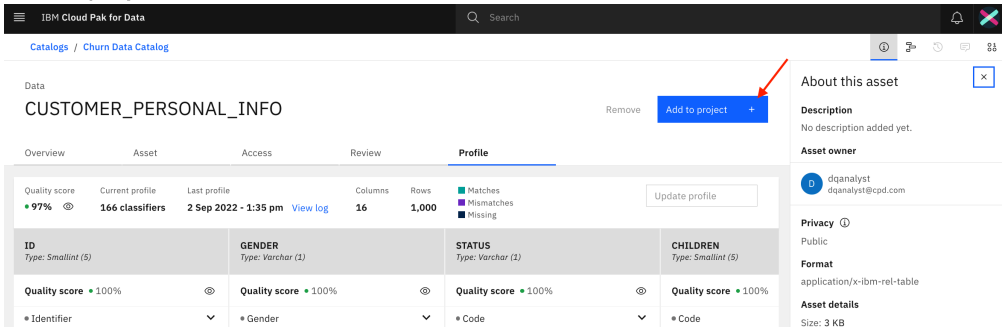
and overall statistics on each data column.




11- Next, review the data quality results at the column level. On the Profile tab, scroll to the right until you find the ZIP4 column (annotated with red rectangle). Note the quality score for ZIP4 column is 89% (annotated with red oval). Click the eye icon (annotated with red arrow) to view the quality dimensions. In this case, quality is low because of data class violations and suspect values. Close out the Data quality dimensions window by clicking the **x** in the top right to return to the asset view.

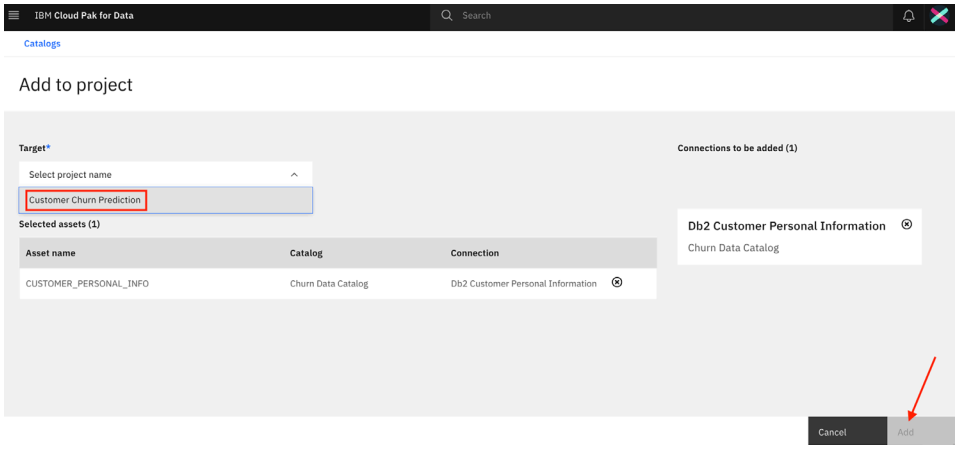



12- Assuming the data scientist thinks the data asset is useful, they would add it to their project. Click on **Add to project** button (annotated with red arrow).

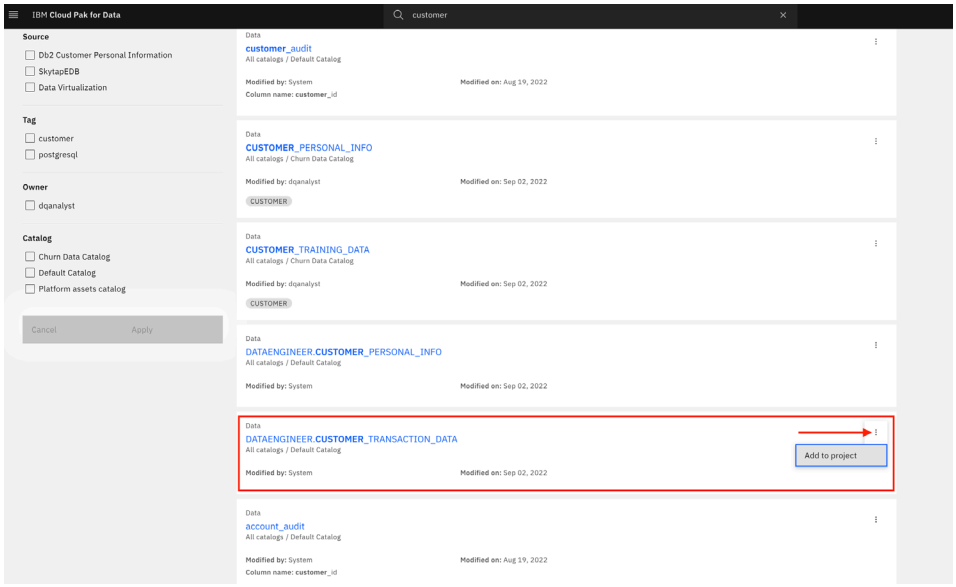



13- On the Add to project page, click the drop down under **Target** field and select the **Customer Churn Prediction** project (annotated with red rectangle) you had created earlier. Click **Add** button (annotated

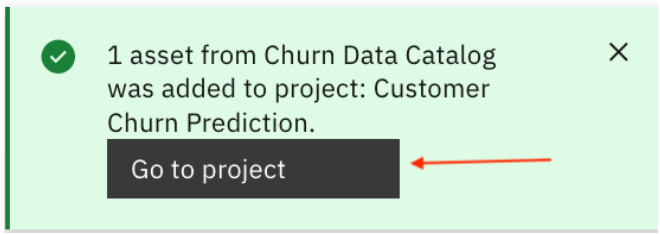
with red arrow) which becomes active once a target project is selected.



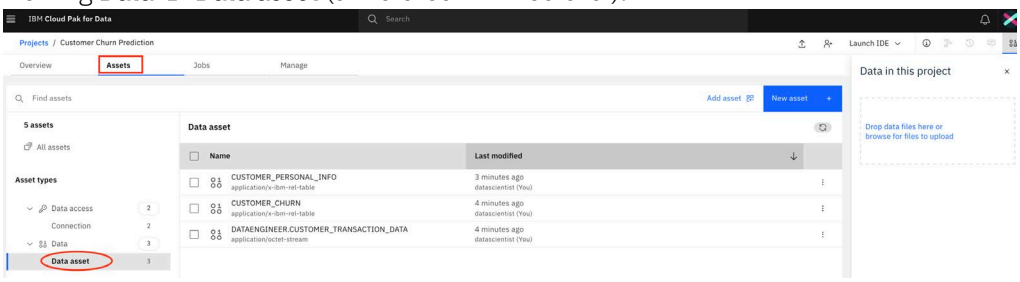

14- Alternatively, you can add data sets to your project directly from the global search results view. Go back and type customer in the global search field and on the search results view, select the **Data** type as filter and click **Apply**. Then scroll down to find the virtualized data asset for customer transaction data, **DATAENGINEER.CUSTOMER_TRANSACTION_DATA** (annotated with red rectangle) and click the menu to the right and select **Add to project** (annotated with red arrow).
On the Add to project page, repeat actions highlighted in step 13.
Afterwards, repeat one more time to add the **CUSTOMER_CHURN** data set which includes churn labels for which customers had previously churned.




15- After adding the 3 data sets, click Go to project on the pop-up notification or navigate back to your Customer Churn Prediction project by going to **Navigation Menu ➔ Projects ➔ All Projects** and selecting your Customer Churn Prediction project.



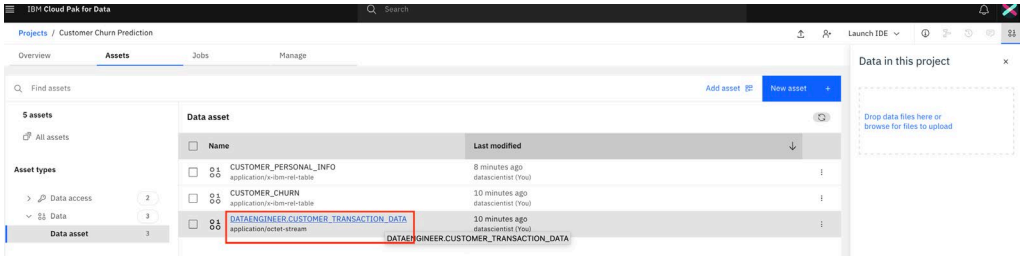
Verify your project has the 3 data sets by clicking the **Assets** tab (annotated with red rectangle) and the viewing **Data ➔ Data asset** (annotated with red oval).



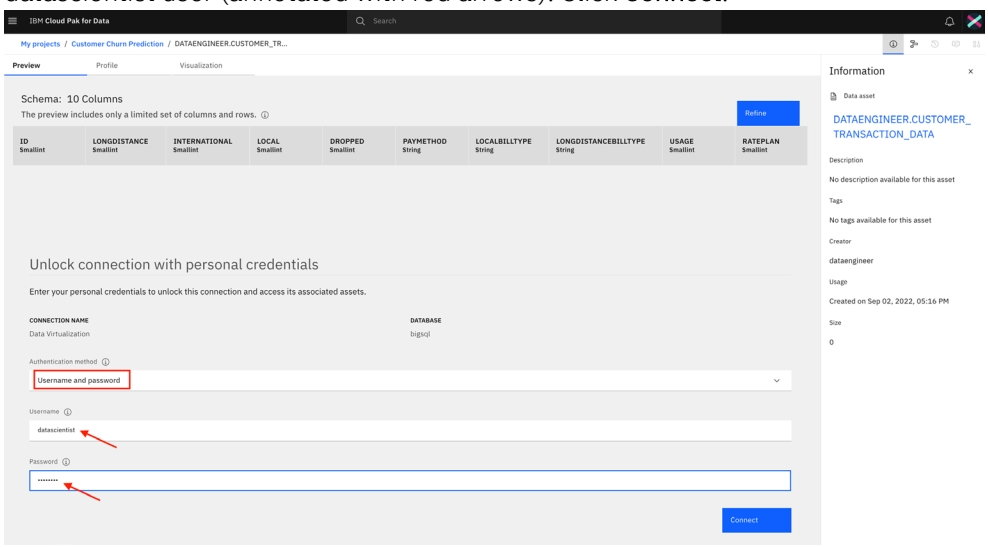
# Data Preparation with Data Refinery in Watson Studio

Now that the relevant data assets are added to your project, the next task you will execute is data preparation using Data Refinery, a visual UI based tool that enables users to interactively discover, cleanse and transform data with over 100 built-in operations.

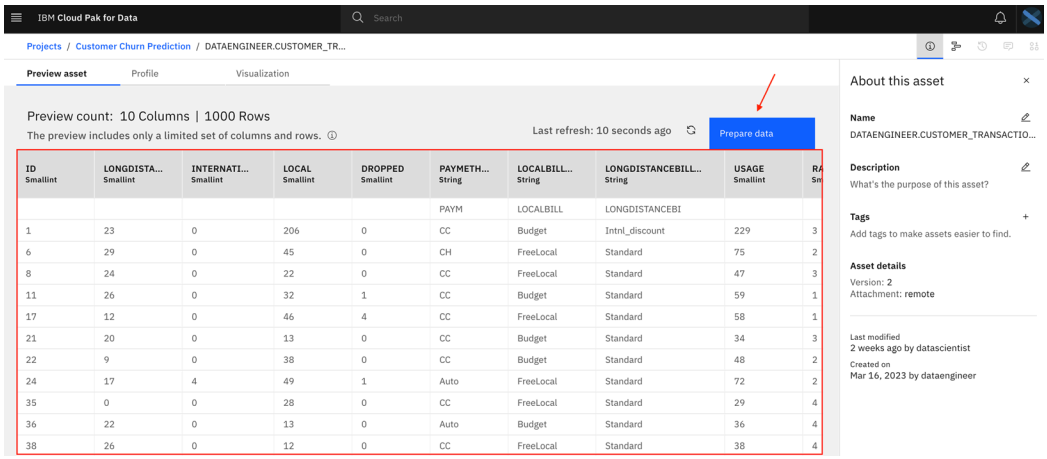
16- Next step is to shape the data to get it ready to be used for training ML models. Cloud Pak for Data supports multiple approaches for data wrangling and transformation. In this lab, you will use Data Refinery to cleanse and shape the data with a graphical flow editor and create a joined data set of the customer personal information, customer transaction data, and the labeled churn data set.
In the project, click the DATAENGINEER.CUSTOMER_TRANSACTION_DATA to view that data set.




17- If you are prompted to Unlock connection with personal credentials, provide the credentials for your datascientist user. These credentials confirm that this user has access to the Data Virtualization connection and can access virtualized data. Select the Authentication method as **Username and password** (annotated with red rectangle) and provide the username and password for your datascientist user (annotated with red arrows). Click **Connect**.



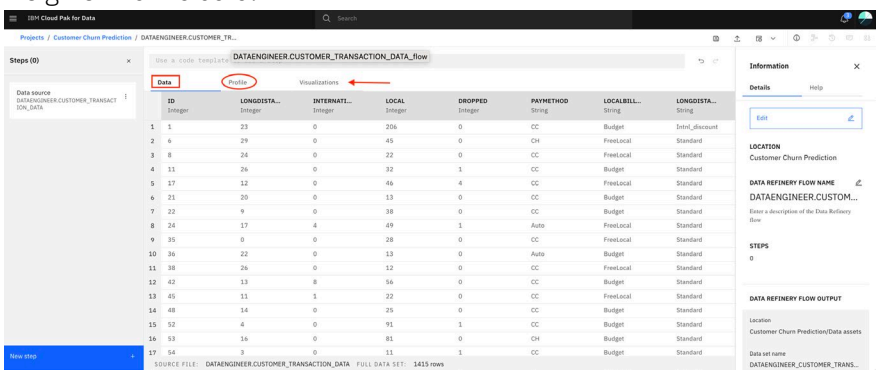

18- After successful connection, the data set sample is viewable. Click **Prepare data** (annotated with red arrow) to start Data Refinery flow.



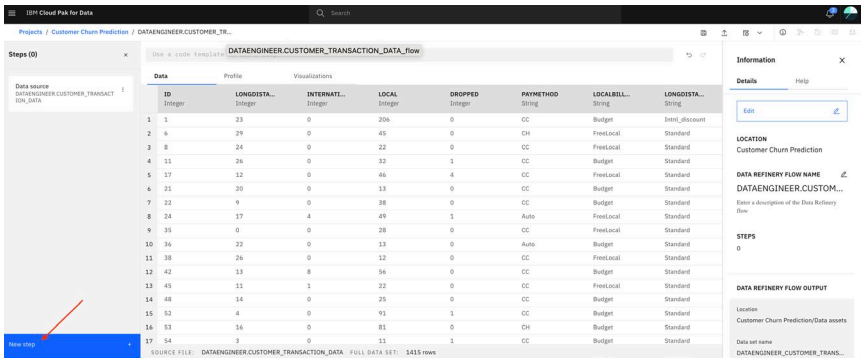

| ID Smallint | LONGDISTA... Smallint | INTERNATI... Smallint | LOCAL Smallint | DROPPED Smallint | PAYMETH... String | LOCALBILL... String | LONGDISTANCEBILL... String | USAGE Smallint | RA Sn |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | PAYM | LOCALBILL | LONGDISTANCEBI | | |
| 1 | 23 | 0 | 206 | 0 | CC | Budget | Intnl_discount | 229 | 3 |
| 6 | 29 | 0 | 45 | 0 | CH | FreeLocal | Standard | 75 | 2 |
| 8 | 24 | 0 | 22 | 0 | CC | FreeLocal | Standard | 47 | 3 |
| 11 | 26 | 0 | 32 | 1 | CC | Budget | Standard | 59 | 1 |
| 17 | 12 | 0 | 46 | 4 | CC | FreeLocal | Standard | 58 | 1 |
| 21 | 20 | 0 | 13 | 0 | CC | Budget | Standard | 34 | 3 |
| 22 | 9 | 0 | 38 | 0 | CC | Budget | Standard | 48 | 2 |
| 24 | 17 | 4 | 49 | 1 | Auto | FreeLocal | Standard | 72 | 2 |
| 35 | 0 | 0 | 28 | 0 | CC | FreeLocal | Standard | 29 | 4 |
| 36 | 22 | 0 | 13 | 0 | Auto | Budget | Standard | 36 | 4 |
| 38 | 26 | 0 | 12 | 0 | CC | FreeLocal | Standard | 38 | 4 |

19- This loads the dataset in Data Refinery. Note the **Data** (annotated with red rectangle), **Profile** (annotated with red oval), and **Visualizations** (annotated with red arrow) tabs.

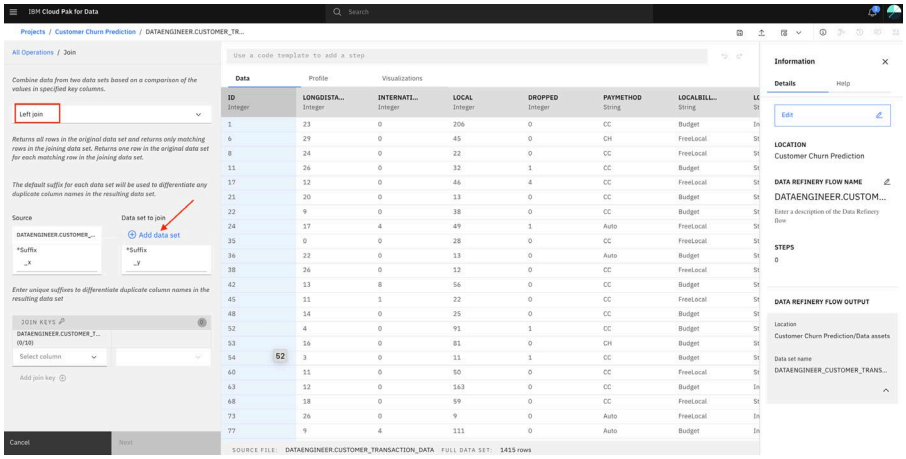
- ⇨ The Data tab displays the data and enables you to apply a number of common operations to cleanse and shape the data in a graphical editor. It also supports deploying R library operations, functions, and logical operators via the command line.
- ⇨ The Profile tab shows useful summary statistics including a histogram of each of the data fields. This is useful to understand the statistical distribution of the data as well as potential skew that may exist.
- ⇨ The Visualizations tab provides over 20 customizable charts to provide perspective and insights into the data.



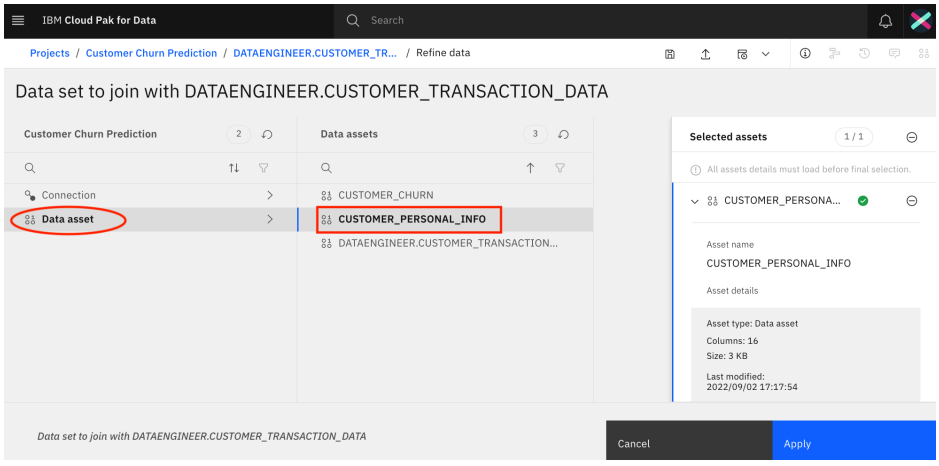

20- Next, you will add a step to join this data set with CUSTOMER_PERSONAL_INFO data set to capture additional features that may impact the likelihood of a customer to churn.
Click the **New Step** button (annotated with red arrow) which will open the operations column, scroll down to find the Join operation and click **Join**.
You can also type Join in the Search operations field and it will filter the list of operations to find Join.



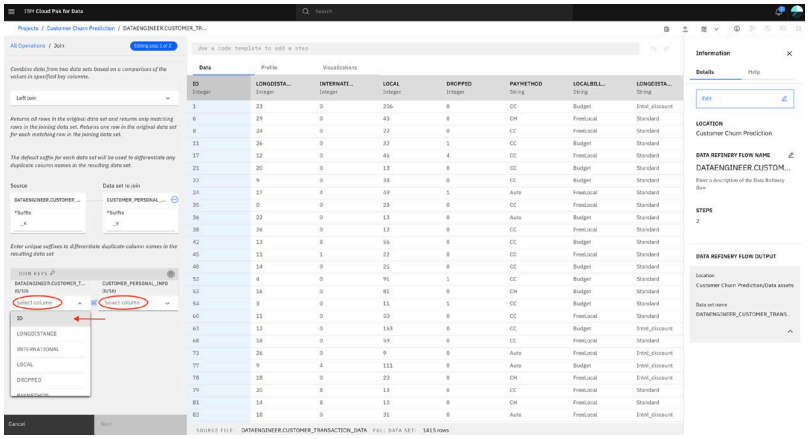
21- On the Join operations window, keep the type of join as “Left join” (annotated with red rectangle) and then click **Add data set** (annotated with red arrow).



22- On the Data set page, click **Data asset** (annotated with red oval), select the **CUSTOMER_PERSONAL_INFO** data set and click **Apply**. In case you’re not familiar with the Left Join operation, Data Refinery provides an explanation of what that operation does; specifically, a Left Join returns all rows in the original data set and returns only the matching rows in the joining data set.

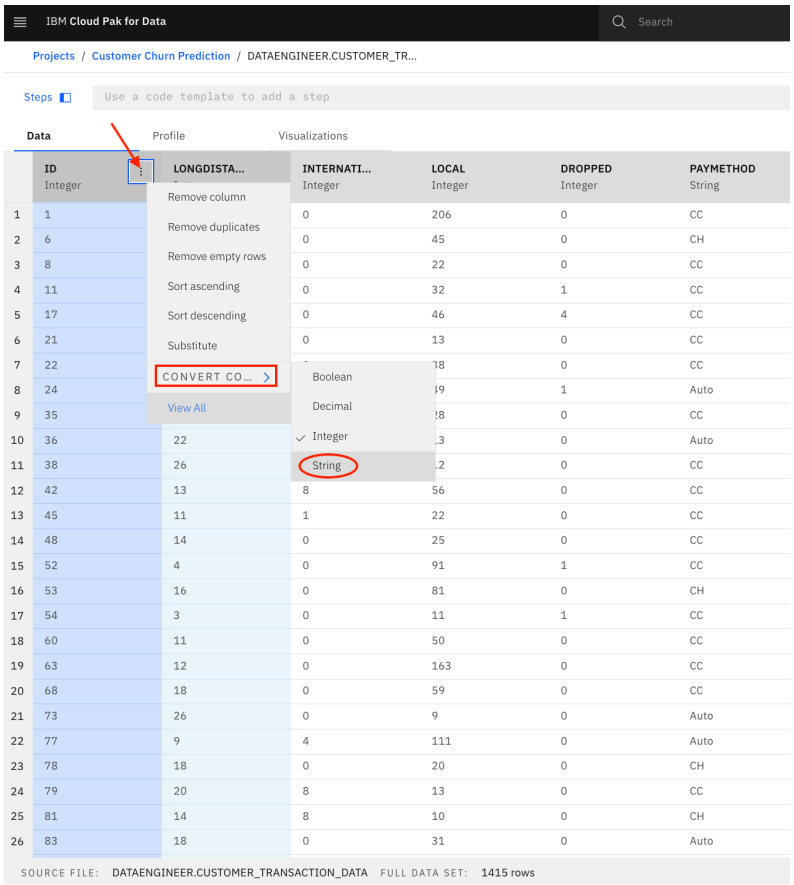



Back on the Join operation window, click **Select column** (annotated with red oval) to specify **ID** (annotated with red arrow) as the field to use for joining the two data sets. Then click **Next**.



*NOTE*: After selecting the ID field under **DATAENGINEER.CUSTOMER_TRANSACTION_DATA** dataset, if you don't see the ID field under the **CUSTOMER_PERSONAL_INFO** dataset, this is caused by the fact that the data types for ID field do not match between the two datasets.

If you run into the issue, please change the data type for the ID field in the loaded dataset by clicking the actions menu (annotated with red arrow) next to the ID field, selecting the **CONVERT COLUMN TYPE** option (annotated with red rectangle) and changing the data type. If the data type is already Integer, change it to String and if it is String, change it to Integer.

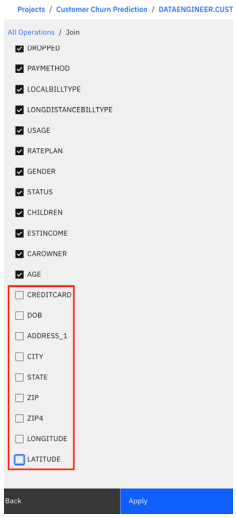



| | ID Integer | LONGDISTA... Integer | INTERNATI... Integer | LOCAL Integer | DROPPED Integer | PAYMETHOD String |
|---|---|---|---|---|---|---|
| 1 | 1 | | 0 | 206 | 0 | CC |
| 2 | 6 | | 0 | 45 | 0 | CH |
| 3 | 8 | | 0 | 22 | 0 | CC |
| 4 | 11 | | 0 | 32 | 1 | CC |
| 5 | 17 | | 0 | 46 | 4 | CC |
| 6 | 21 | | 0 | 13 | 0 | CC |
| 7 | 22 | | | 18 | 0 | CC |
| 8 | 24 | | | 19 | 1 | Auto |
| 9 | 35 | | | 28 | 0 | CC |
| 10 | 36 | 22 | | 3 | 0 | Auto |
| 11 | 38 | 26 | | 2 | 0 | CC |
| 12 | 42 | 13 | 8 | 56 | 0 | CC |
| 13 | 45 | 11 | 1 | 22 | 0 | CC |
| 14 | 48 | 14 | 0 | 25 | 0 | CC |
| 15 | 52 | 4 | 0 | 91 | 1 | CC |
| 16 | 53 | 16 | 0 | 81 | 0 | CH |
| 17 | 54 | 3 | 0 | 11 | 1 | CC |
| 18 | 60 | 11 | 0 | 50 | 0 | CC |
| 19 | 63 | 12 | 0 | 163 | 0 | CC |
| 20 | 68 | 18 | 0 | 59 | 0 | CC |
| 21 | 73 | 26 | 0 | 9 | 0 | Auto |
| 22 | 77 | 9 | 4 | 111 | 0 | Auto |
| 23 | 78 | 18 | 0 | 20 | 0 | CH |
| 24 | 79 | 20 | 8 | 13 | 0 | CC |
| 25 | 81 | 14 | 8 | 10 | 0 | CH |
| 26 | 83 | 18 | 0 | 31 | 0 | Auto |

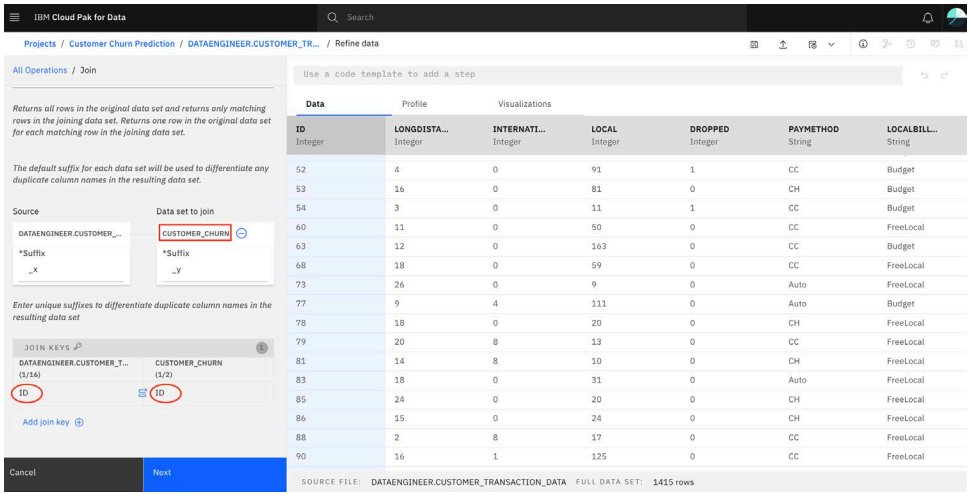
SOURCE FILE: DATAENGINEER.CUSTOMER_TRANSACTION_DATA FULL DATA SET: 1415 rows

23- On the next window, it shows all the fields that will result from the join operation. At this point, you can remove fields you do not wish to include in the final data set. Remove the bottom nine fields as shown below (annotated with red rectangle) and click **Apply.** These fields are removed as they don't contribute any meaningful information for predicting the likelihood of a customer to churn. The fields to remove are:

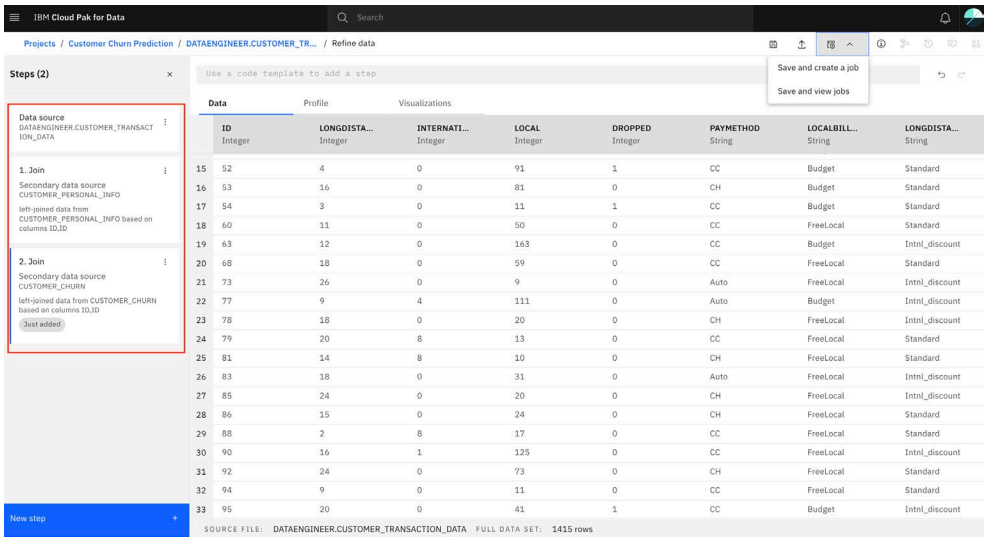
***CREDITCARD, DOB, ADDRESS_1, CITY, STATE, ZIP, ZIP4, LONGITDUE, LATITUDE***




24- Repeat the process (steps 22-24) to apply a join operation on the resulting data set and the **CUSTOMER_CHURN** dataset. The join field is **ID** (annotated with red oval) and the data set to join is **CUSTOMER_CHURN** (annotated with red rectangle).




25- Note that the Data Refinery flow (annotated with red rectangle) has been augmented with all the executed operations. As you perform more operations to shape the data, they get added to the Data Refinery flow. For this lab, we will just perform the Join operations but typically, you'd perform several other operations to transform the data and make it ready for analytics insights and training machine learning models.



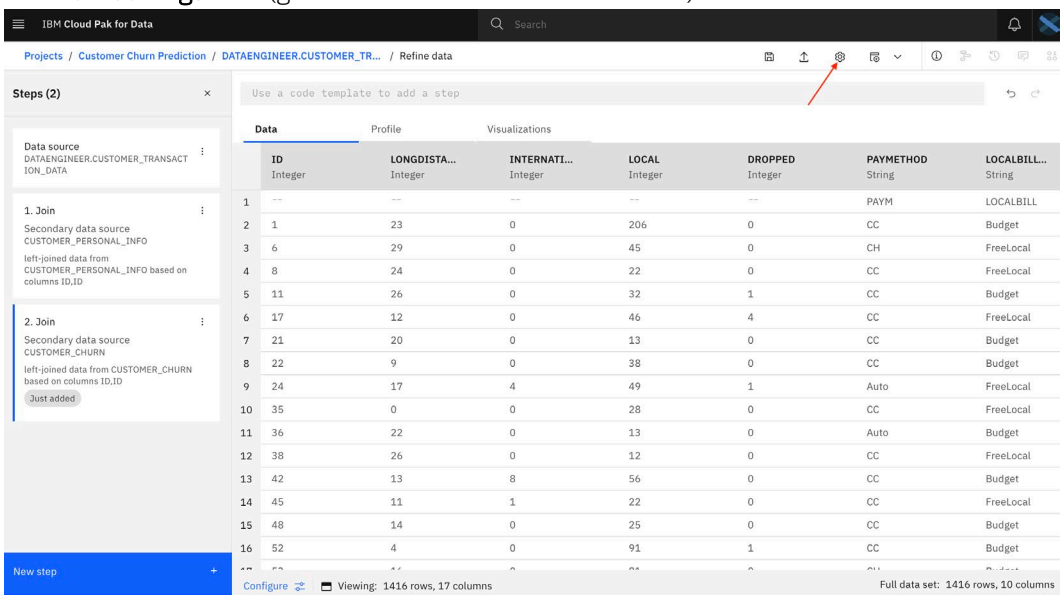

| | ID Integer | LONGDISTA... Integer | INTERNATI... Integer | LOCAL Integer | DROPPED Integer | PAYMETHOD String | LOCALBILL... String | LONGDISTA... String |
|---|---|---|---|---|---|---|---|---|
| 15 | 52 | 4 | 0 | 91 | 1 | CC | Budget | Standard |
| 16 | 53 | 16 | 0 | 81 | 0 | CH | Budget | Standard |
| 17 | 54 | 3 | 0 | 11 | 1 | CC | Budget | Standard |
| 18 | 60 | 11 | 0 | 50 | 0 | CC | FreeLocal | Standard |
| 19 | 63 | 12 | 0 | 163 | 0 | CC | Budget | Intnl_discount |
| 20 | 68 | 18 | 0 | 59 | 0 | CC | FreeLocal | Standard |
| 21 | 73 | 26 | 0 | 9 | 0 | Auto | FreeLocal | Intnl_discount |
| 22 | 77 | 9 | 4 | 111 | 0 | Auto | Budget | Intnl_discount |
| 23 | 78 | 18 | 0 | 20 | 0 | CH | FreeLocal | Intnl_discount |
| 24 | 79 | 20 | 8 | 13 | 0 | CC | FreeLocal | Standard |
| 25 | 81 | 14 | 8 | 10 | 0 | CH | FreeLocal | Standard |
| 26 | 83 | 18 | 0 | 31 | 0 | Auto | FreeLocal | Intnl_discount |
| 27 | 85 | 24 | 0 | 20 | 0 | CH | FreeLocal | Intnl_discount |
| 28 | 86 | 15 | 0 | 24 | 0 | CH | FreeLocal | Standard |
| 29 | 88 | 2 | 8 | 17 | 0 | CC | FreeLocal | Standard |
| 30 | 90 | 16 | 1 | 125 | 0 | CC | FreeLocal | Intnl_discount |
| 31 | 92 | 24 | 0 | 73 | 0 | CH | FreeLocal | Standard |
| 32 | 94 | 9 | 0 | 11 | 0 | CC | FreeLocal | Standard |
| 33 | 95 | 20 | 0 | 41 | 1 | CC | Budget | Intnl_discount |

In practice, the data typically requires several more operations to cleanse by removing nulls, filtering rows with missing data, aggregating data across fields, and/or applying a number of different operations. In this lab, the dataset we're using is already in good shape and the only operations you will apply is to join the customer data (which was already a join of customer personal information and transaction data) and labeled churn data set.

Take a minute to browse the set of supported operations.

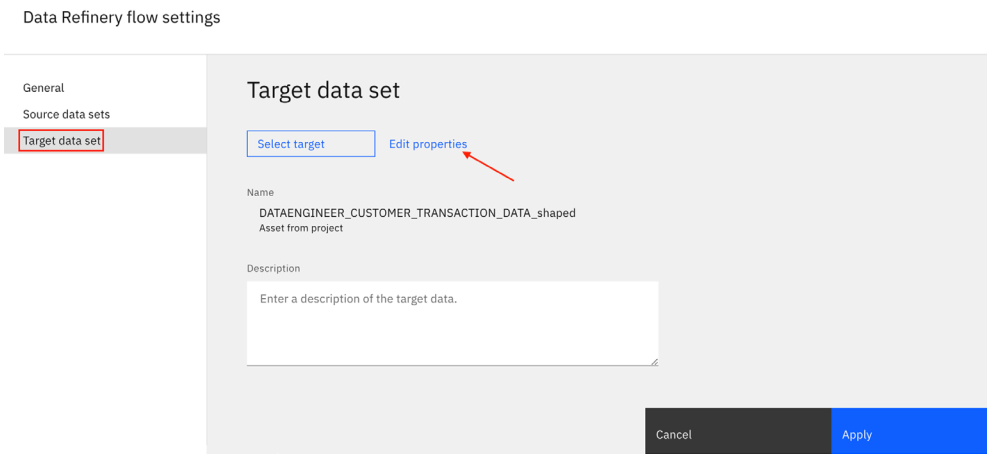
This video provides more examples and details of operations for data transformation using Data Refinery.

26- Once you've applied all the operations to transform the data, next step is to edit the properties of the flow to change the default name of the output data set produced by the refinery flow. To do so, click the **Flow settings** icon (gear icon annotated with red arrow).

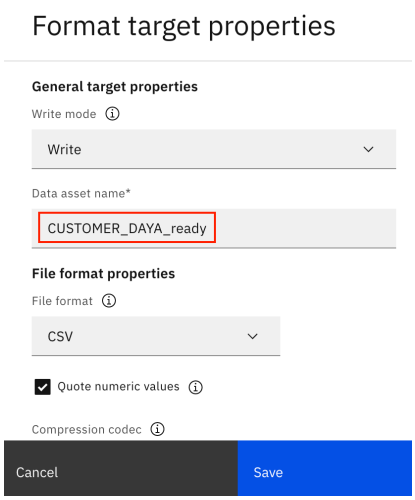



| | ID Integer | LONGDISTA... Integer | INTERNATI... Integer | LOCAL Integer | DROPPED Integer | PAYMETHOD String | LOCALBILL... String |
|---|---|---|---|---|---|---|---|
| 1 | -- | -- | -- | -- | -- | PAYM | LOCALBILL |
| 2 | 1 | 23 | 0 | 206 | 0 | CC | Budget |
| 3 | 6 | 29 | 0 | 45 | 0 | CH | FreeLocal |
| 4 | 8 | 24 | 0 | 22 | 0 | CC | FreeLocal |
| 5 | 11 | 26 | 0 | 32 | 1 | CC | Budget |
| 6 | 17 | 12 | 0 | 46 | 4 | CC | FreeLocal |
| 7 | 21 | 20 | 0 | 13 | 0 | CC | Budget |
| 8 | 22 | 9 | 0 | 38 | 0 | CC | Budget |
| 9 | 24 | 17 | 4 | 49 | 1 | Auto | FreeLocal |
| 10 | 35 | 0 | 0 | 28 | 0 | CC | FreeLocal |
| 11 | 36 | 22 | 0 | 13 | 0 | Auto | Budget |
| 12 | 38 | 26 | 0 | 12 | 0 | CC | FreeLocal |
| 13 | 42 | 13 | 8 | 56 | 0 | CC | Budget |
| 14 | 45 | 11 | 1 | 22 | 0 | CC | FreeLocal |
| 15 | 48 | 14 | 0 | 25 | 0 | CC | Budget |
| 16 | 52 | 4 | 0 | 91 | 1 | CC | Budget |

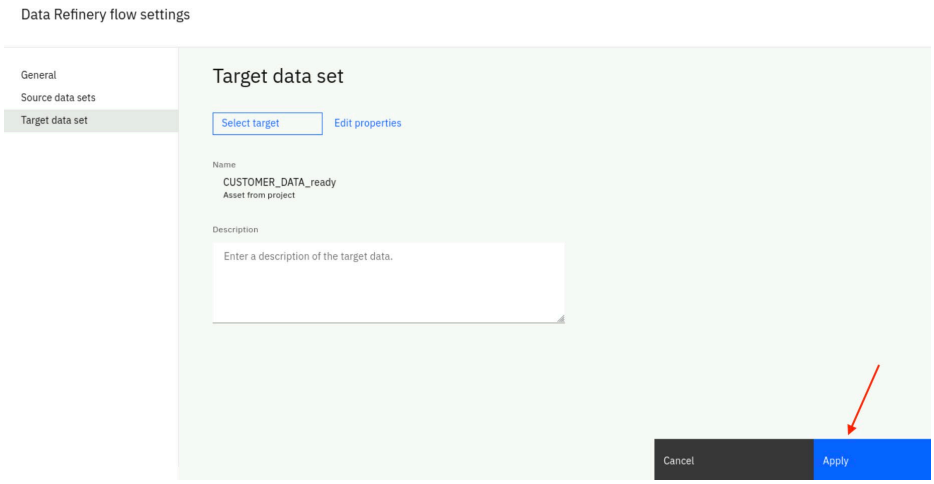
27- Select **Target data set** (annotated with red rectangle) and click **Edit properties** (annotated with red arrow).



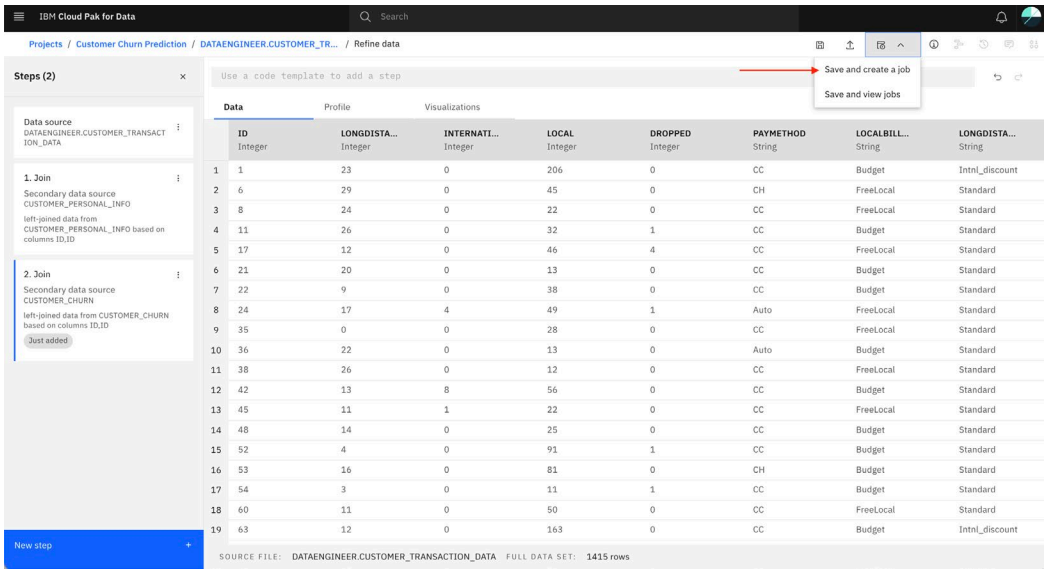

28- Change Data asset name to be **CUSTOMER_DATA_ready** (annotated with red rectangle). You can edit the name to whatever you like. However, the rest of the lab, assumes you called it CUSTOMER_DATA_ready. Click **Save**.



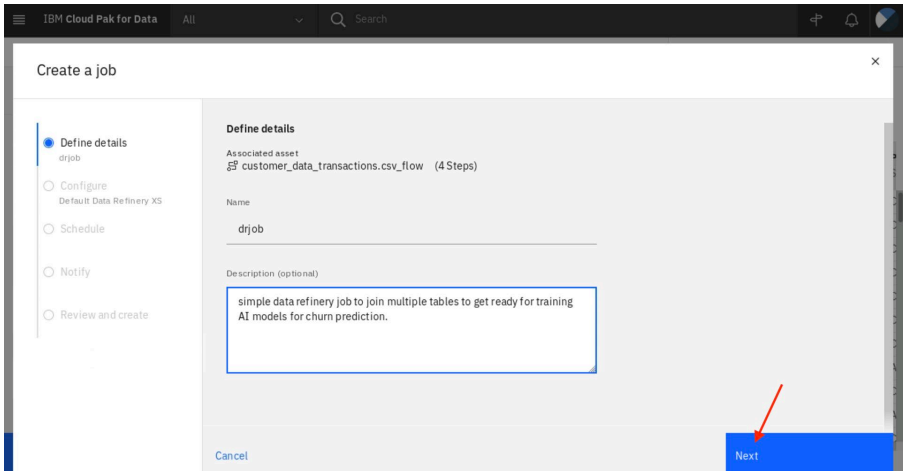

29- Click **Apply** (annotated with red arrow) to save out the settings changes.



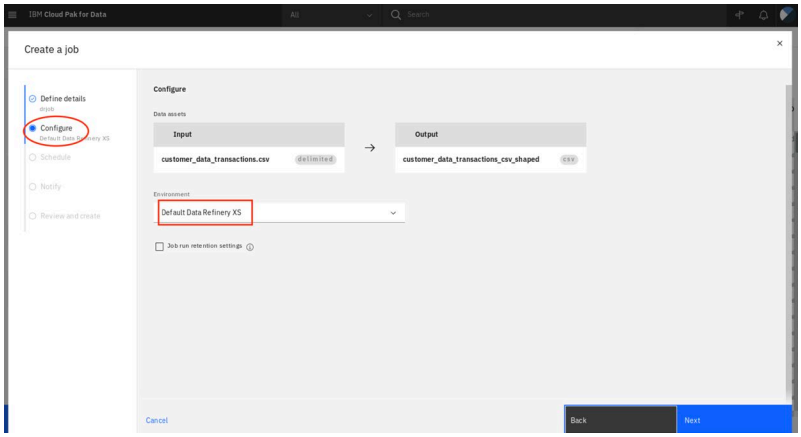

30- Next step is to save the flow and create a job to apply this data refinery flow against the complete data set. To do so, click on **Save and create a job** (annotated with red arrow).



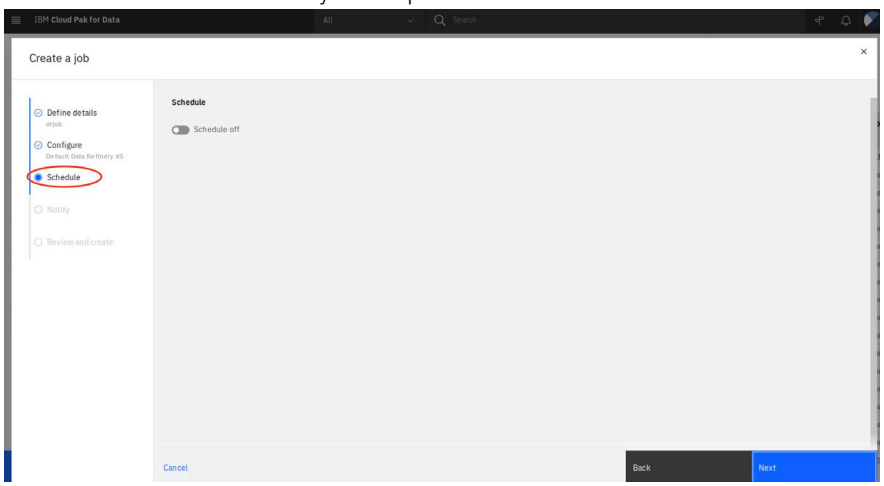

31- Provide a Name for the job <drjob>, add a Description (optional), and click Next (annotated with red arrow).



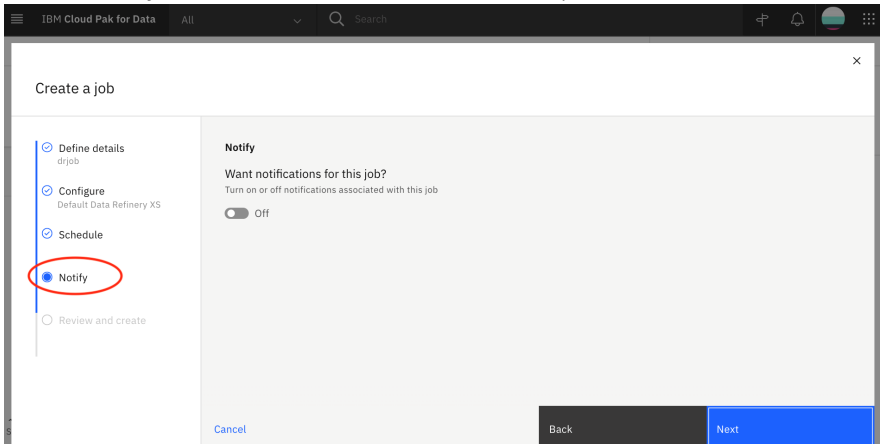

On the **Configure** tab (annotated with red oval), review the **Environment,** and keep the default selection (annotated with red rectangle), and click **Next**. For jobs that require more resources, you can select a larger Environment to run the job.



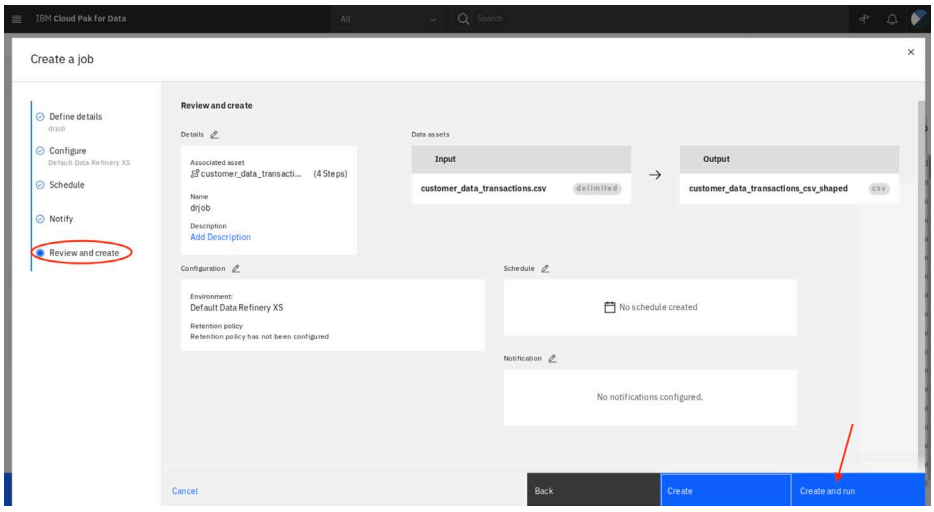
On the **Schedule** tab (annotated with red oval), keep the Schedule slider set to off (default), and click **Next**. In this lab, we don't need to run the data refinery job at a given schedule but we'll manually run it as needed and that is why we kept the default selection as off.



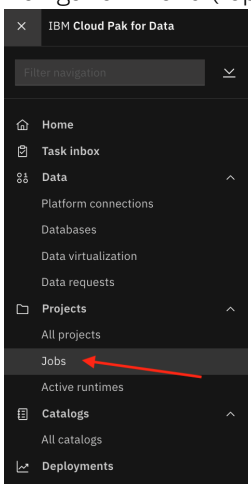

On the Notify tab (annotated with red oval), keep the Notification off as default. Click **Next**.



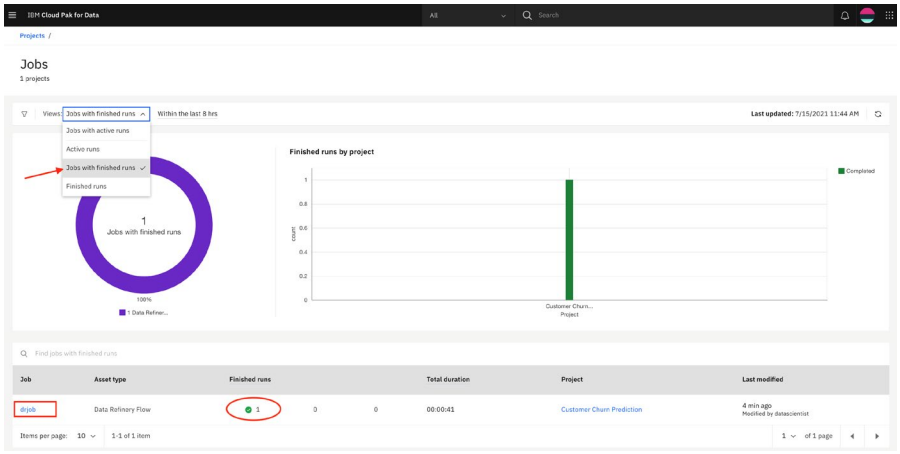

On the **Review and create** tab (annotated with red oval), review the job details and click **Create and run** (annotated with red arrow).



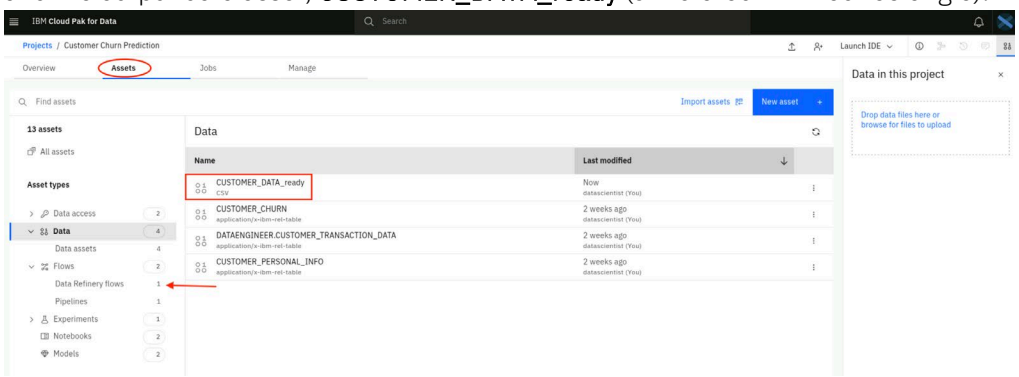
32- After you click Create and Run, navigate to the jobs view to monitor progress by clicking on the Navigation menu (top left hamburger icon) and selecting **Jobs** (annotated with red arrow).




33- On the Jobs page, you can filter the view by selecting whether you want to look at **Active runs**, **Jobs with active runs**, **Jobs with finished runs**, or **Finished runs**. Feel free to filter the different views to see the results. Initially, the job will appear in the view **Jobs with active runs** and when it completes, the job will appear in the view **Jobs with finished runs**. Feel free to click on the job name (annotated with red rectangle) and review the details and status of the run(s).




34- Navigate back to your **Customer Churn Prediction** project and click the **Assets** tab (annotated with red oval). Note the new assets added to your project, one Data Refinery flow (annotated with red arrow) and the output data asset, **CUSTOMER_DATA_ready** (annotated with red rectangle).



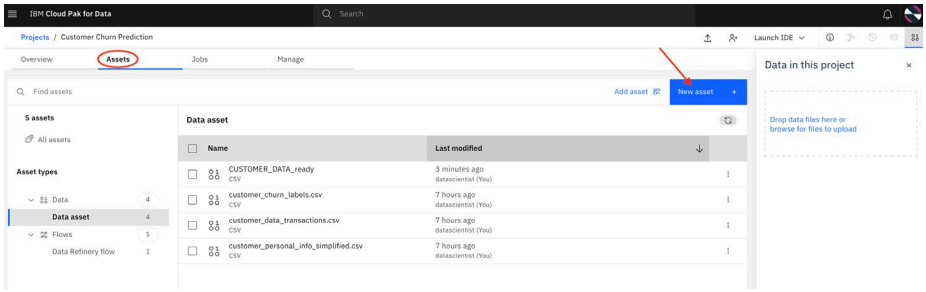
At this point, you have collected data from various sources and leveraged Data Refinery to shape the data using a graphical editor. Now, the data is ready to be used for training a machine learning model for predicting the likelihood of a customer to churn based on his/her demographic and transaction data.

# Option 1: Train AutoAI Model for Churn Prediction

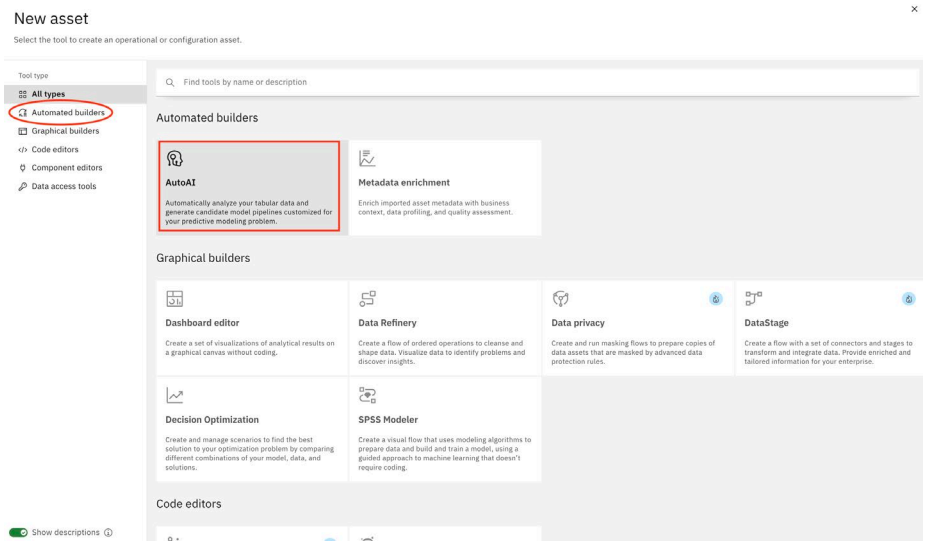
In this section, we illustrate how to leverage AutoAI to quickly train multiple AI models for churn prediction and select the pipeline that delivers best performance.

### Create and run an AutoAI experiment

1- Navigate back to your project, click **Assets** tab (annotated with red oval), and click **New asset** + (annotated with red arrow).

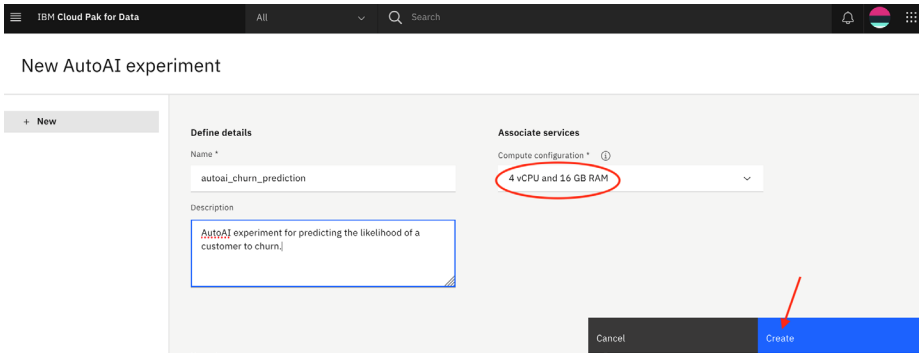



2- Scroll down and click **AutoAI** tile (annotated with red rectangle). Note you can also filter the Tool type to **Automated builders** (annotated with red oval) for quicker access to such tools.

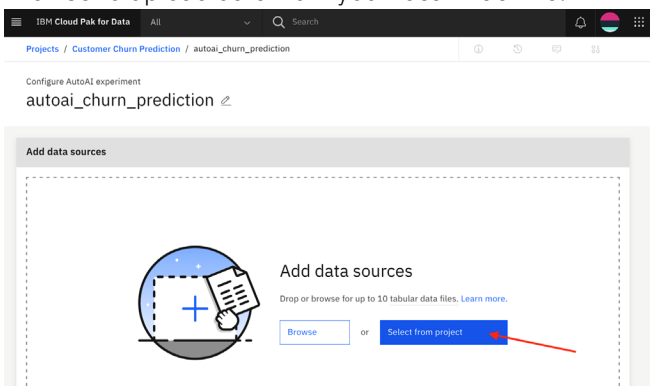



3- Provide a Name and a Description (optional) for your AutoAI experiment, keep the default Compute configuration (annotated with red oval) and click **Create** (annotated with red arrow). You could select

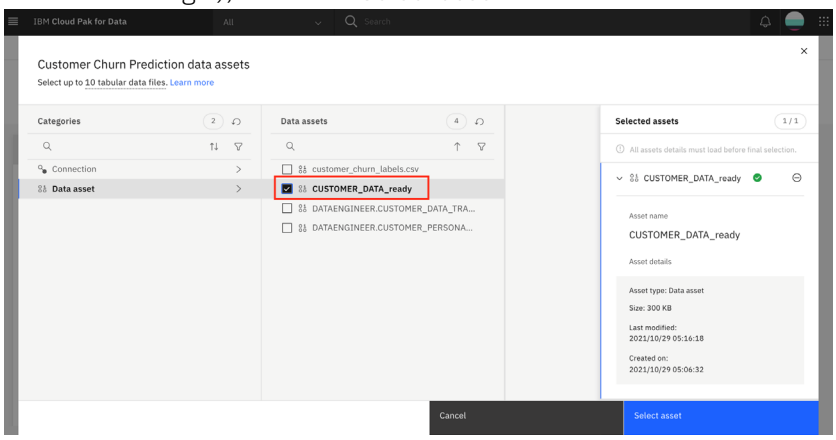
a different configuration if you needed to assign more resources for your AutoAI experiment.



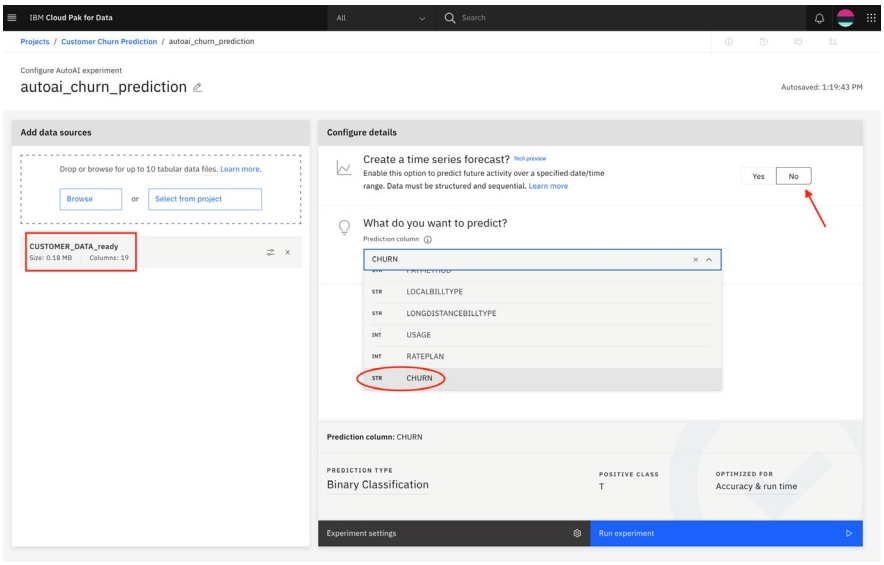

4- On the AutoAI add data sources page, click the **Select from project** button (annotated with red arrow) since you will be using the dataset you had created earlier with Data Refinery. You could also click Browse to upload data from your local machine.




5- Click Data asset and select the checkbox to select the **CUSTOMER_DATA_ready** dataset (annotated with red rectangle); then click **Select asset**.

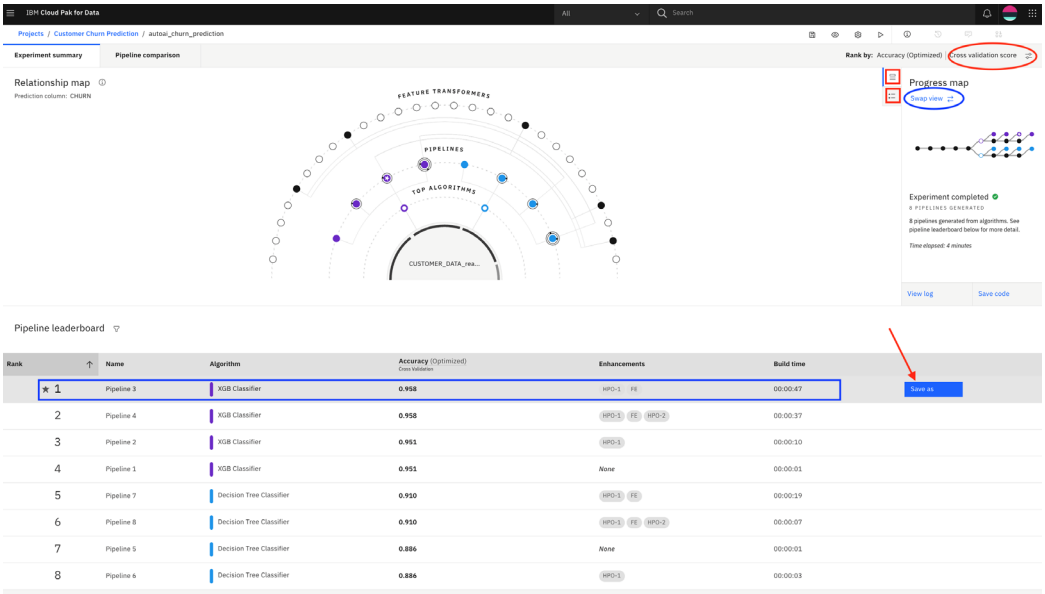



6- On the next page, you will see the selected dataset (annotated with red rectangle) and you will be prompted to select whether you want to Create a time series forecast which is supported by AutoAI. Click **No** (annotated with red arrow) since customer churn prediction is a classification use case and not a time series forecasting use case. Once you click No, you will get the option to select which column to predict. Scroll down the list to select the **CHURN** column (annotated with red oval). At this point, we have provided the data set, indicated it is a classification use case and selected the prediction column.




7- Click **Run experiment** (blue button) to kick off the AutoAI run. Note that you can click the Experiment settings to review the default settings and change some of the configurations if you wish. Please review those settings as they're very informative.

8- AutoAI runs for a few minutes on this dataset and produces a number of pipelines as shown in figure below including training/test data split, data preprocessing, feature engineering, model selection, and hyperparameter optimization. You can dig deeper into any of the pipelines to better understand feature importance, the resulting metrics, the selected model, and any applied feature transformation.
While waiting for AutoAI's run to complete, please review the AutoAI Documentation.

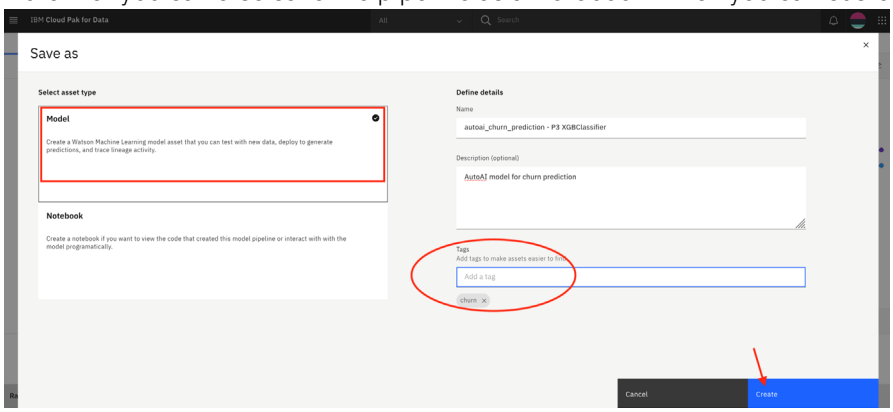



Specifically, review the AutoAI Implementation Details to understand what algorithms are supported, what data transformations are applied and what metrics can be optimized.

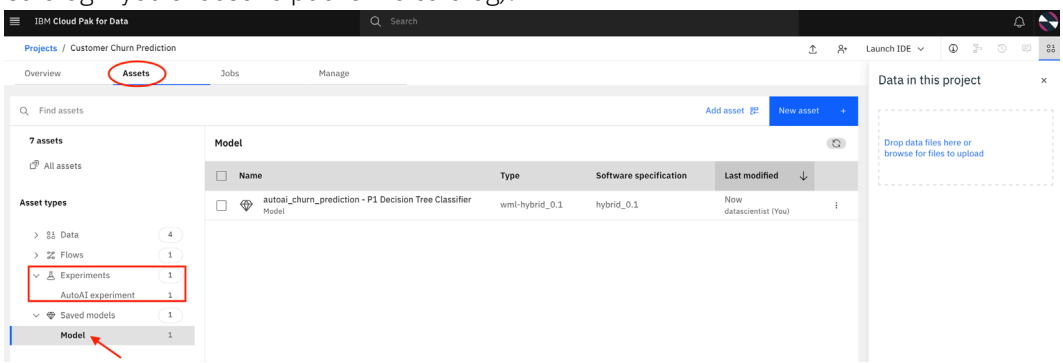
The AutoAI run will take 2-3 minutes to complete. Once complete, please spend a few minutes exploring the dashboard.

- ⇨ Switch between the ***Experiment details*** and the ***Legend*** information (annotated with red rectangles in previous figure) to better understand the generated Relationship map.
- ⇨ Switch between ***Cross Validation*** and ***Holdout*** results by clicking the icon next to Cross Validation (annotated with red oval in previous figure) to see how the pipeline ranking changed depending on which data is being evaluated.
- ⇨ Swap the view between the ***Relationship map*** and the ***Progress map*** (annotated with blue oval in previous figure) to see the different views of the AutoAI pipeline creation process.
- ⇨ Click on the top pipeline (annotated with blue rectangle in previous figure) to review the details for that pipeline. AutoAI reports several valuable evaluation criteria like several performance metrics (Accuracy, Area under ROC, Precision, Recall, F1) as well as the confusion matrix, Precision Recall Curve, and feature importance. If the pipeline also included feature engineering (or feature transformation), the pipeline details will explain what transformations were applied.
- ⇨ Close the pipeline details window by clicking **x** top right of window. After reviewing the trained pipelines, you can decide which one you'd like to save as a model to deploy. Assuming you select the first pipeline, mouse over the first pipeline and click **Save as** (annotated with red arrow in figure above).

9- On the Save As page, select **Model** (annotated with red rectangle), update the default Name if you wish, add a Description (optional) and Tags (optional) and click **Create** (annotated with red arrow). Note that you can also save the pipeline as a Notebook which you can customize further.

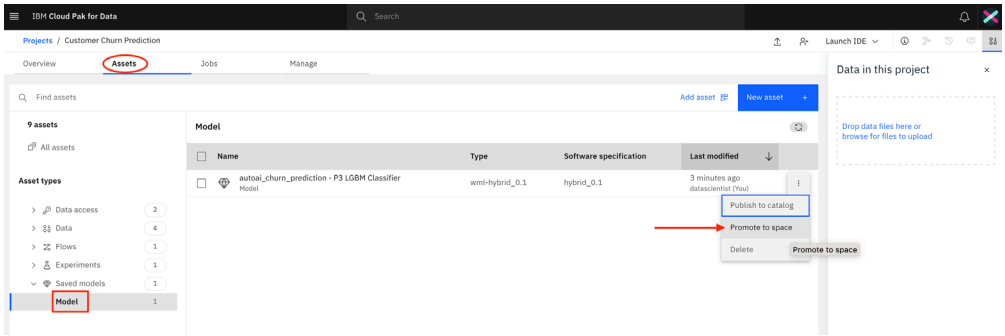



10- Navigate back to the project assets by clicking your project and then clicking the **Assets** tab (annotated with red oval). Notice the new AutoAI experiment (annotated with red rectangle) and the new model you created under the Saved models section (annotated with red arrow). *Optional* It is a good practice to publish the model to the catalog and you can do so by clicking on the actions menu (3 vertical dots) next to the model and selecting **Publish to Catalog** (you can publish to Churn Data Catalog if you choose to publish to catalog).

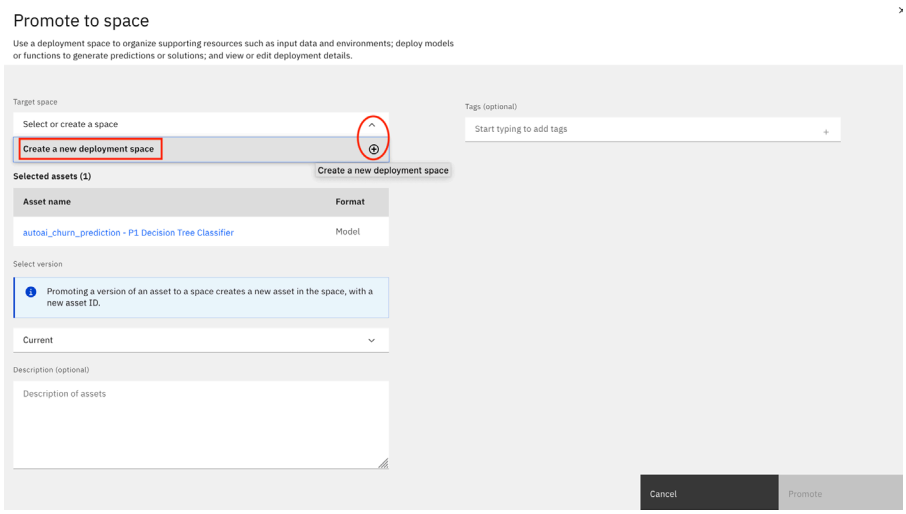


## Deploy your AutoAI model

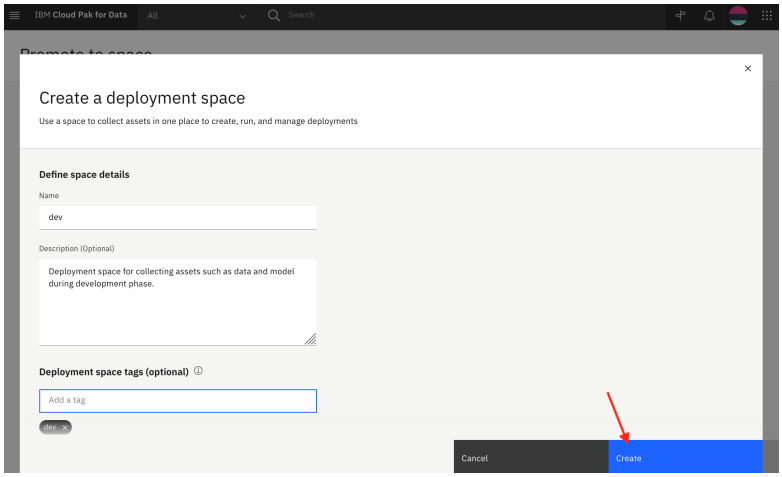
11- On the project **Assets** page (annotated with red oval), select **Model** under Saved models (annotated with red rectangle) and then select **Promote to Space** (annotated with red arrow) from the the actions menu (3 vertical dots) next to the model name.



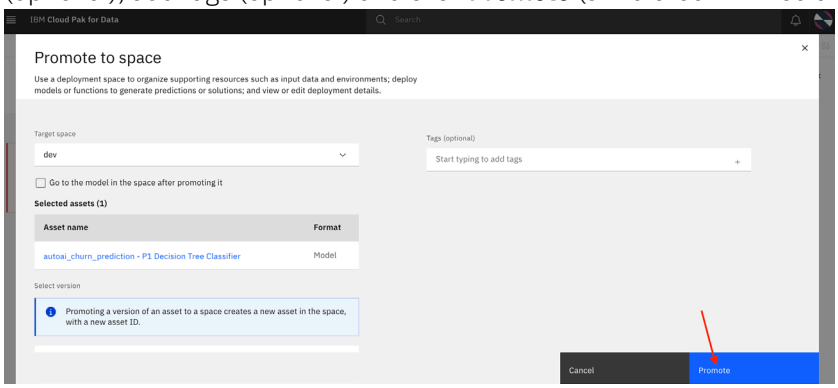

12- On the Promote to space page, select a target deployment space from the drop-down (annotated with red oval) if you had already created spaces before. If not, click on **Create a new deployment space** (annotated with red rectangle) to create a new space.



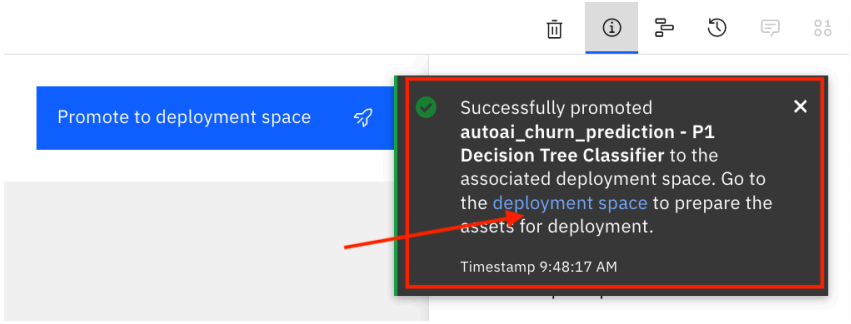

13- Provide a Name and a Description (optional) for the deployment space and click **Create** (annotated with red arrow). You can also add tags to the space.



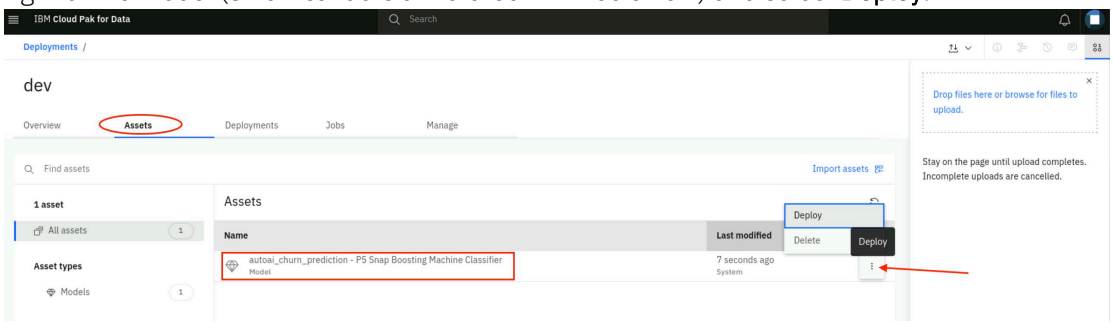
14- Back on the *Promote to space* page, keep the default selected version (Current), add a Description (optional), add tags (optional) and click **Promote** (annotated with red arrow).




15- Once the model is successfully promoted to the deployment space, you will see a notification message (annotated with red rectangle). Click on the **deployment space** link (annotated with red arrow) to navigate to the deployment space.

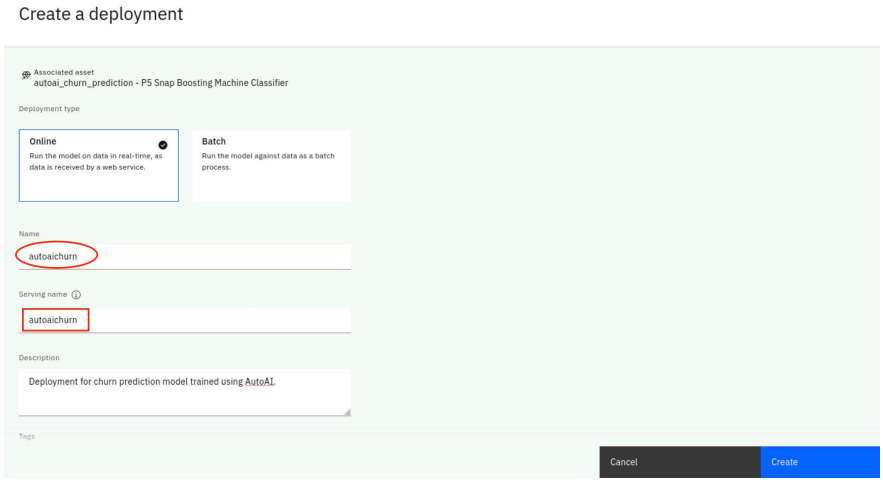



16- On the **dev** deployment space page, select the **Assets** tab (annotated with red oval) and you will see the AutoAI model (annotated with red rectangle) you just promoted to the deployment space (your model name may be different from what is in the screenshot below). Click on the list of options to the right of the model (3 vertical dots annotated with red arrow) and select **Deploy**.

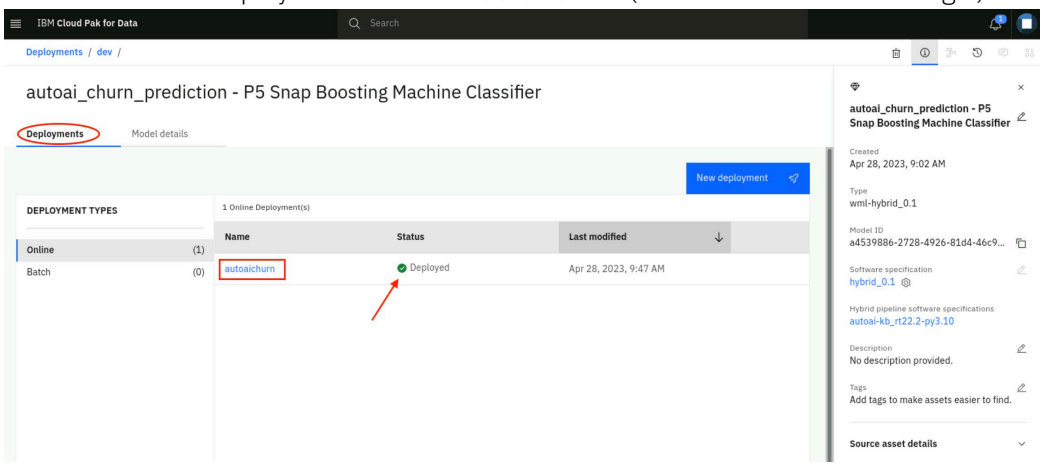



17- On the **Create a deployment** page, select **Online** (annotated with red rectangle), add a Name and a Description (optional) as well as any tags you like (optional) and click **Create** (annotated with red

arrow).



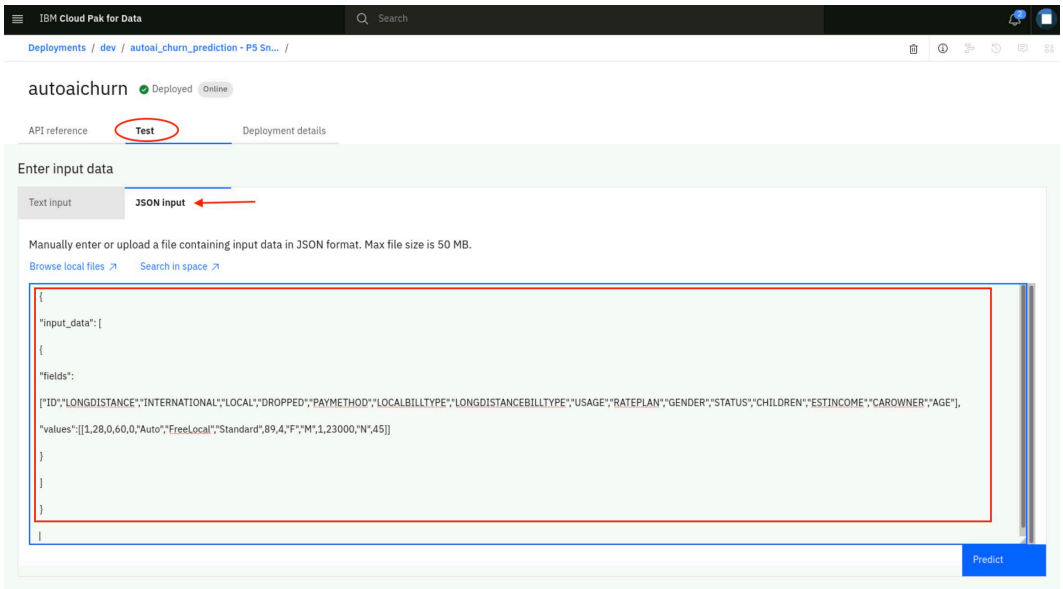

18- Wait for the Deployment Status to change from **In progress** to **Deployed** (annotated with red arrow). Then click on the deployed model name **autoaichurn** (annotated with red rectangle).




19- On the model page **API reference** tab (annotated with red rectangle), review the model Endpoint and the various Code snippets (in different languages) to illustrate how to make an API call to the deployed model. Then select the **Test** tab (annotated with red oval), click on the **JSON input** tab (annotated with red arrow), paste the following JSON sample in the window and click **Predict** (blue button). *Note* please be careful as you copy / paste the value below. Special characters like double quotes may not copy correctly which may cause the prediction to give an error. Feel free to copy/paste from AutoAI Payload Sample box note.

```
{
"input_data": [
{

"fields":["ID","LONGDISTANCE","INTERNATIONAL","LOCAL","DROPPED","PAYMETHOD","LOCALBILLTYPE","LONGDISTANCEBILLTYPE","USAGE","RATEPLAN","GENDER","STATUS","CHILDREN","ESTINCOME","CAROWNER","AGE"],
"values":[[1,28,0,60,0,"Auto","FreeLocal","Standard",89,4,"F","M",1,23000,"N",45]]
}
]
}
```



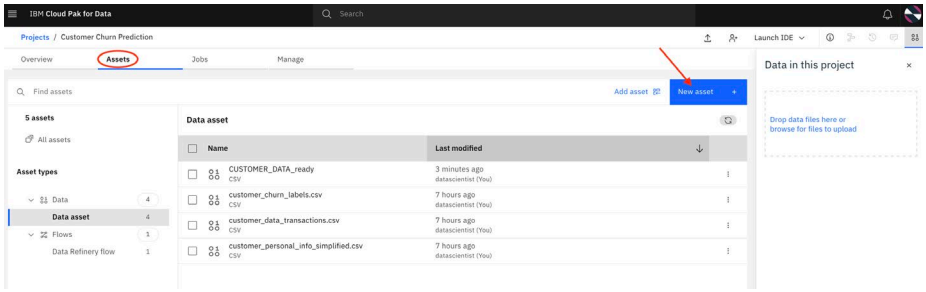

The deployed model will predict the likelihood of the user to churn given the specific values for the various features. The model returns the predicted churn label as “T” (true) or “F” (false) and the probability of that prediction which effectively expresses the likelihood of that user to churn (or not). A “T” label returned by the model indicates the user is likely to churn and the corresponding probability. These probabilities can be used in conjunction with the predicted label to better serve customers on a more granular basis. Your application can be customized to make decisions based on the predicted label and the probabilities of that prediction.

# Option 2: Train Churn Prediction Model with Jupyter Notebook

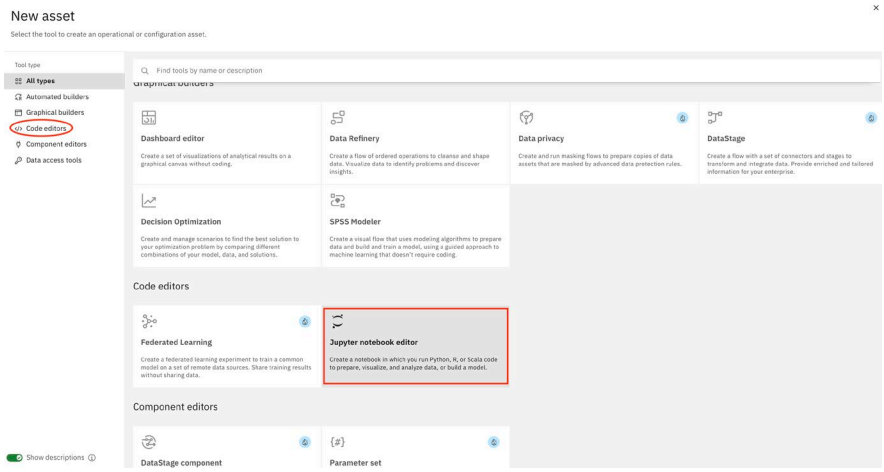
In this section, we illustrate an alternate method for training AI models in Cloud Pak for Data, namely by using Jupyter notebook and open-source libraries. This is a very common and mostly preferred method by data scientists as it provides them with the utmost flexibility in exploring different algorithms for training best performing AI models.

**Review and run the Notebook**

1- Log into Cloud Pak for Data as **datascientist** user.
2- Navigate to your **Customer Churn Prediction** project.
3- Click **Assets** tab (annotated with red oval) and click **New asset** + (annotated with red arrow).



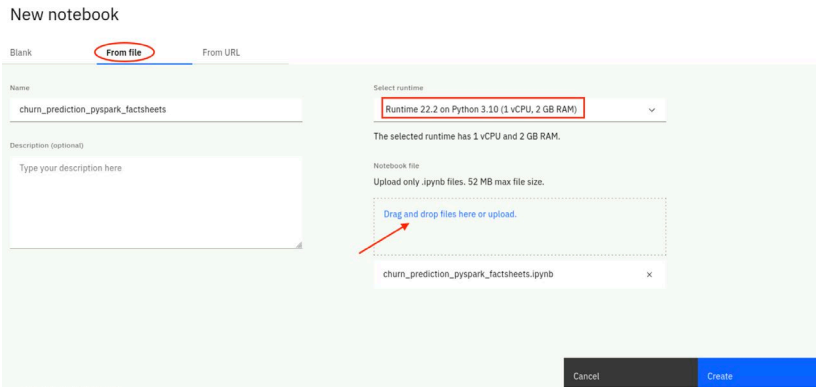

4- Scroll down and select the **Jupyter notebook editor** (annotated with red rectangle). Note that you can filter asset types by selecting the **Code editors** (annotated with red oval) to quickly find Jupyter notebook editor.



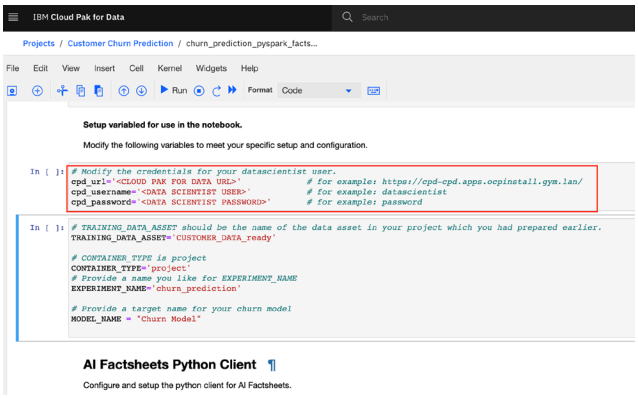

5- On the New notebook page, click **From file** tab (annotated with red oval) and click the **Drag and drop files here or upload** (annotated with red arrow). Select the churn_prediction_pyspark_factsheets.ipynb notebook to upload. You should have downloaded this notebook from box to your computer (if not, download churn_prediction_pyspark_factsheets.ipynb now).

   a. Verify the selected runtime is IBM Runtime 22.2 on Python 3.10 (1 vCPU, 2GB RAM) (annotated with red rectangle).

b. Click **Create**.



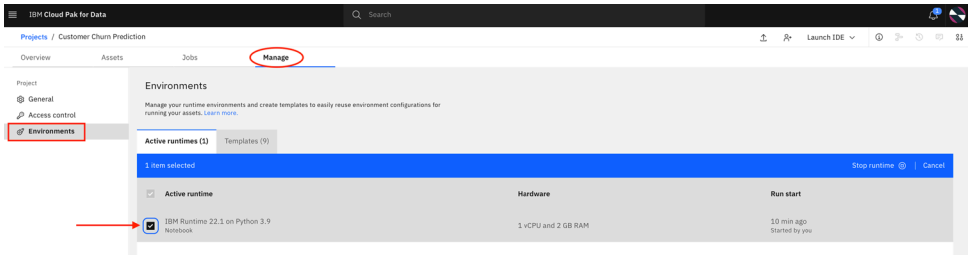

6- Before running the notebook, you need to make a few edits.
   a. Update **cpd_url**, **cpd_username** and **cpd_password** variables to reflect the specifics of your environment. This is required.
   b. Update the TRAINING_DATA_ASSET name if you used a name other than CUSTOMER_DATA_ready for the output of your data refinery job in the earlier section.
   c. Optional: Update the names for EXPERIMENT_NAME and MODEL_NAME to any names of your choice.



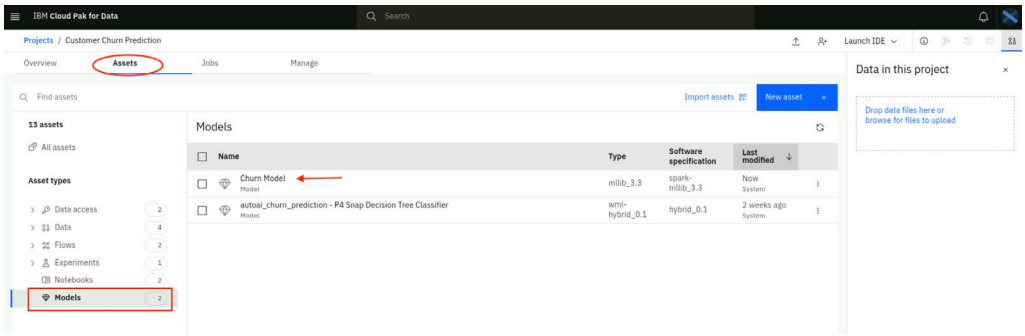

7- Starting from the first cell in the notebook, review the documentation and run through the steps in the notebook. When you complete all the steps, you would have done the following:
   a. Use the Factsheets Python client to register the model with the model entry and auto-log the training facts.
   b. Accessed the CUSTOMER_DATA_ready data from your project.
   c. Processed the data to prepare features relevant for the prediction.
   d. Trained a Random Forest ML model to predict the likelihood of customers to churn using a sample of the data.
   e. Evaluated the model against test data not used in training.
   f. Associated Watson Machine Learning with your Customer Churn Prediction project.
   g. Stored the model in the project.
   h. Collected factsheet data for the model.

8- As you read and execute the various cells in the notebook, note that the notebook was written in a manner to support being executed independently (as you're currently doing) as well as trigger by a Watson Pipeline which you will run in the next module. The difference between the two approaches is minor and is mainly related to how the training data is provided for the model.

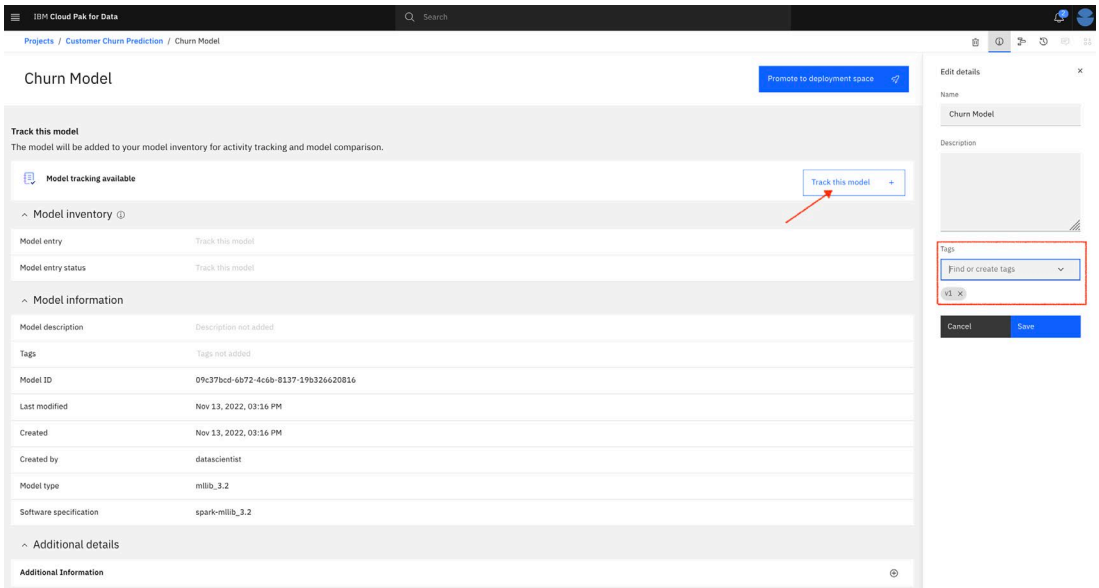
9- After the notebook completes, stop the notebook environment to save resources since this is a shared environment. To do so, navigate back to the **Customer Churn Prediction** project, select the **Manage** tab (annotated with red oval), and click **Environments** (annotated with red rectangle). Then select the check box next to the active environment and click **Stop runtime**.
Click **Stop** on the confirmation pop-up window.



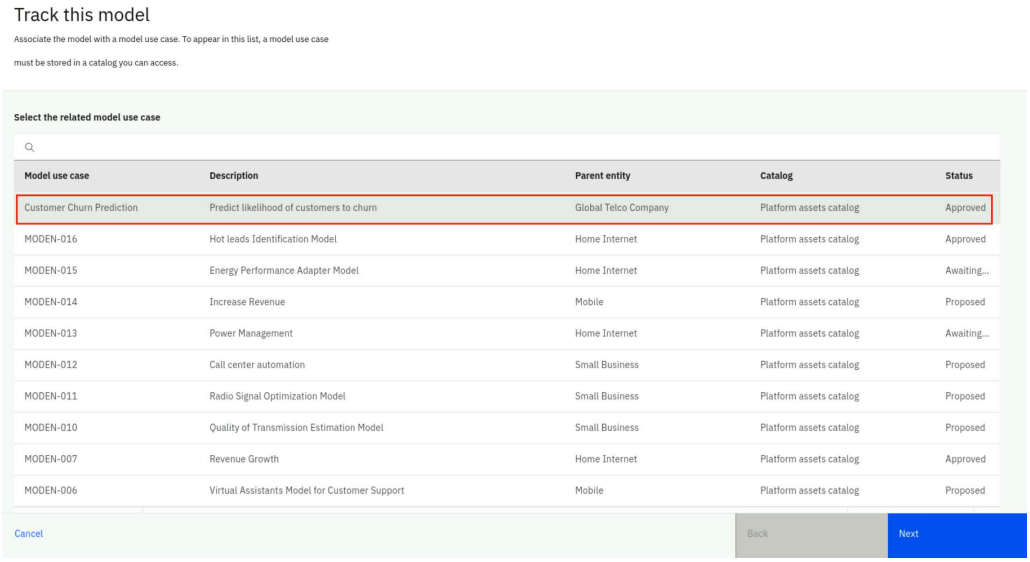

10- In your project, navigate back to the **Assets** tab (annotated with red oval), select **Models** (annotated with red rectangle) and then click on the **Churn Model** (annotated with red arrow) to open up that model.



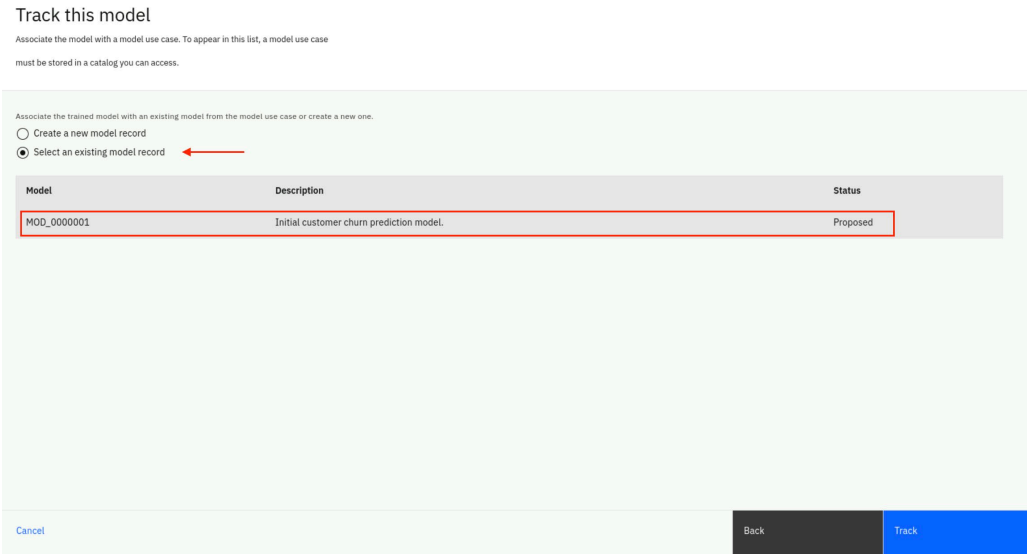

11- On the Churn Model page, review the collected information about the model (scroll down and up to review the factsheet data). In the Tags field (annotated with red rectangle), add the tag v1 (by typing v1 and hitting Enter) and click **Save**.
After that, click **Track this model** button (annotated with red arrow).



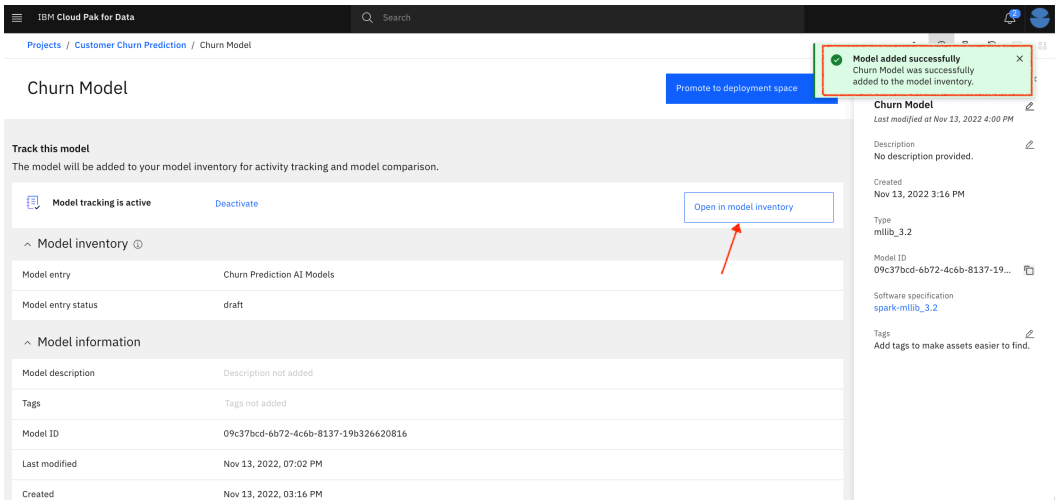
12- On the *Track this model* page, select the **Customer Churn Prediction** model use case (annotated with red rectangle) and click **Next**.



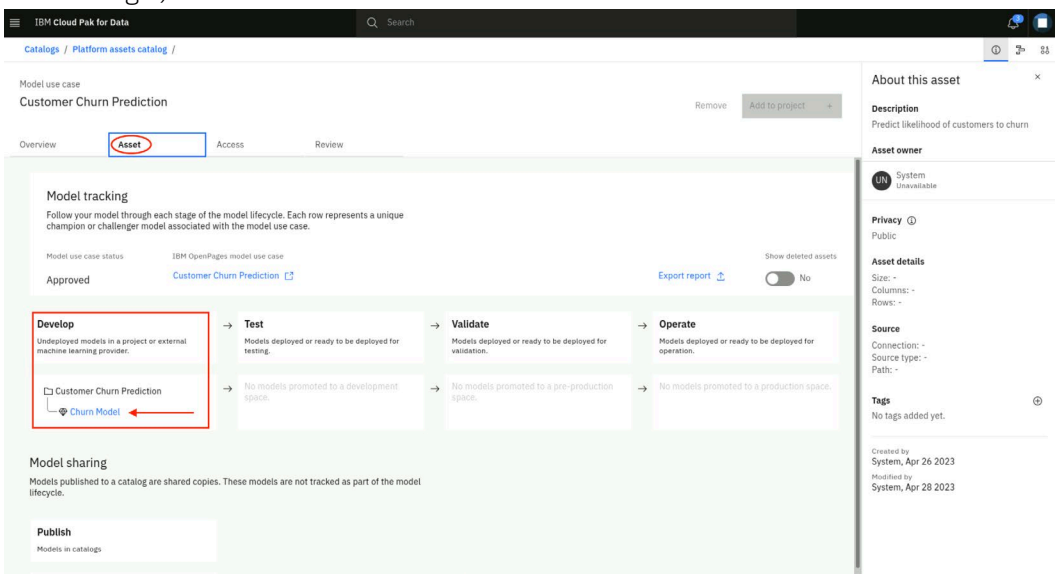

13- On the Track this model page, click the radio button next to **Select an existing model record** (annotated with red arrow) which would like of existing models. Then, select the **MOD_0000001** model (annotated with red rectangle) and click **Track**. Remember **MOD_0000001** is the model that you had created earlier in OpenPages.



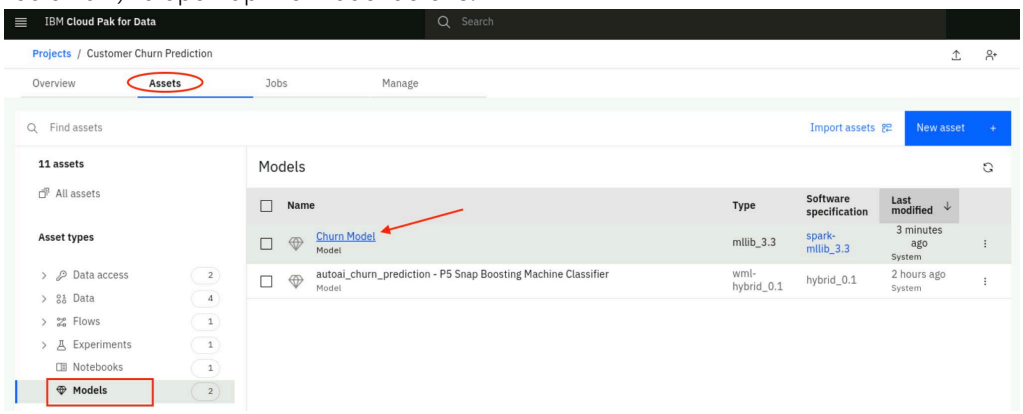

14- Verify you get the notification that that model was added successfully to the inventory (annotated with red rectangle). Click the **Open in model inventory** button (annotated with red arrow) to view the model details and status in its lifecycle as tracked by the model inventory.



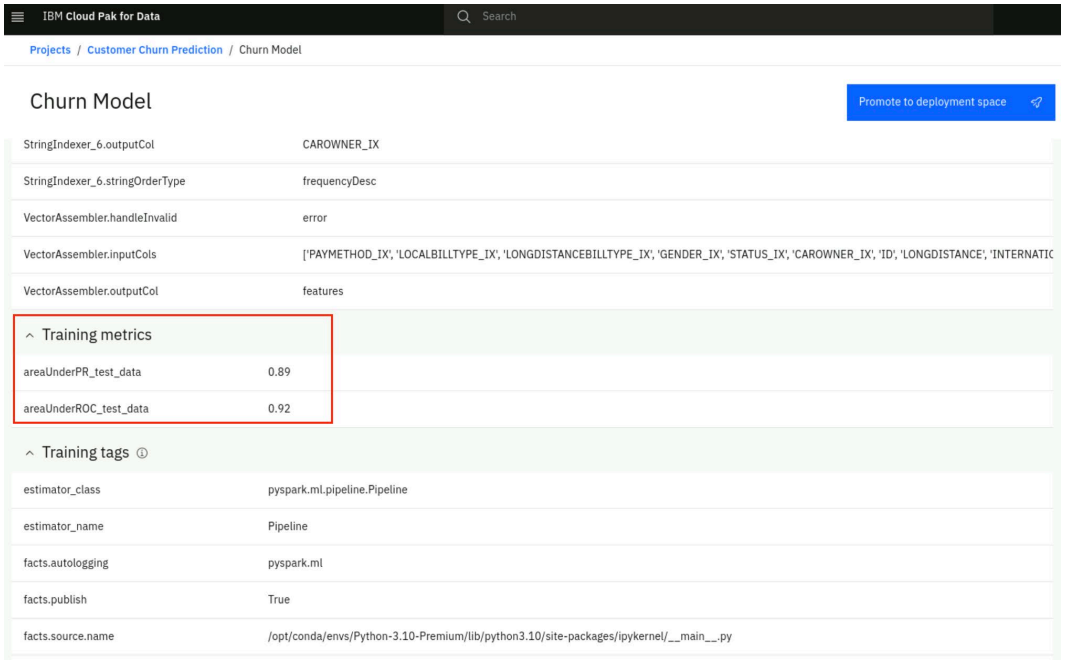

15- On the model use case page, click the **Asset** tab (annotated with red oval) and note that the **Churn Model** (annotated with red arrow) is now being tracked and it is in the **Develop** stage (annotated with red rectangle).



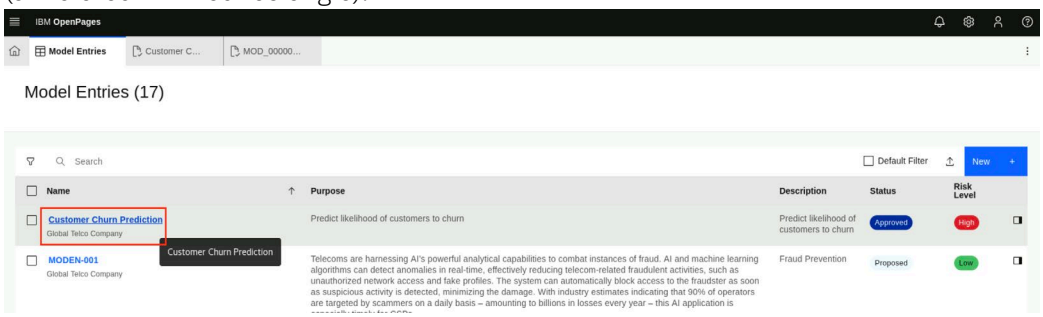

16- Navigate back to your Customer Churn Prediction project, click the **Asset** tab (annotated with red oval) and filter by **Model** asset type (annotated with red rectangle). Click the **Churn Model** (annotated with red arrow) to open up the model details.



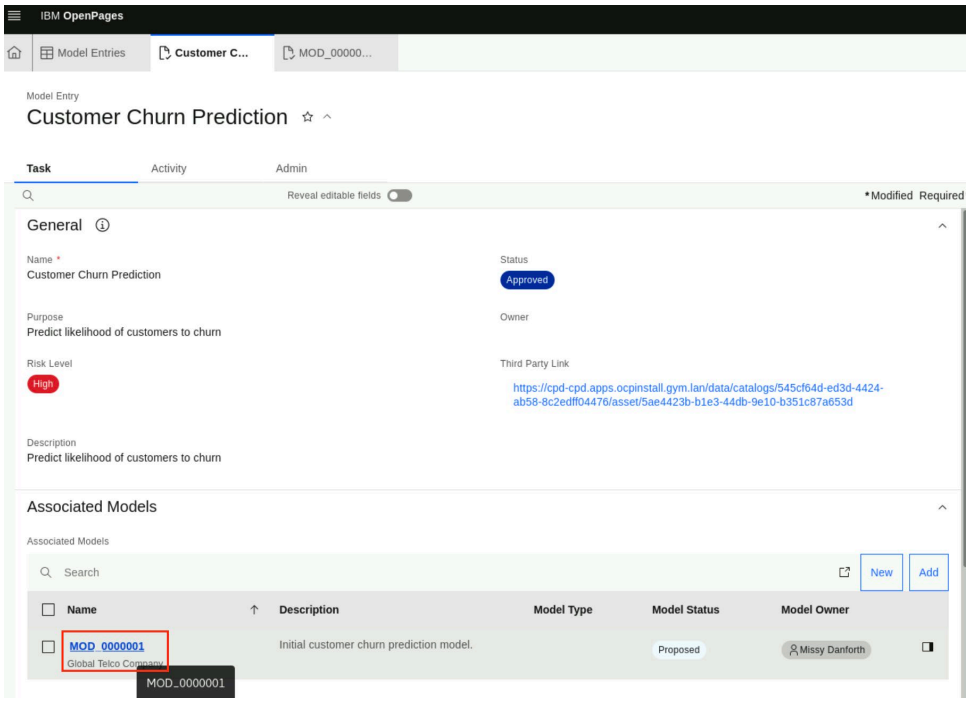
17- Review the captured details of the model such as who created the model, when it was created/modified, current status, what OpenPages model it maps to and several other details. Scroll down to review training information and specifically, check the Training metrics section where important metrics such as areaUnderROC_test_data and areaUnderPR_test_data have been automatically captured.



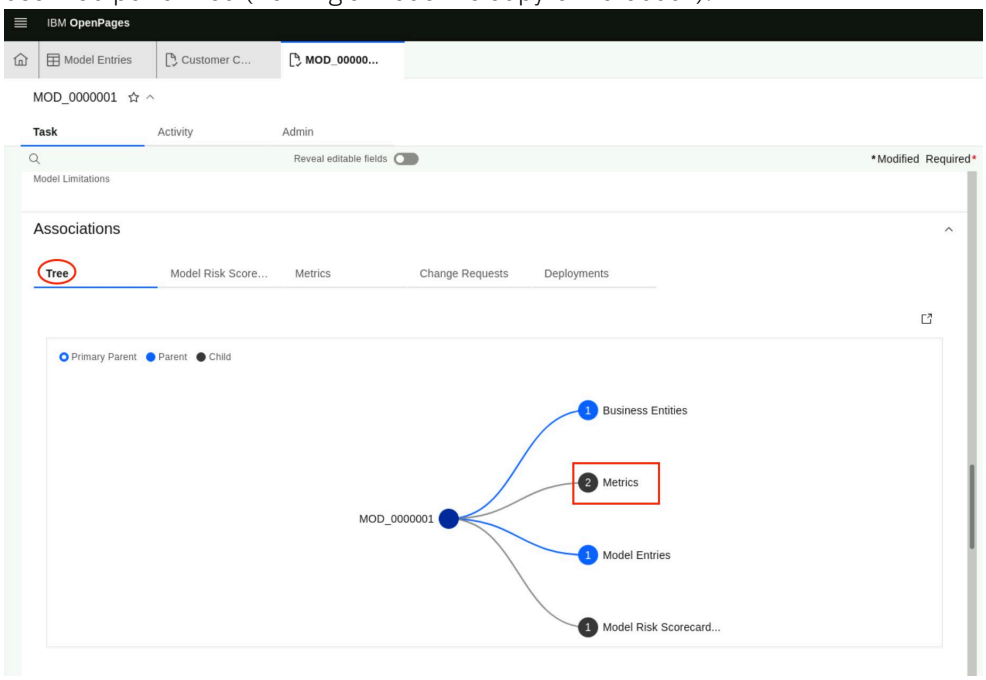

18- In a different window, navigate to OpenPages if it is still open. If closed, log back into Cloud Pak for Data as admin user, navigate to Services ➔ Instances and then click the link to launch OpenPages. In OpenPages, navigate to Model Entries, either from Home dashboard view or from navigation menu (top left) ➔ Inventory ➔ Model Entries, then select the **Customer Churn Prediction** model entry (annotated with red rectangle).



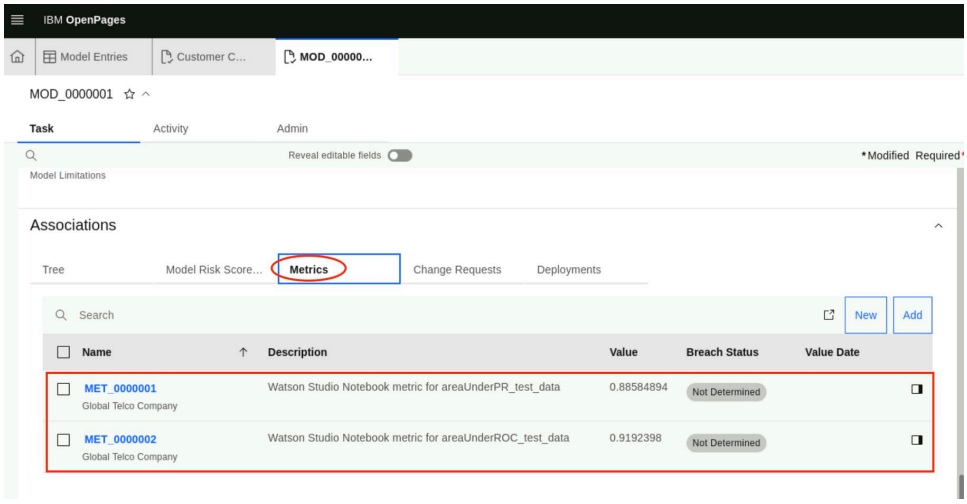

On the Customer Churn Prediction model entry, scroll down and click **MOD_0000001** model (annotated with red rectangle) to launch that.



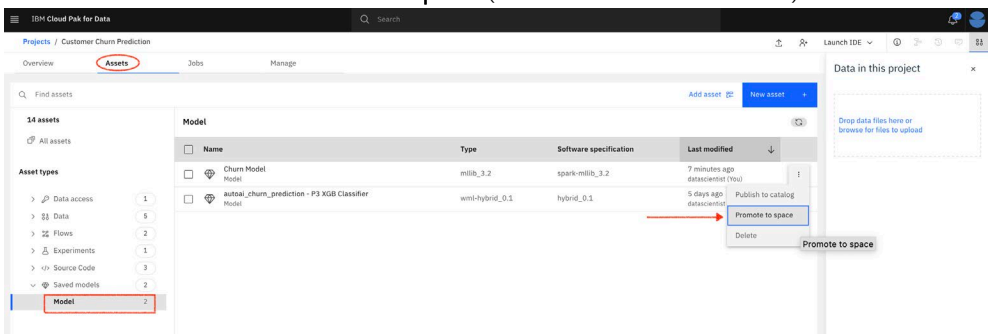

19- In the Associations section, under the Tree tab (annotated with red oval), observe there is a new component associated with this mode, the Metrics component (annotated with red rectangle). This component was imported and synchronized automatically based on the activity that the datascientist user had performed (training a model via Jupyter notebook).



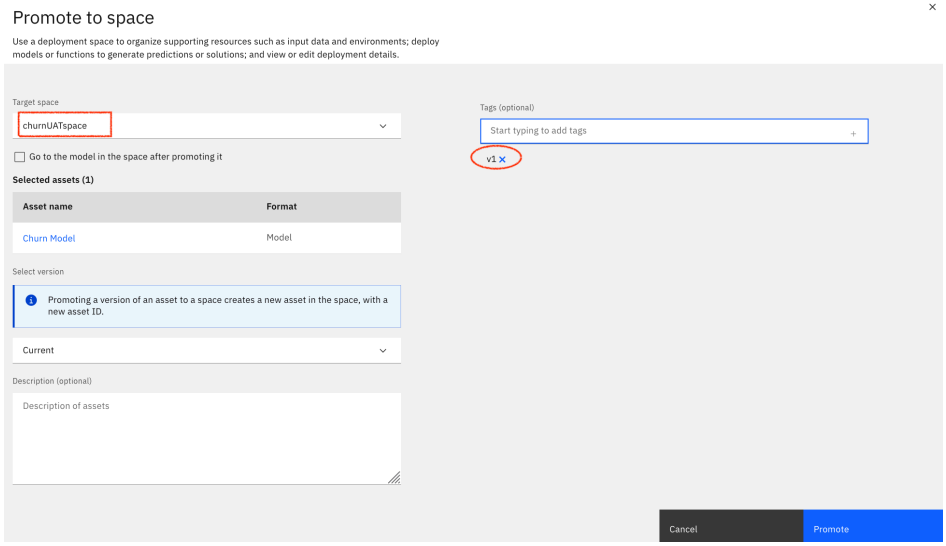

20- Click the **Metrics** tab (annotated with red oval) and review the metrics (annotated with red rectangle) which were automatically captured from the model training step (automation).



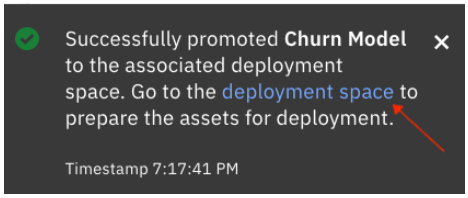

21- Back on your project page in Cloud Pak for Data, click the actions menu (3 vertical dots) next to the **Churn Model** and click **Promote to space** (annotated with red arrow).



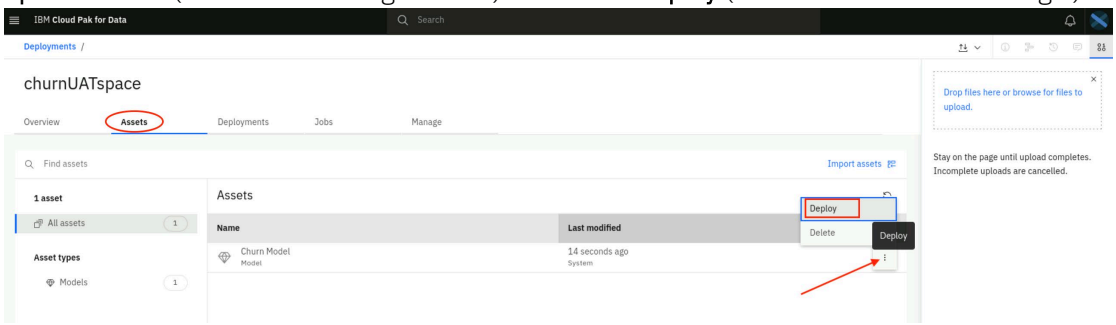

22- On the **Promote to space** page, select the **Target space** as **churnUATspace** (annotated with red rectangle), add a tag of **v1** (annotated with red oval) and click **Promote**.
If the churnUATspace does not show up in the drop-down list, create it in a similar manner to how you created the **dev** space earlier (Select **Create a new deployment space**, and on the pop-up window, provide a name and a description (optional) and click **Create.** After space is created, click **Close**).



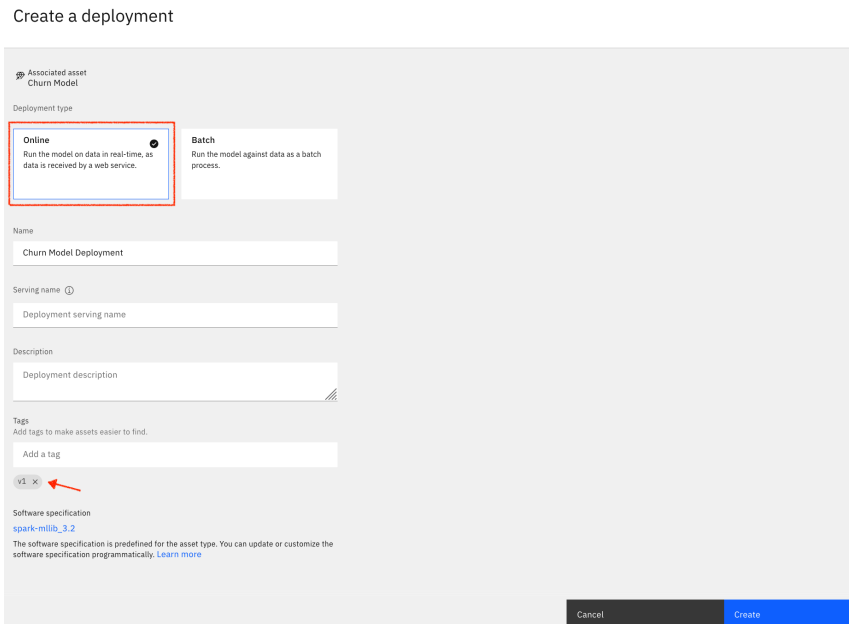
23- Verify you get the notification that the Churn Model is successfully promoted and click the **deployment space** link (annotated with red arrow) to view the model in the churnUATspace.



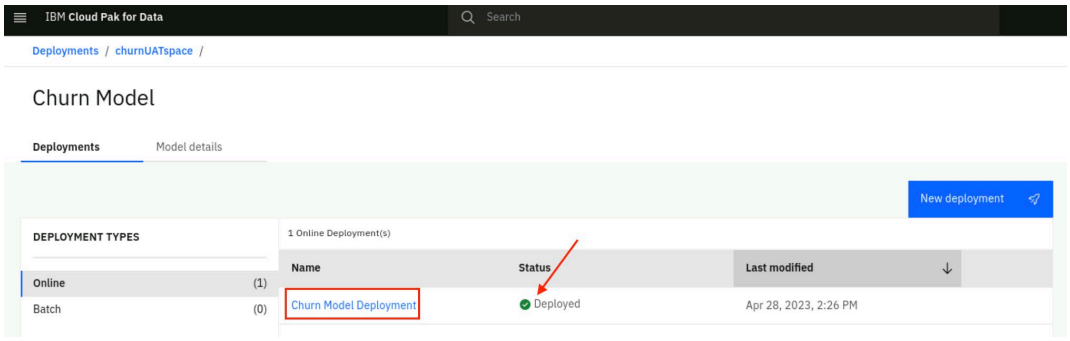

24- On the churnUATspace, click **Assets** tab (annotated with red oval), click the **open and close list of options** menu (annotated with right arrow) and select **Deploy** (annotated with red rectangle).



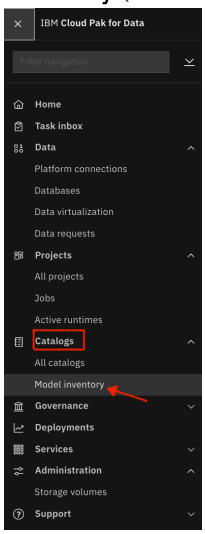

25- On the **Create a deployment** page, select **Online** (annotated with red rectangle) for the type of deployment, provide a **Name** (for ex., Churn Model Deployment), add a **v1** tag (annotated with red arrow) and click **Create**.



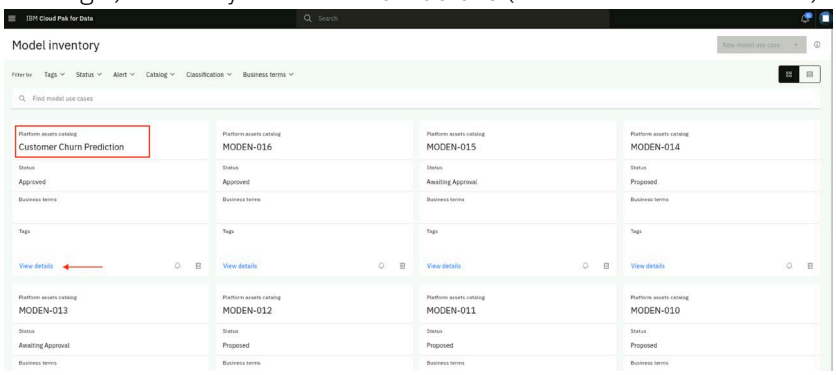

26- Wait for the status to change from **In progress** to **Deployed** (annotated with red arrow) and then click the deployment name, **Churn Model Deployment** (annotated with red rectangle). *Optional Once that page loads, you can try testing the deployed model like you did previously.



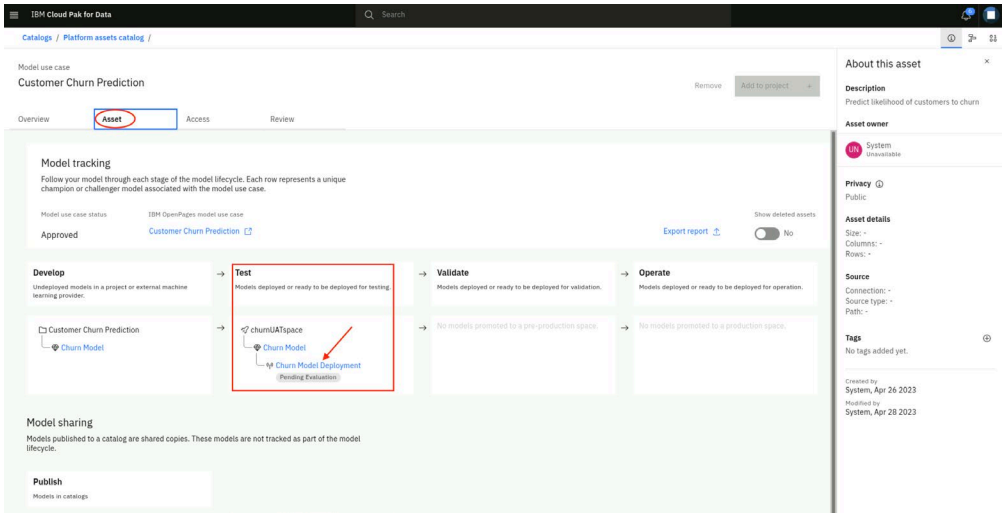

27- At this time, navigate back to the Model inventory by clicking the navigation menu and selecting **Model inventory** (annotated with red arrow) under **Catalogs** (annotated with red rectangle).



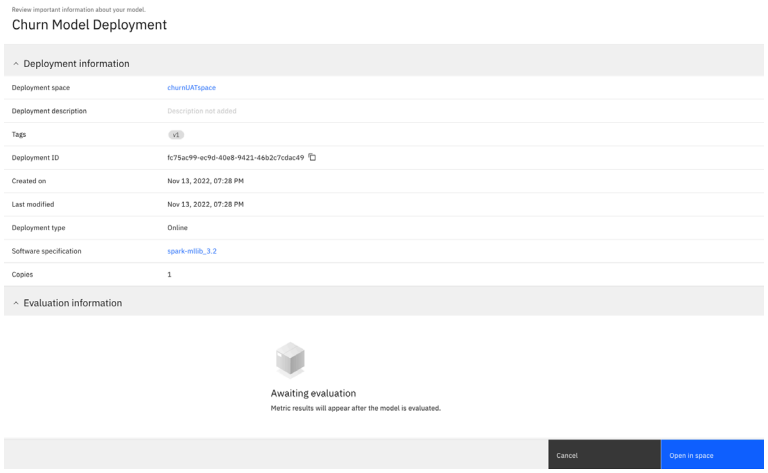

28- On the *Model Inventory* page, find the **Customer Churn Prediction** model use case (annotated with red rectangle) inventory and click **View details** (annotated with red arrow).



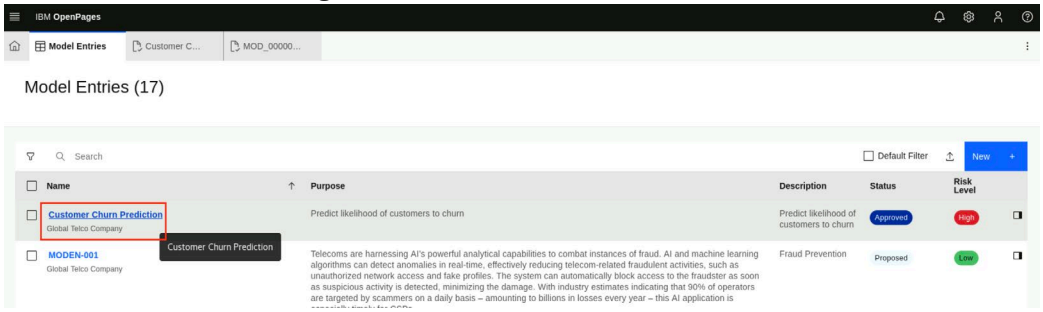

29- On the *Customer Churn Prediction* model use case page, select the **Asset** tab (annotated with red oval) and note that the Churn Model is now in **Deploy** state (annotated with red rectangle). Effectively, once the Churn Model was deployed in a deployment space, its status moves to the Deploy state. Also note that the model inventory now highlights that the model is Pending Evaluation which we'll cover in a subsequent module where OpenScale is used for validating AI models for quality, fairness, and explainability. Click the **Churn Model Deployment** (annotated with red arrow) to review the details of the model deployment.



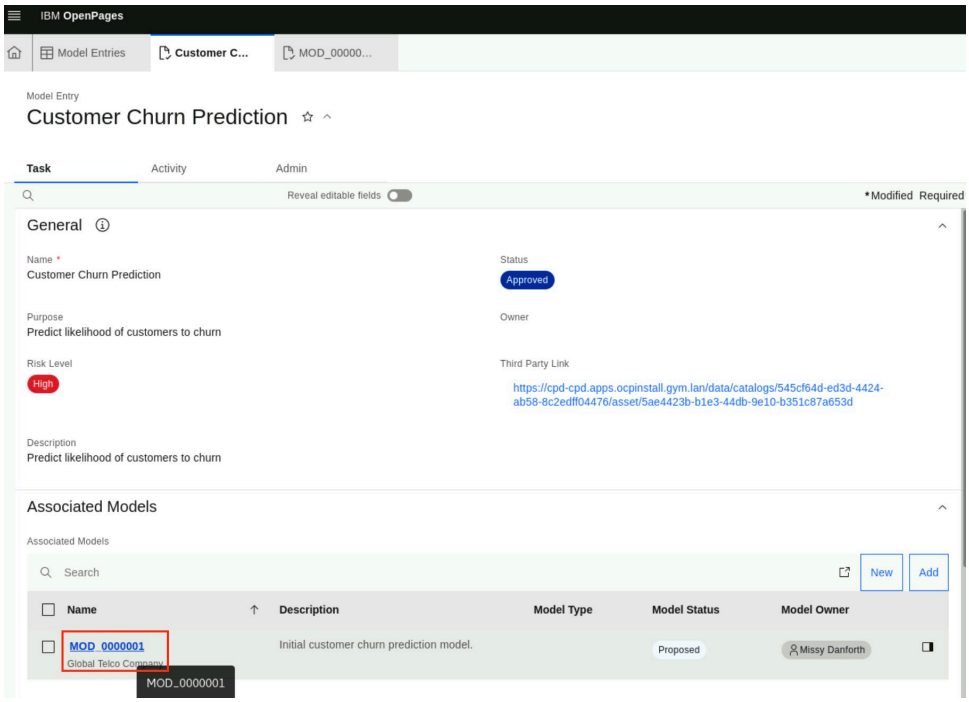

30- Take a minute to review the Churn Model Deployment details.



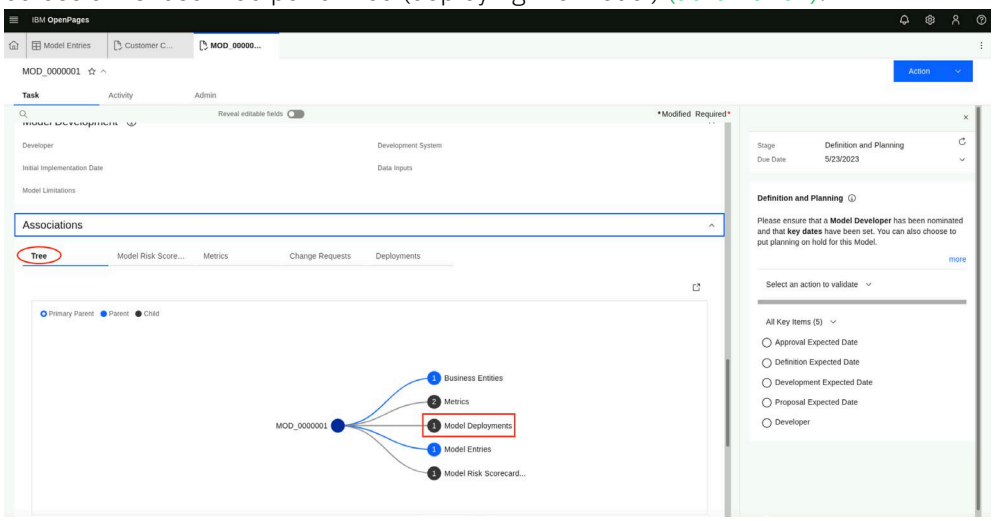

31- Navigate to the browser window with OpenPages if it is still open. If closed, log back into Cloud Pak for Data as admin user, navigate to Services ➔ Instances and then click the link to launch OpenPages. In OpenPages, navigate to Model Entries, either from Home dashboard view or from navigation menu (top left) ➔ Inventory ➔ Model Entries, then select the **Customer Churn Prediction** model entry (annotated with red rectangle).



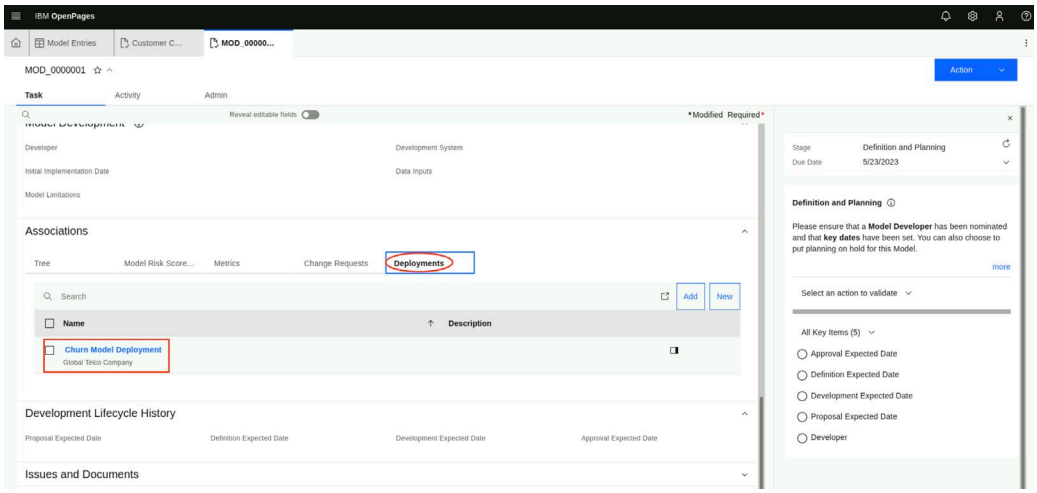

On the Customer Churn Prediction model entry, scroll down and click **MOD_0000001** model (annotated with red rectangle) to launch that.




32- In the Associations section, under the Tree tab (annotated with red oval), observe there is a new component associated with this mode, the Deployments component (annotated with red rectangle). This component was imported and synchronized automatically based on the activity that the datascientist user had performed (deploying the model) (automation).




33- Click the **Deployments** tab (annotated with red oval) and note the Churn Model Deployment (annotated with red rectangle) which is automatically captured from the model deployment step executed by the datascientist user.




This concludes this module. In the next modules we'll continue with additional functions achieved with different services.

## Summary

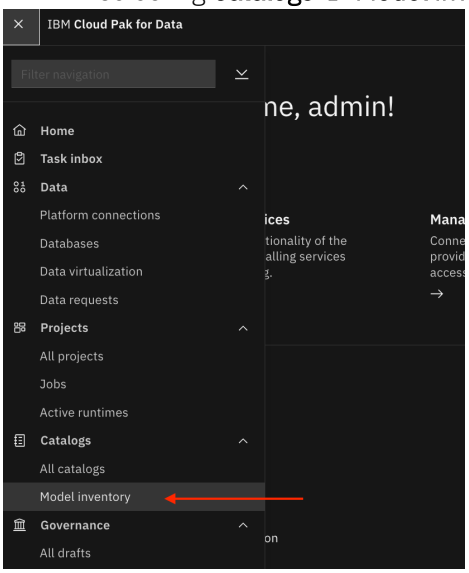
In this module, we've reviewed a typical flow for training AI models starting with finding relevant data assets from the catalog, preparing the data using Data Refinery, and training AI models for churn prediction using different approaches including AutoAI and Jupyter notebooks. We also demonstrated how the model factsheets are tracked through its lifecycle (covered develop and test steps) as well as the details are automatically synchronized to the AI Governance tool (OpenPages).

# Appendix

### Model Inventory Config

In this section, you configure Model Inventory to disable synchronization with OpenPages.

1- If you're logged out, navigate your favorite browser to Cloud Pak for Data url and login as ***admin*** user.
2- Navigate to Model inventory by clicking the Navigation menu (top left hamburger icon) and selecting **Catalogs ➔ Model inventory** (annotated with red rectangle




3- On the *Model inventory* page, select the **Manage** tab (annotated with red oval) and disable synchronization with OpenPages by making sure the **Sync with IBM OpenPages** slider is off (annotated with red arrow).