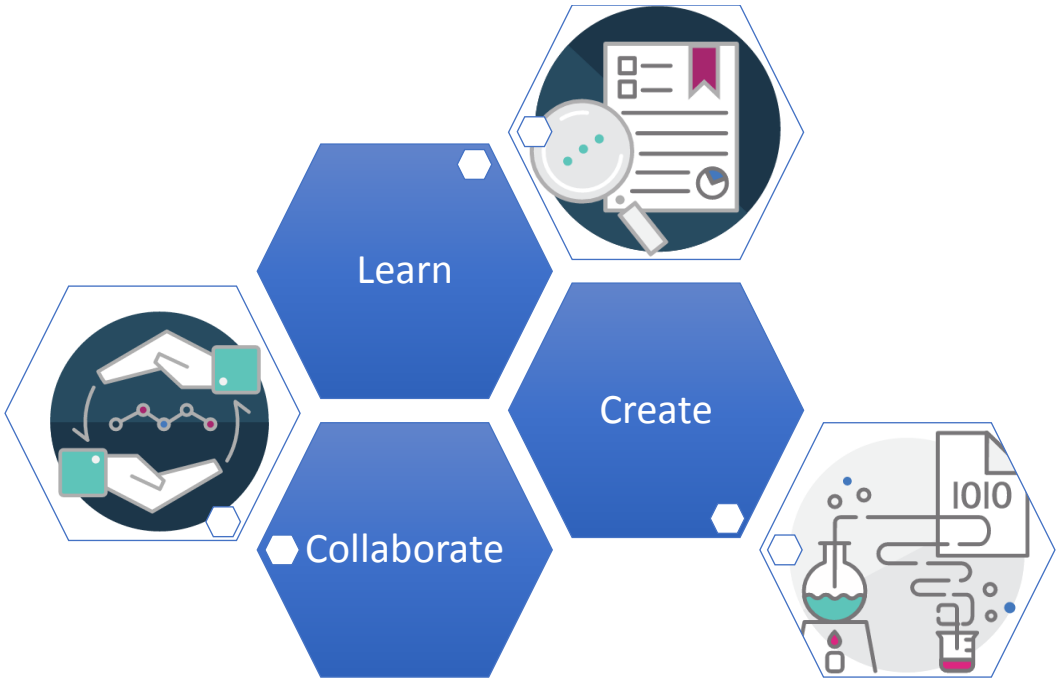
IBM Cloud

Data Science Experience (DSX)

Practical Hands-on Introduction

Lab Guide



Data Science

Machine Learning

Notices and Disclaimers

The information contained in these materials is provided for informational purposes only, and is provided AS IS without warranty of any kind, express or implied. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, these materials. Nothing contained in these materials is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software. References in these materials to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates. This information is based on current IBM product plans and strategy, which are subject to change by IBM without notice. Product release dates and/or capabilities referenced in these materials may change at any time at IBM's sole discretion based on market opportunities or other factors, and are not intended to be a commitment to future product or feature availability in any way.

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

IBM, the IBM logo and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at ibm.com/legal/copytrade.shtml

Other company, product and service names may be trademarks or service marks of others

Document Revision History

Rev #	File Name	Date
1.0	DSX Hands-on Workshop.docx	11/1/2017

Prepared & Revised by:
Louis Frolio – louis.frolio@ibm.com

Table of Contents

Lab Environment Overview 5

Lesson 1: DSX Signup & Home Page 6

- **Lesson 1: Workflow Overview** 7
- **Lesson 1: Instructions** 8

Lesson 2: Jupyter Notebook 13

- **Lesson 2: Workflow Overview** 14
- **Lesson 2: Instructions** 15

Lesson 3: Machine Learning Flows 26

- **Lesson 3: Workflow Overview** 27
- **Lesson 3: Instructions** 28

Lesson 4: Watson Machine Learning 36

- **Lesson 4: Workflow Overview** 37
- **Lesson 4: Instructions** 38

Lab Environment Overview

Installed Software and Tools

Software	Link
IBM Data Science Experience (DSX)	https://datascience.ibm.com/
IBM SPSS Statistics	http://www-03.ibm.com/software/products/no/spss-stats-base
Jupyter	http://jupyter.org/
GitHub	https:/github.org/
R	https://www.r-project.org

Lesson 1: DSX Signup & Home Page

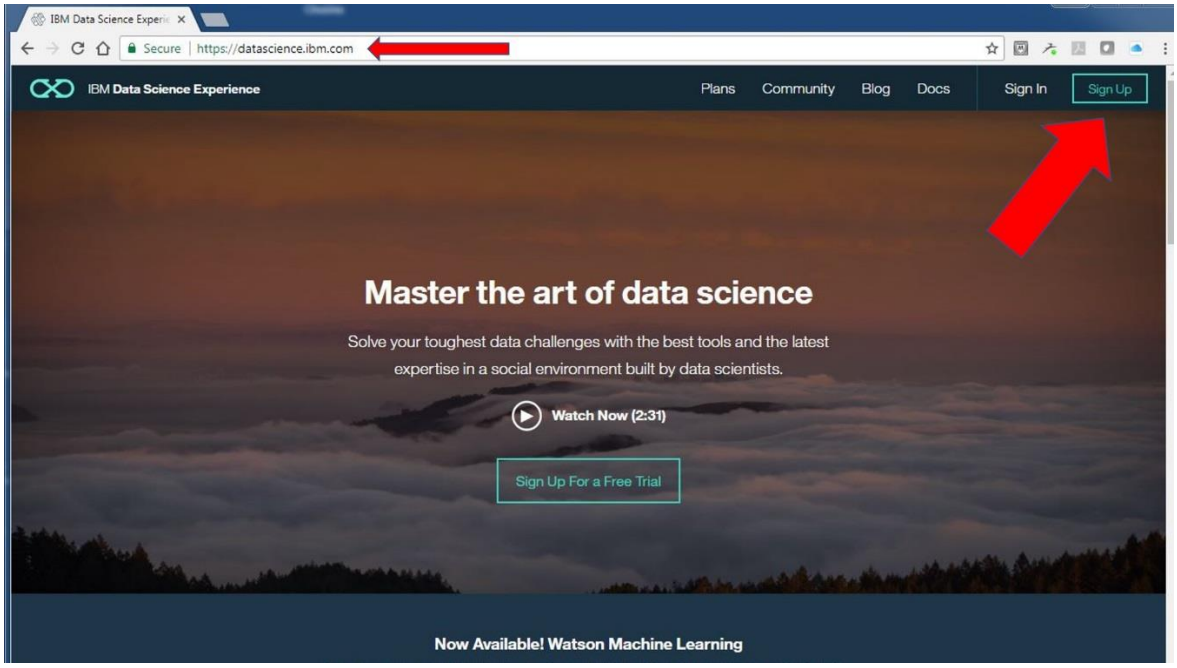
Purpose:	This lab introduces DSX, its sign up and walk-through of the features and functions starting at the Home Page.

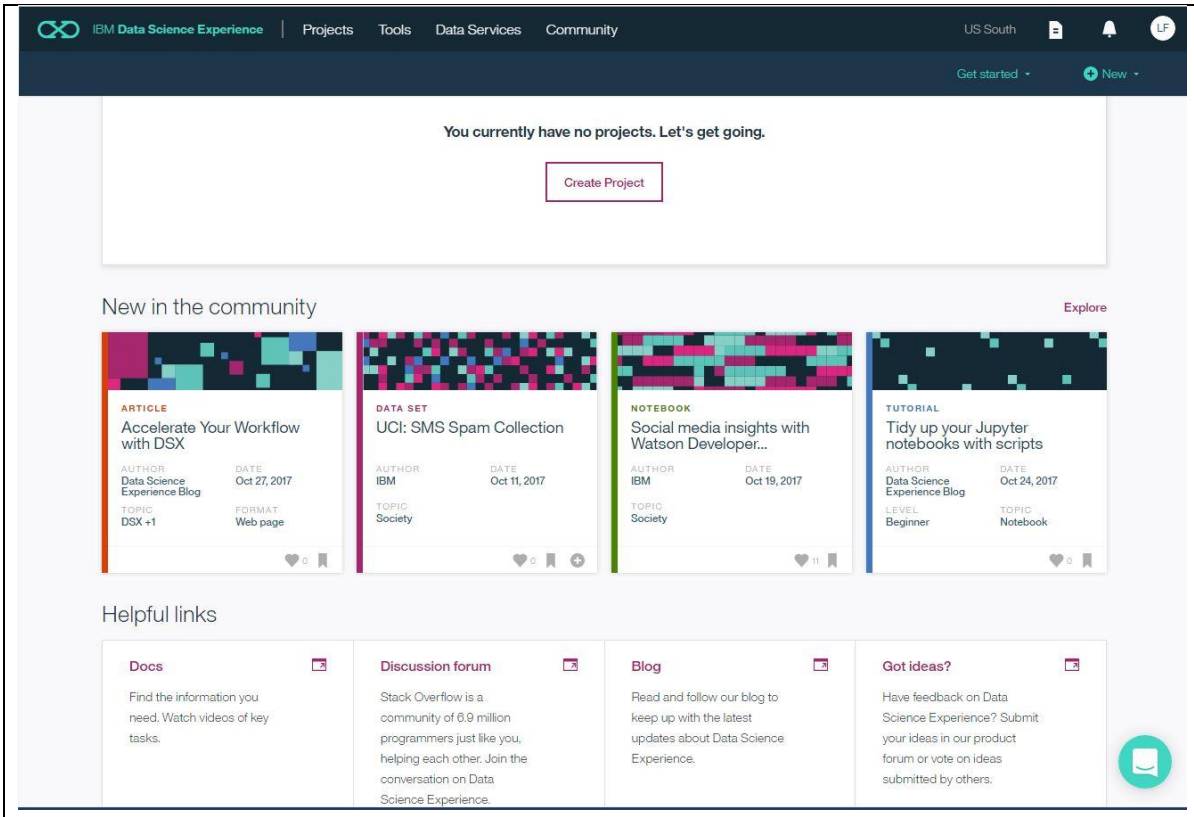
Tasks:	Tasks you will complete in this lab exercise include: • Create/Sign In to DSX Account • Engage Live Chat • Differentiate Four Types of Community Cards • Explore Personal Profile, Apps/Services, and Integrations

Lesson 1: Workflow Overview

1. Create Account / Sign In to DSX
2. Live Chat
3. Community Cards
4. Profile Settings
5. Apps and Services

Lesson 1: Instructions

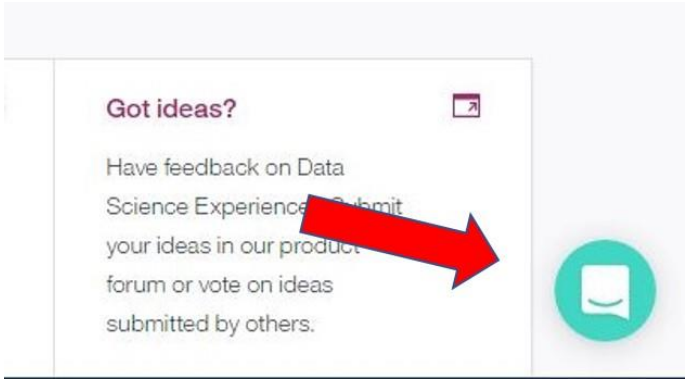
Action
1. Create Account/Sign In to DSX • Open web browser and navigate to: **https://datascience.ibm.com**      • Click on “Sign Up” and you will be prompted for several items of information. After a few moments of self-configuration, you will be brought to your new Home Page:

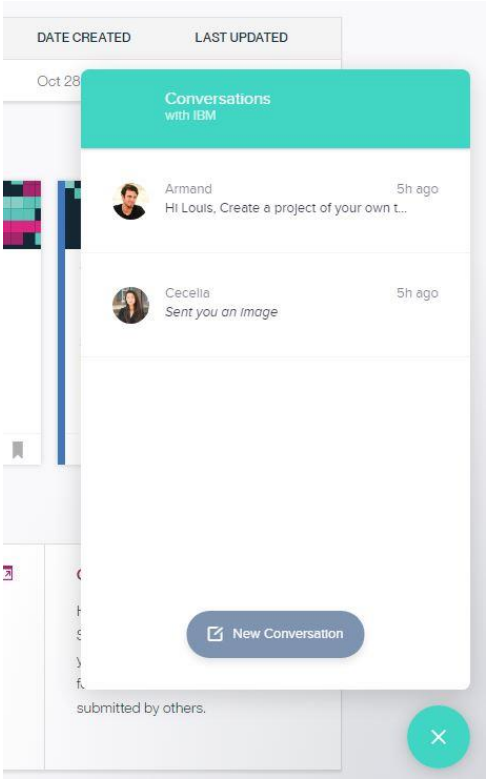


2. Live Chat

This is the home page of IBM Data Science Experience(DSX). Here you have all the tools that you need in a single place to **Learn, Create, and Collaborate**.

- On the bottom right-hand corner, you will see a **Live Chat** feature. Click on the **Chat** icon to launch Live Chat:



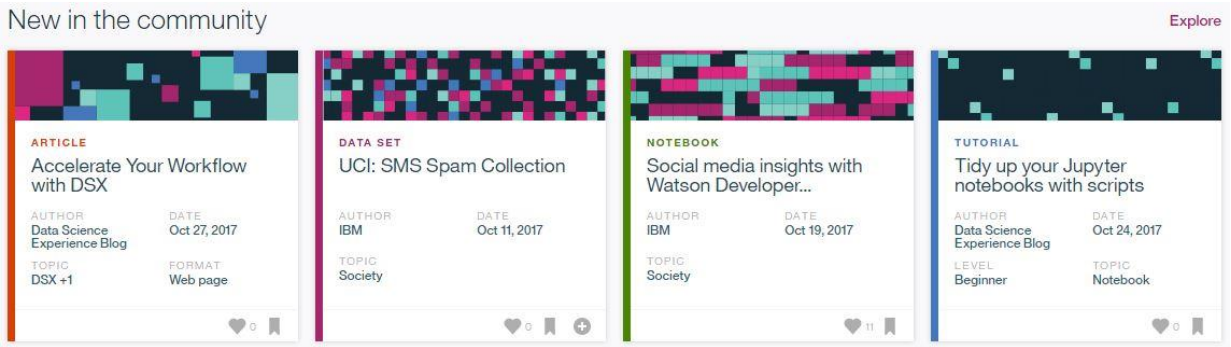


If you need assistance, you need only click on **New Conversation** to connect with a live person. Through this Live Chat feature, you can also continue conversations the next time you log into DSX.

We use feedback captured through **Live Chat** and the offerings instrumentation to guide our decisions in designing and developing **Data Science Experience**. We perform this analysis using DSX.

3. Community Cards

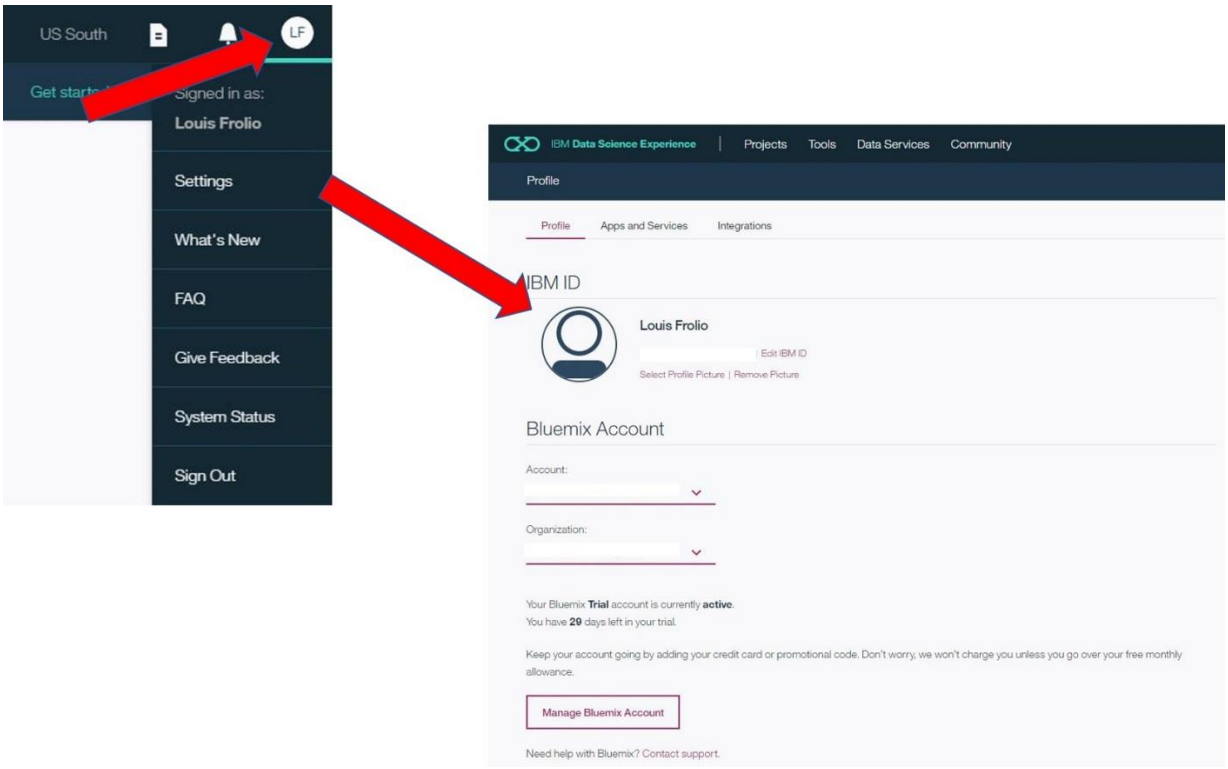
In the middle of the Home Page you see the **Community Cards**:



There are four types of cards – **Articles, Data Sets, Notebooks, and Tutorials**. These are designed to make it easier for you to learn about data science and experiment with its various tools and techniques.

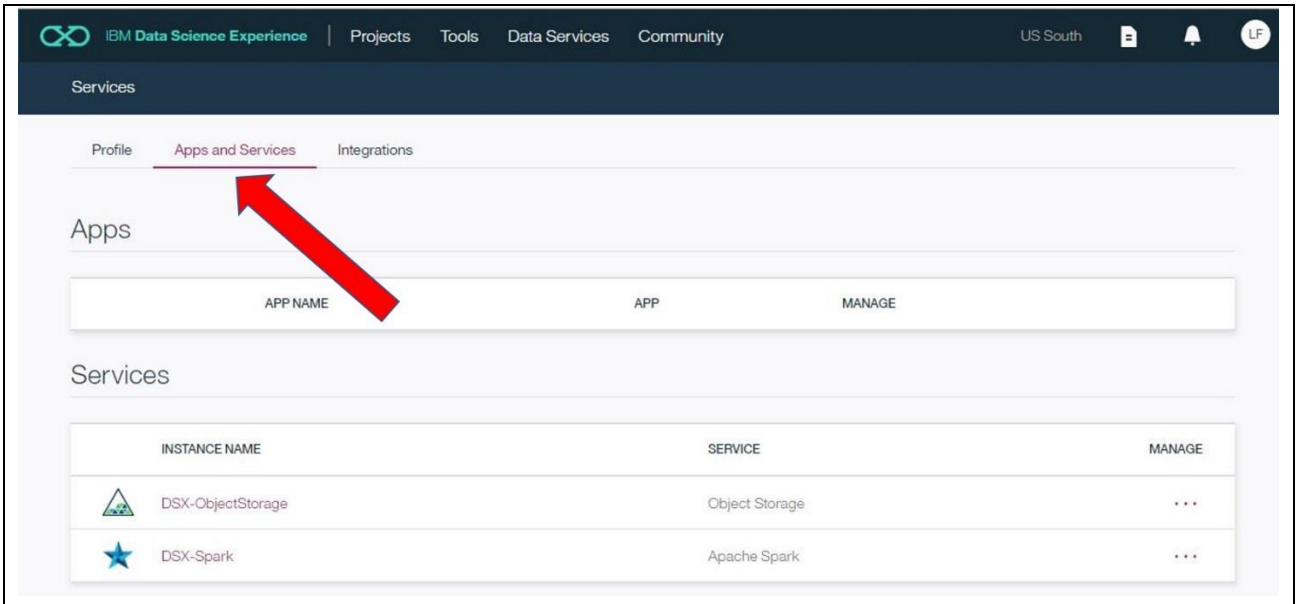
4. Profile Settings

- Click on **Settings** to look at your **Profile, Apps and Services, and Integrations**. This is where you see the details of your Bluemix Account:



5. Apps and Services

- Click on **Apps and Services** to view all your current Bluemix account service instances:



Above is the default for the brand-new account, there is an instance of DSX Object Storage, and DSX Spark.

Integrations is where you configure DSX for GitHub integration.

End of Lesson 1

Lesson 2: Jupyter Notebook

Purpose:	This lesson introduces projects within DSX, their purpose, value, and how they are used to support collaboration. Also, Jupyter notebooks are introduced and used as part of a customer churn analysis using the R programming language.

Tasks:	Tasks you will complete in this lab exercise include: • Create and Configure DSX Project • Add Notebook Asset • Add Data Asset • Create Notebook Reference to Data Asset • Predict Customer Churn using Machine Learning Techniques • Evaluate Model Performance

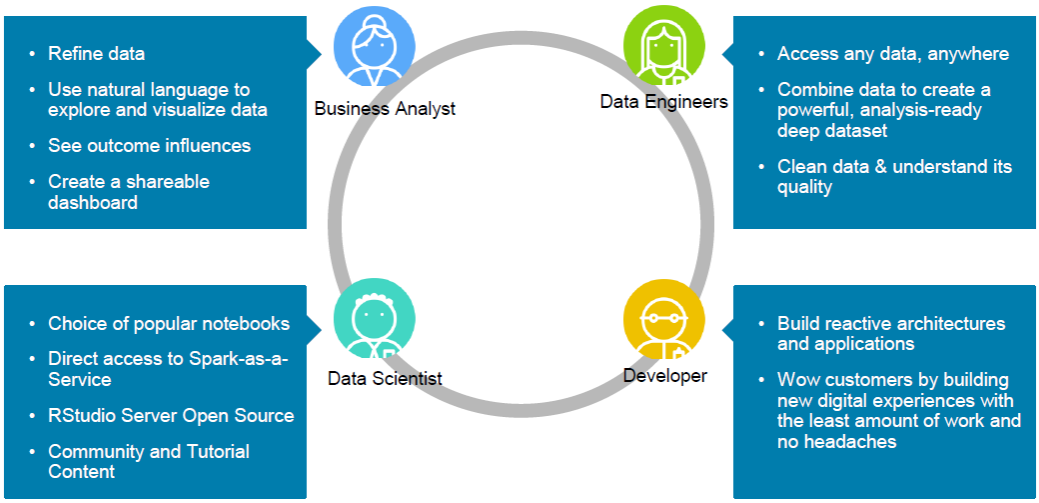
Lesson 2: Workflow Overview

1. Project Overview
2. Create New Project
3. Create Notebook
4. Load Data
5. Bind Notebook to Data Asset
6. Build and Evaluate Customer Churn Model

Lesson 2: Instructions

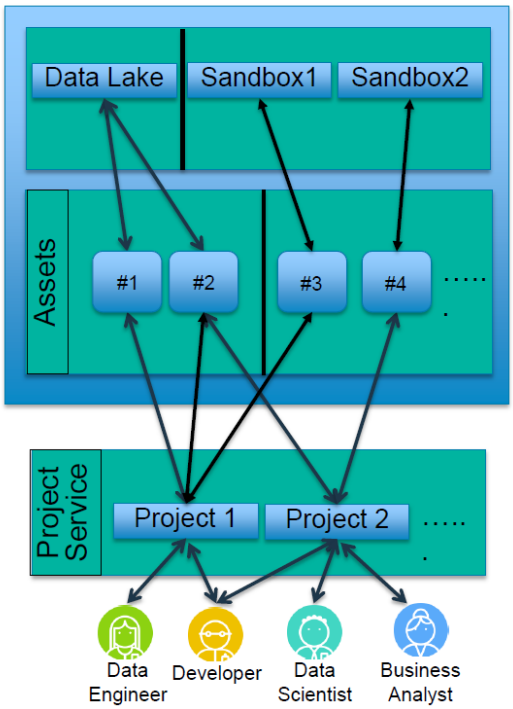
1. Project Overview

Data professionals need purpose-built, self-service communities that enable them to seamlessly collaborate across personas.



Projects make collaboration easier by:

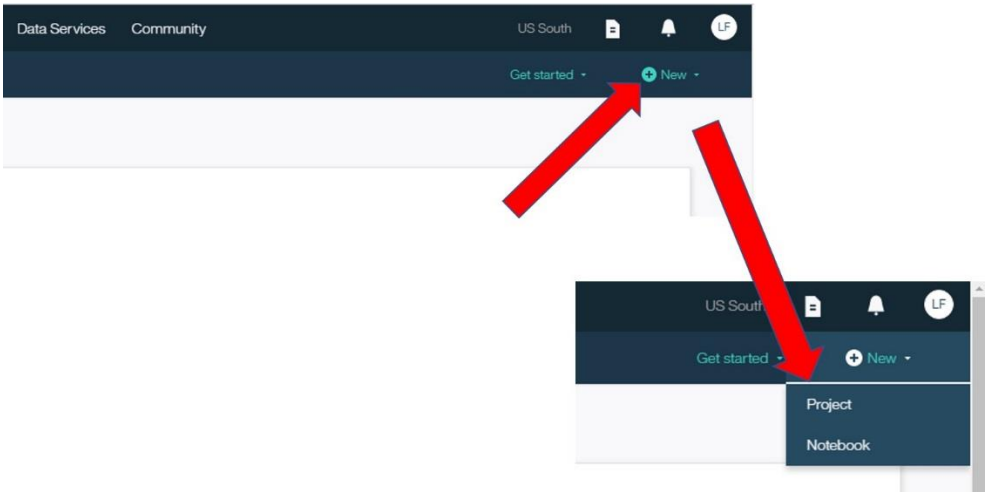
- Allowing different users and personas to share a set of assets
- Enabling users to collaborate and manage their notebooks, artifacts, plus more
- Providing three levels of rights: Viewers, Editors, and Admins



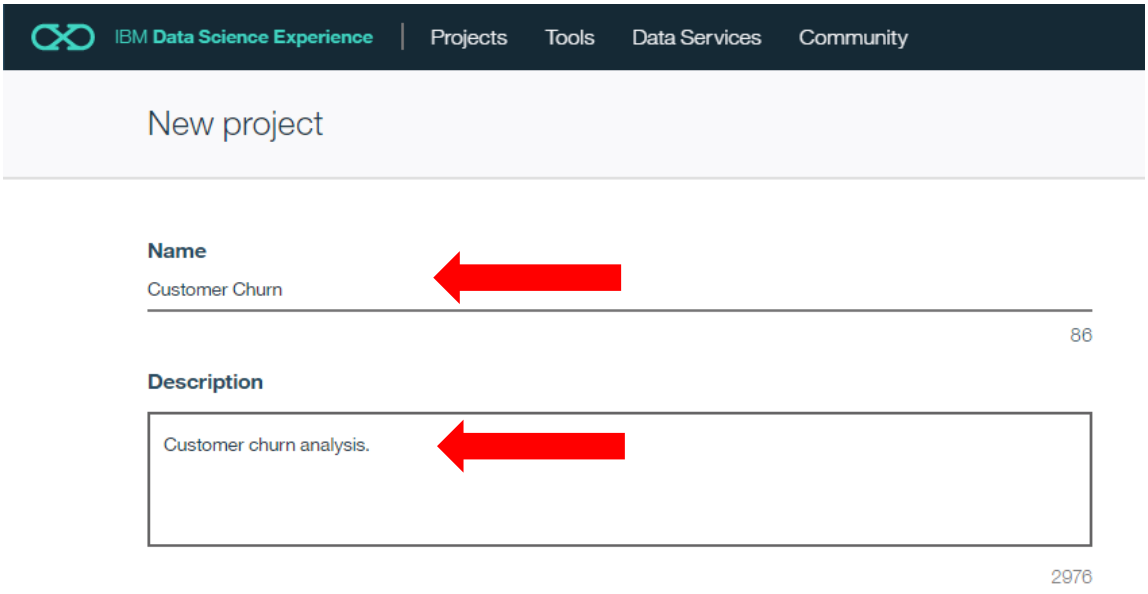
Action

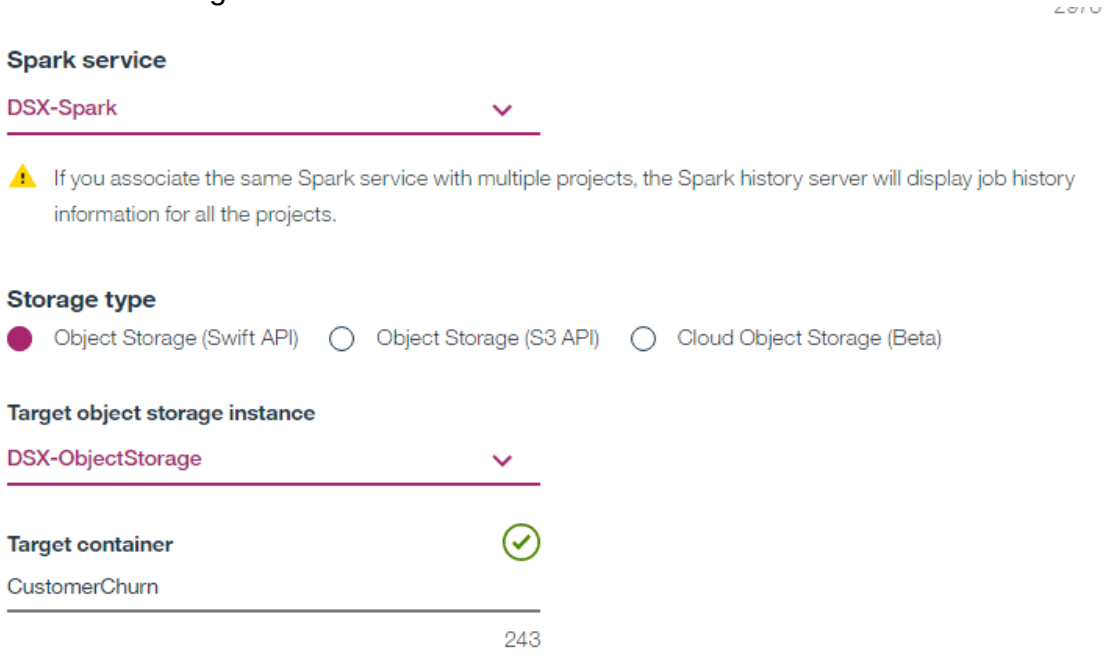
2. Create New Project

- Navigate to **https://datascience.ibm.com**
- Login to DSX
- On the top right side, click **Create New and select project**

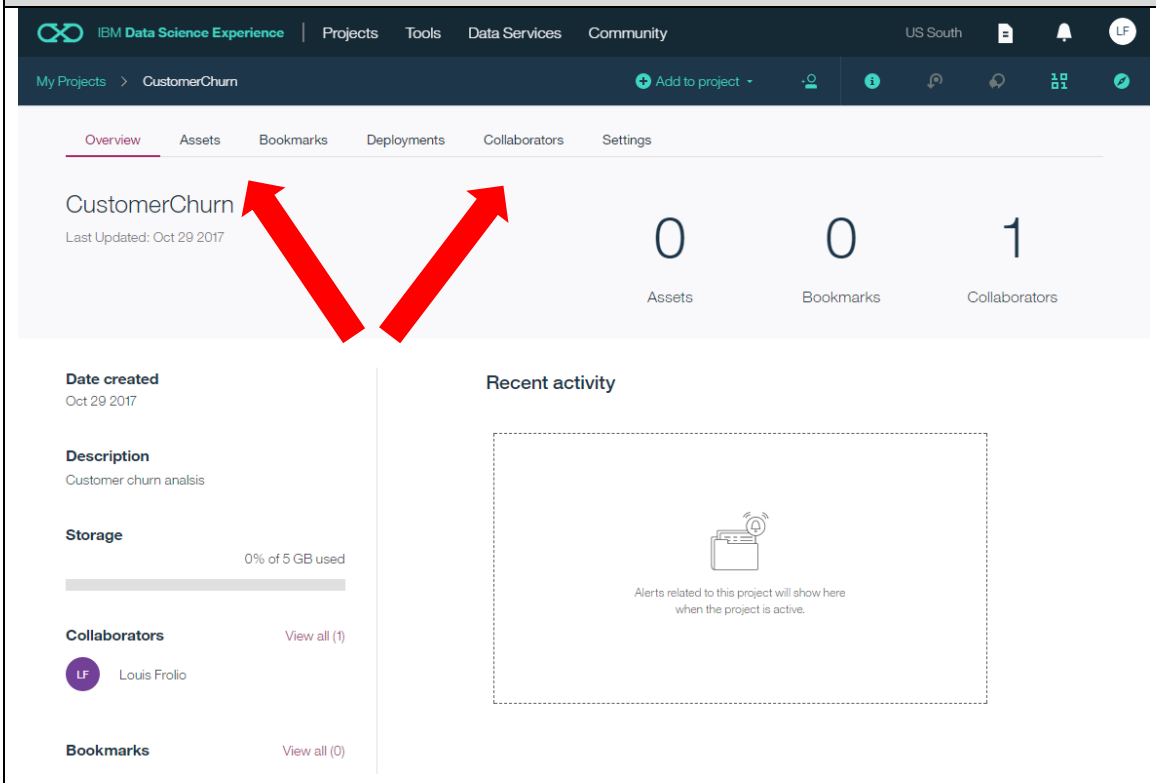


- Type the Project Name **Customer Churn,** add a meaningful description:



Action
Ensure the defaults are selected as follows: • Select your Spark Service **DSX-Spark** • Select **Object Storage (Swift API)** • Select Target Object Storage Instance **DSX-ObjectStorage** • Default Target Container **CustomerChurn**  • Click **Create**

Action

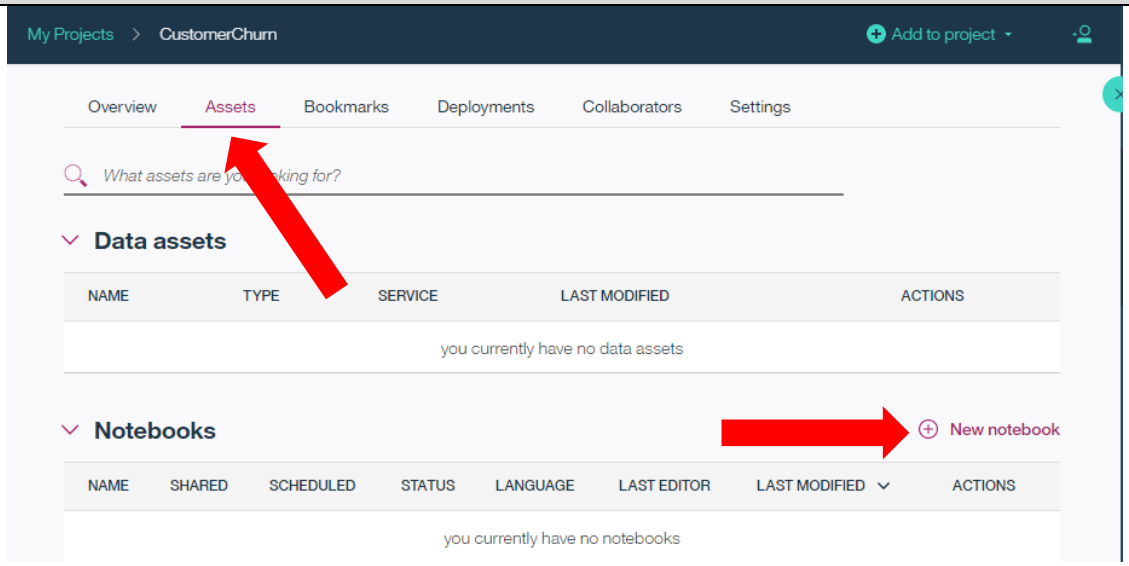


You now have a **Project** that is empty. You can use the tabs along the top to **add assets** to your project such as Connections, Notebooks, Data Assets, etc. You can also **add collaborators** to the Project.

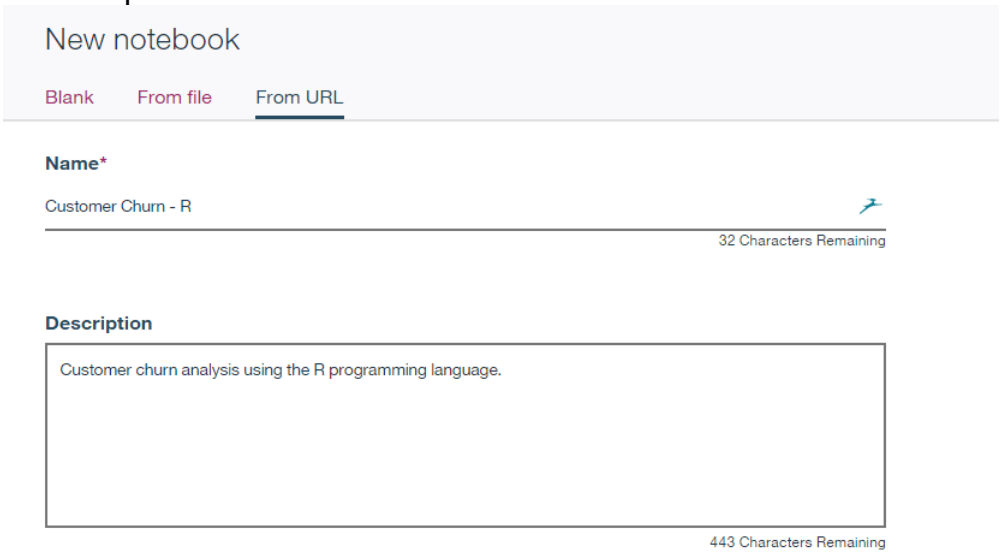
3. Create Notebook

- Click **Assets,** then **Add Notebooks**

Action



- Choose **From URL** from the tab, give the notebook a name and meaningful description:

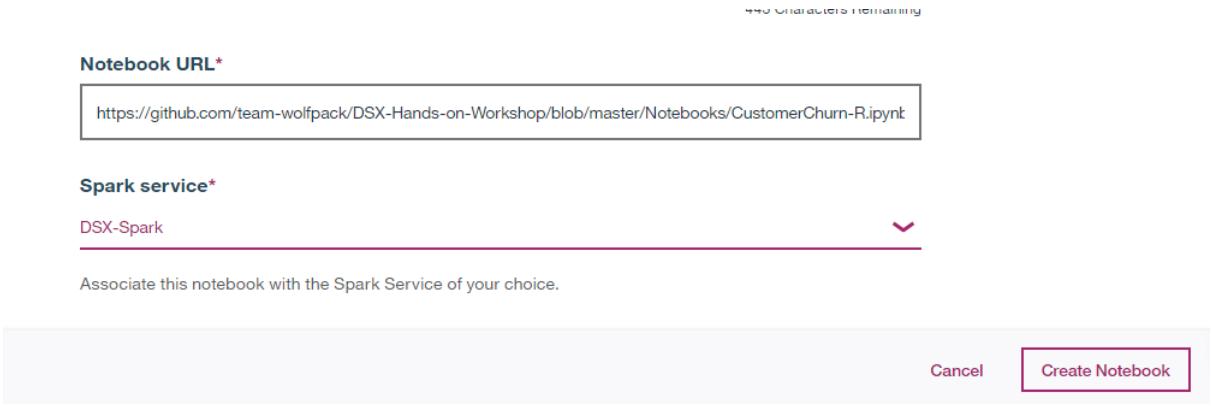


- In a separate browser window navigate to: https://github.com/team-wolfpack/DSX-Hands-on-Workshop

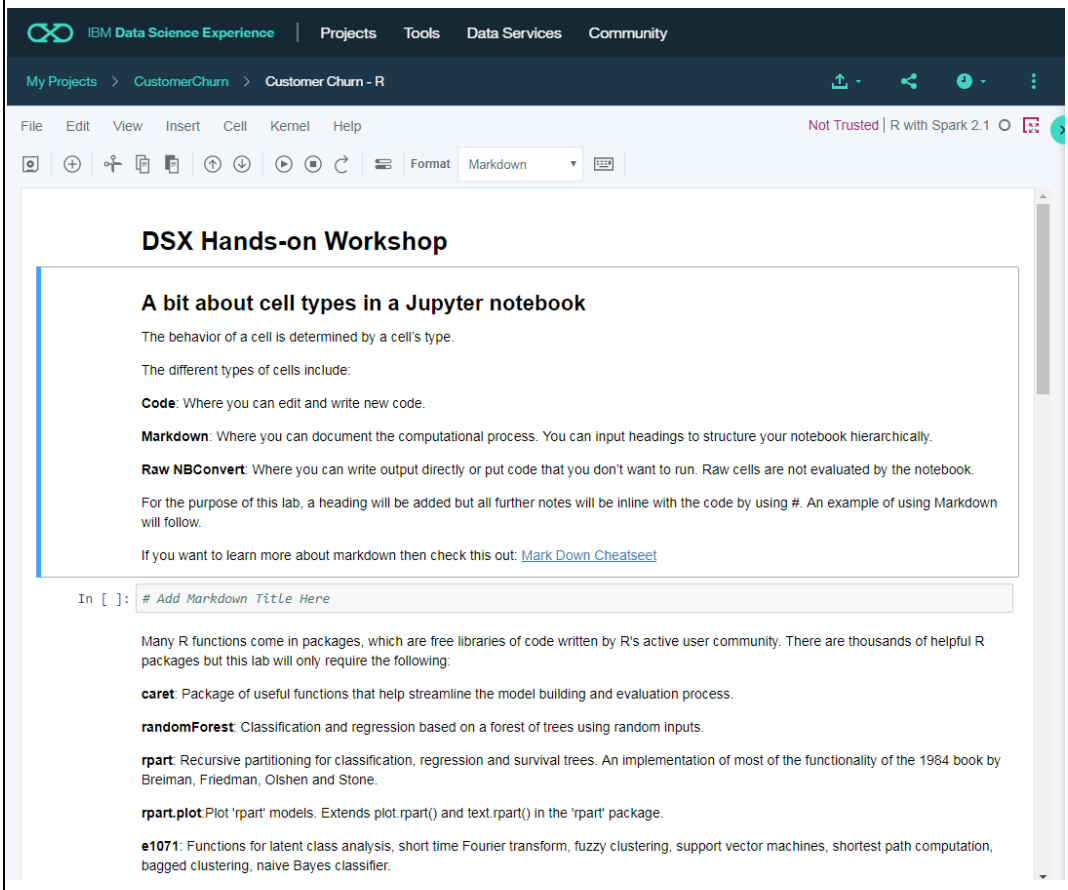
- Click on Notebooks, right click on **CustomerChurn-R.ipynb** then choose **Copy link address**. Go back to the **DSX New Notebook** page.

Action

Paste URL into **Notebook URL** text box then choose **Create Notebook**:



You should now see something like this:



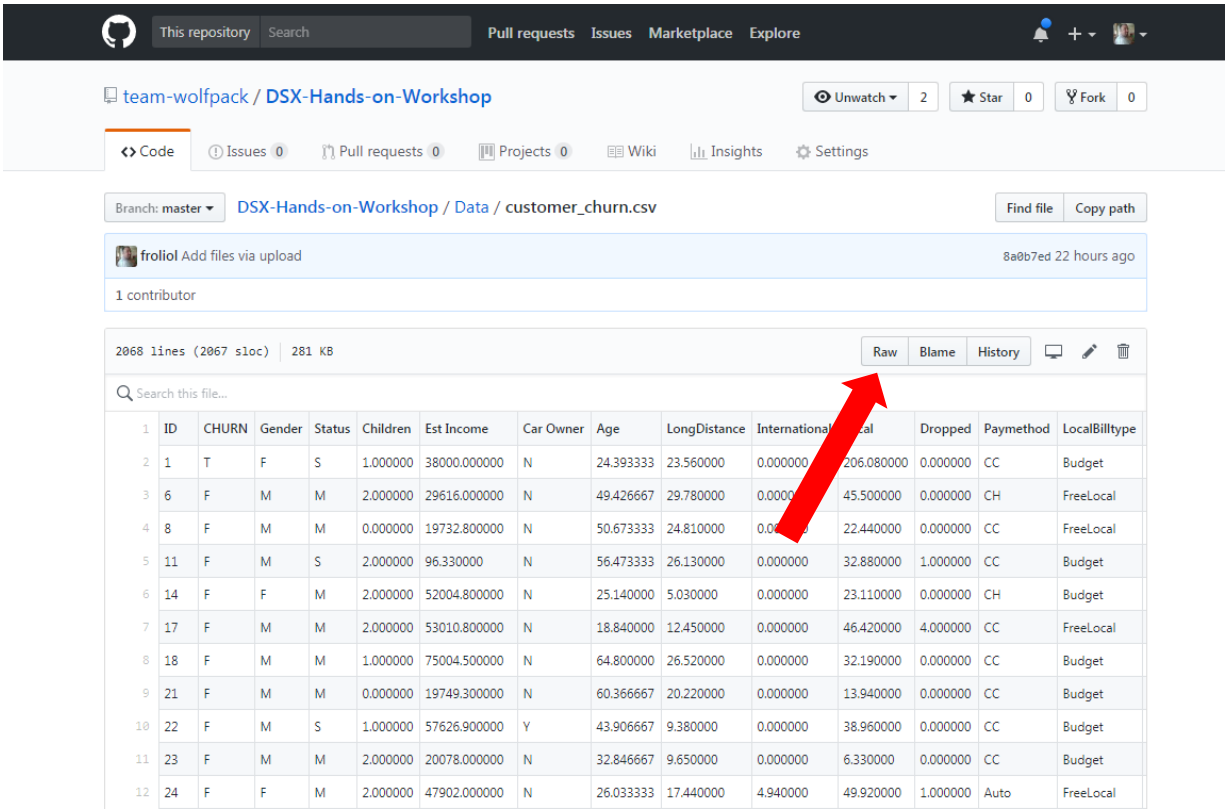
Action

4. Load Data

- In a separate browser, navigate once again to Github:

  https://github.com/team-wolfpack/DSX-Hands-on-Workshop

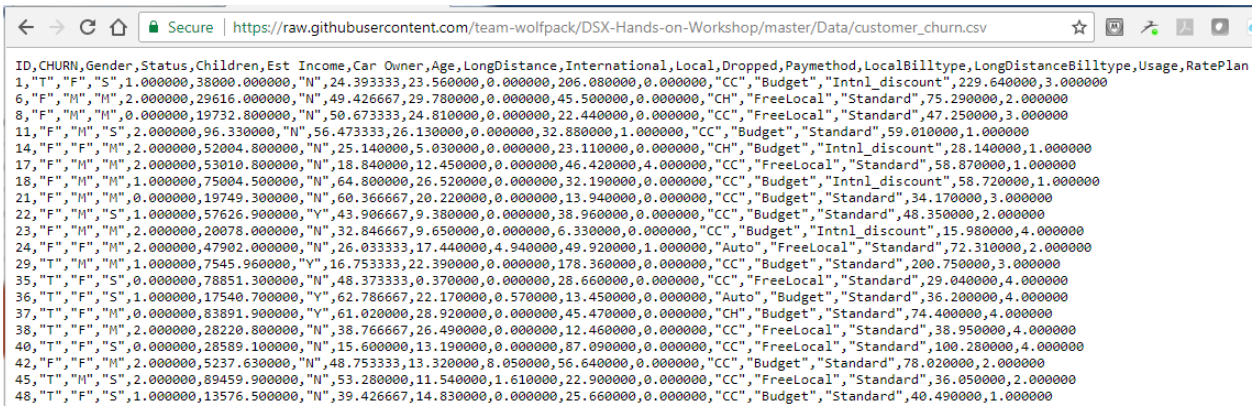
- Click on customer_churn.csv, you should see a tabular list of data. Click on the Raw button:



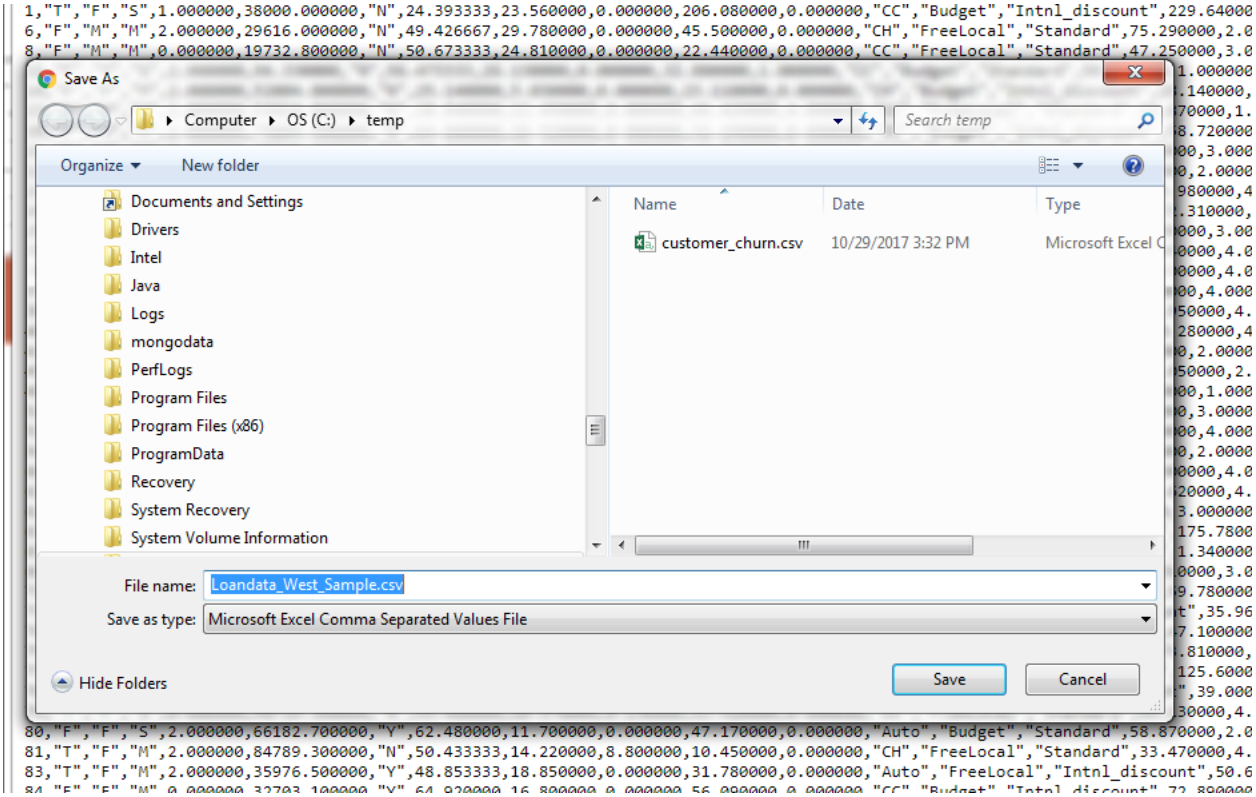
	ID	CHURN	Gender	Status	Children	Est Income	Car Owner	Age	LongDistance	International	cal	Dropped	Paymethod	LocalBilltype
2	1	T	F	S	1.000000	38000.000000	N	24.393333	23.560000	0.000000	206.080000	0.000000	CC	Budget
3	6	F	M	M	2.000000	29616.000000	N	49.426667	29.780000	0.0000	45.500000	0.000000	CH	FreeLocal
4	8	F	M	M	0.000000	19732.800000	N	50.673333	24.810000	0.0	22.440000	0.000000	CC	FreeLocal
5	11	F	M	S	2.000000	96.330000	N	56.473333	26.130000	0.000000	32.880000	1.000000	CC	Budget
6	14	F	F	M	2.000000	52004.800000	N	25.140000	5.030000	0.000000	23.110000	0.000000	CH	Budget
7	17	F	M	M	2.000000	53010.800000	N	18.840000	12.450000	0.000000	46.420000	4.000000	CC	FreeLocal
8	18	F	M	M	1.000000	75004.500000	N	64.800000	26.520000	0.000000	32.190000	0.000000	CC	Budget
9	21	F	M	M	0.000000	19749.300000	N	60.366667	20.220000	0.000000	13.940000	0.000000	CC	Budget
10	22	F	M	S	1.000000	57626.900000	Y	43.906667	9.380000	0.000000	38.960000	0.000000	CC	Budget
11	23	F	M	M	2.000000	20078.000000	N	32.846667	9.650000	0.000000	6.330000	0.000000	CC	Budget
12	24	F	F	M	2.000000	47902.000000	N	26.033333	17.440000	4.940000	49.920000	1.000000	Auto	FreeLocal

Action

You should now see the raw output from the csv file:

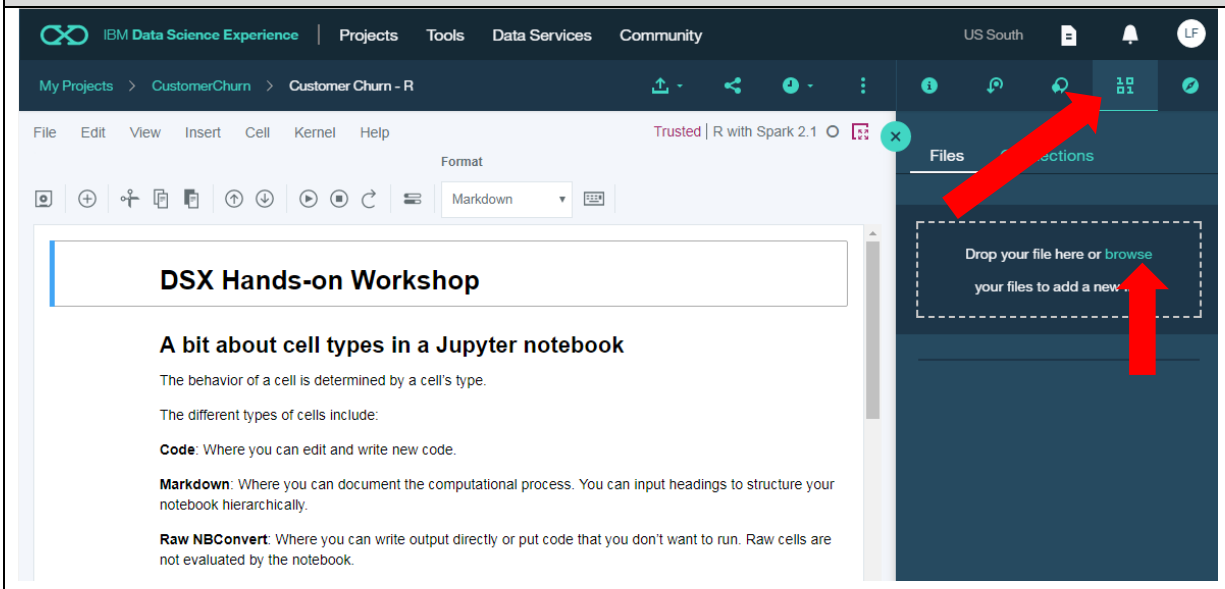


- Right click on the output and choose “Save as” and then save the output to your local file system with the file’s original name “customer_churn.csv.”

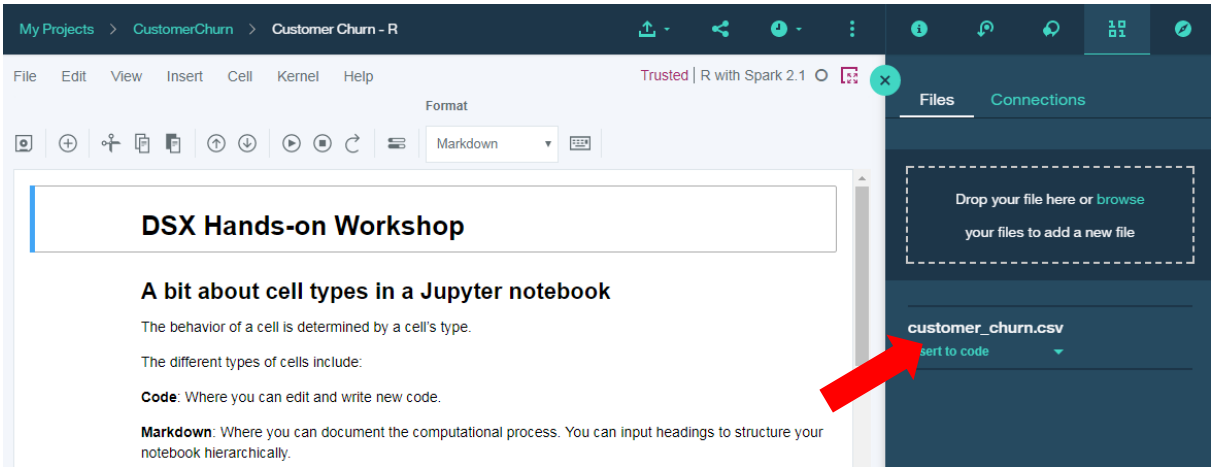


- Go back to the CustomerChurn-R notebook and then click on the Data icon at the top right of the screen:

Action



A new panel will be presented with Files highlighted. Click on browse, navigate to the customer_churn.csv file and select it. You should now see that the file has been imported into the project:



5. Bind Notebook to Data Asset

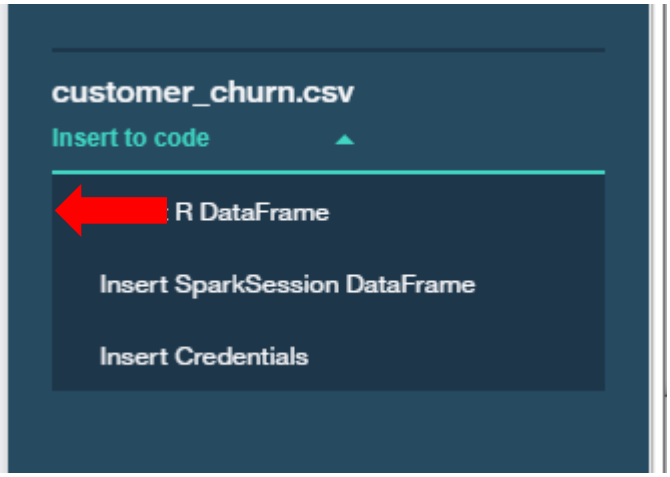
Although the data is part of the project the notebook has no reference to it. Let's now add a reference so that we can analyze the data.

Action

- In the notebook scroll down to the fifth cell:

```
In [ ]: # Placeholder for R Data Frame Auto-code
```

- Place the cursor in the cell and beneath the comment. Navigate the side panel where the data set is displayed, click on “insert into code”:



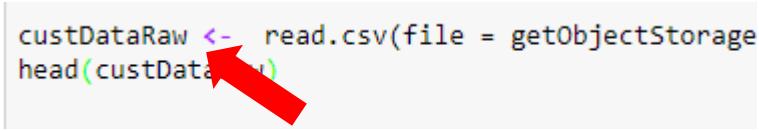
- Choose “Insert R DataFrame” to insert auto-generated code that will allow the notebook to access the data stored on Bluemix:

```
In [ ]: # Placeholder for R Data Frame Auto-code
# @hidden_cell
# This function accesses a file in your Object Storage. The definition contains your credentials.
# You might want to remove those credentials before you share your notebook.
getObjectStorageFileWithCredentials_d5fa59dc7e72461489ab1f8be43ed5a0 <- function(container, filename) {
    # This functions returns a textConnection object for a file
    # from Bluemix Object Storage.

    if(!require(httr)) install.packages('httr')
    if(!require(RCurl)) install.packages('RCurl')
    library(httr, RCurl)
    auth_url <- paste("https://identity.open.softlayer.com",'/v3/auth/tokens', sep= '')
```

Look at the last line of the newly inserted data frame, particularly the name assigned to it:

```
df.data.1 <-  read.csv(file = getObjectStorageFileWithCrede
head(df.data.1)
```

Action
Let's make it more friendly. Change "df.data.1" to "custDataRaw."  **6. Build and Evaluate Customer Churn Model** **Lesson 2 Continued in Jupyter Notebook**

Lesson 3: Machine Learning Flows

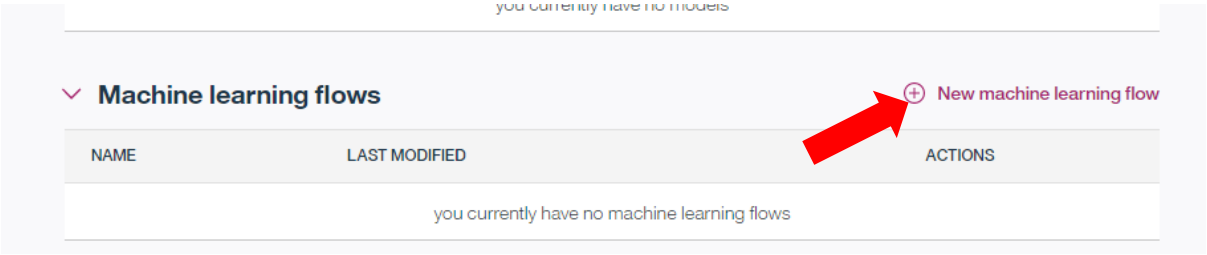
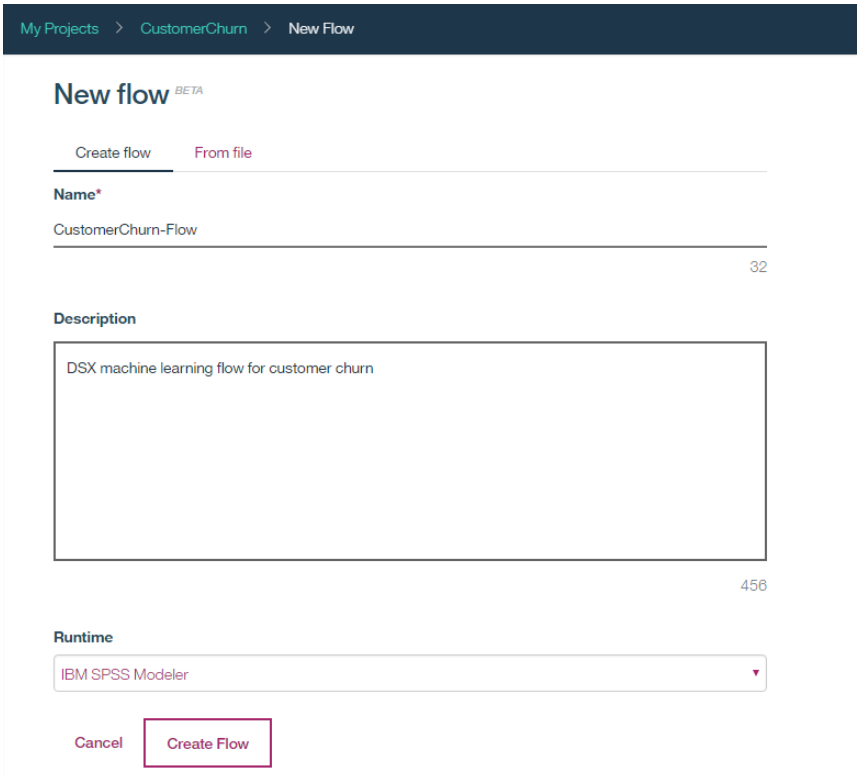
Purpose:	This lesson introduces Machine Learning Flows in DSX. Flows provide a graphical approach to machine learning like that of SPSS Modeler.

Tasks:	Tasks you will complete in this lab exercise include: • Create Machine Learning Flow • Import Data • Leverage Flows' Palette to Orchestrate Customer Churn Machine Learning Pipeline • Evaluate Customer Churn Model

Lesson 3: Workflow Overview

1. Create Machine Learning Flow
2. Add Data Asset
3. Add & Configure Type Object
4. Add & Configure Model Objects
5. Run Flow to Create Nuggets
6. Add & Configure Analysis Object - Measure Performance
7. Add Second Model Technique to Flow

Lesson 3: Instructions

Action
1. Create Machine Learning Flow • Navigate to CustomerChurn project page • Click on "**New machine learning flow**"    • Choose "**Create flow**" on the top menu. Give the flow a meaningful name and description. For "Runtime" choose "**IBM SPSS Modeler**": 

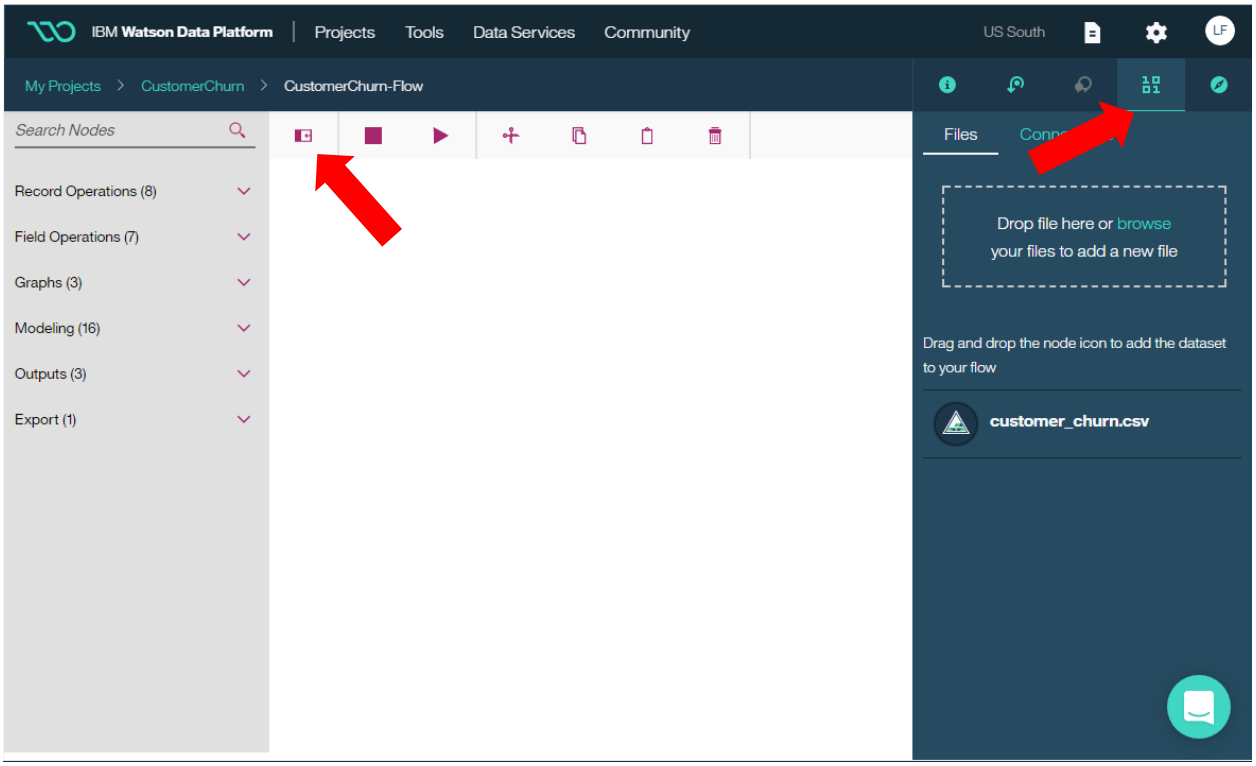
Action

Click on “Create flow.”

2. Add Data Asset

You should now see an empty workspace.

- On the top left click on the “Palette” icon, and on the top right click on the “**Find and Add Data**” icon.



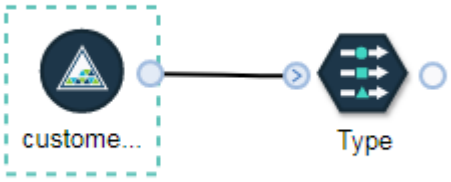
The palette represents the set of tools available for use with DSX flows. The menu of the right should look familiar.

- Let’s start by dragging and dropping the “**customer_churn.csv**” file onto the workspace.

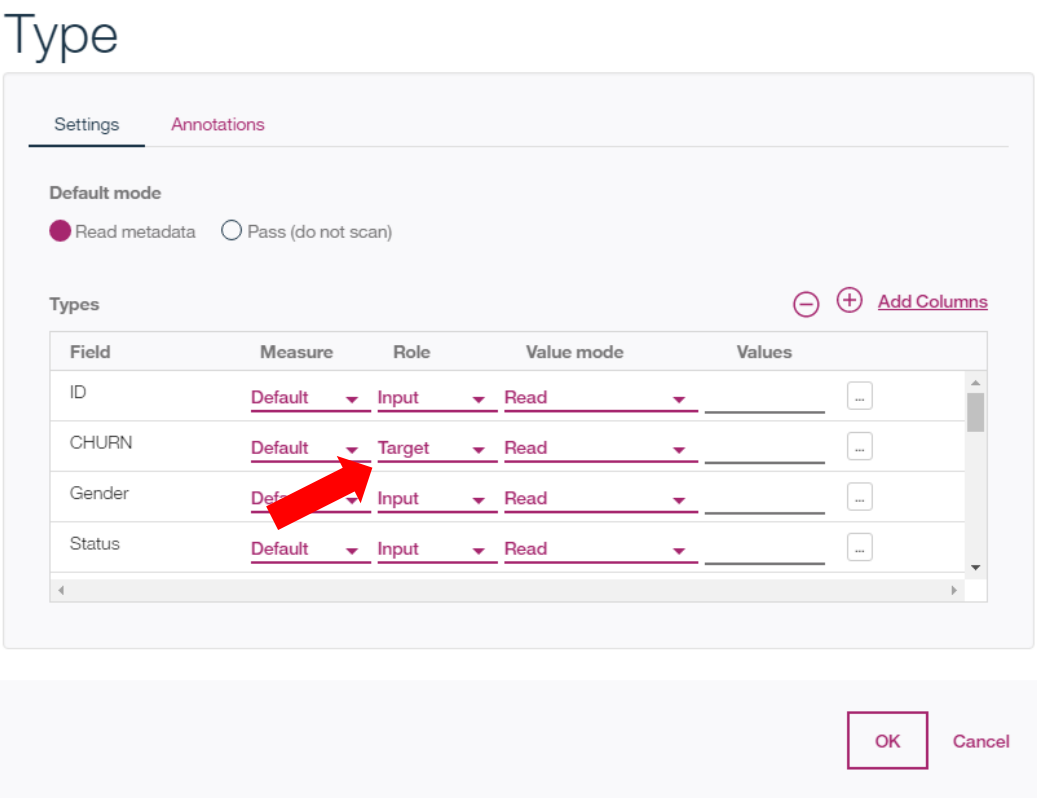
3. Add & Configure Type Object

- From the palette, expand “**Field Operations**”, then drag and drop “**Type**” onto the workspace and to the right of “customer_churn.csv. Connect the two objects:

Action



- Double click on “Type”, click on “**Add Columns**” then add all the columns. Click on the arrow back to the “Settings” page of the “Type” object. For the “CHURN” column, change its Role to that of “**Target.”** Leave the default for the remaining columns:



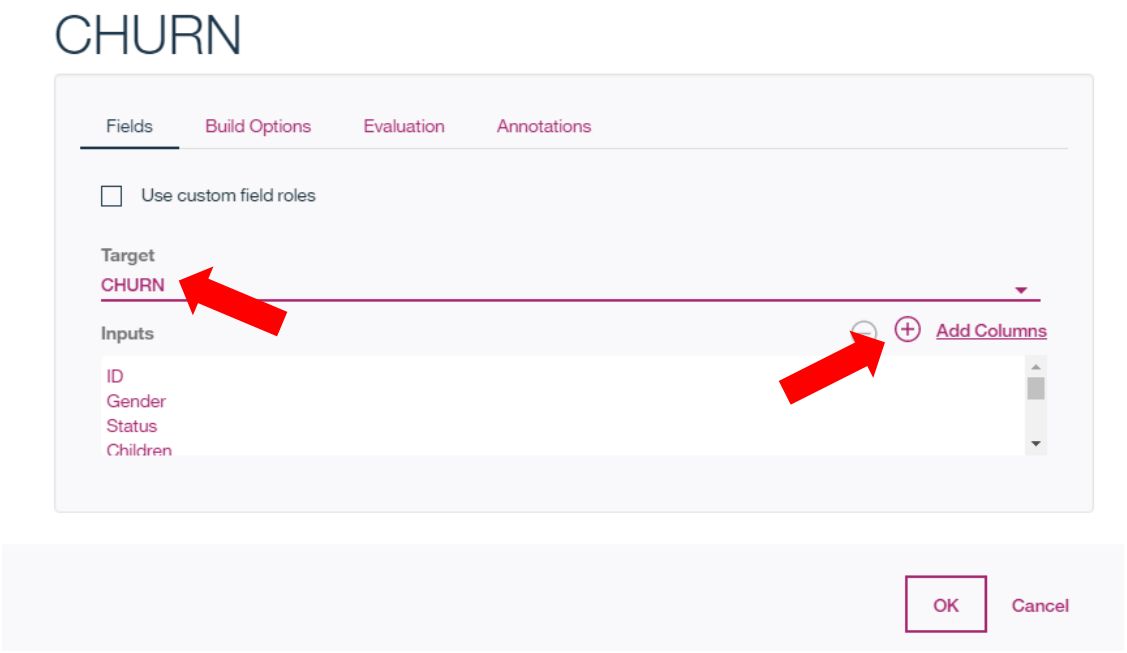
- Click “OK” to exit.

Action

4. Add & Configure Model Object

- From the palette, expand the "**Modeling**" branch then drag "**C&R Tree**" onto the workspace to the right of "Type."
- Connect the two then double click on "C&R Tree" to edit its properties.

The "Target" should identify "CHURN" automatically:



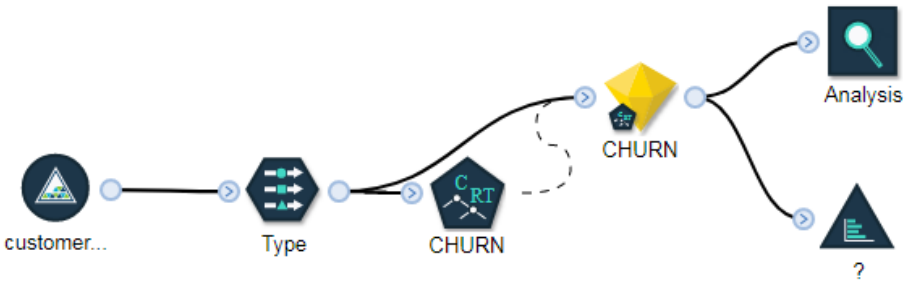
- Click on "Add Columns." Recall from the notebook exercise you were asked to jot down the top 10 fields that were identified as the greatest influencers. Choose those columns as inputs to the decision tree model. Click "OK" to return to the workspace:

5. Run Flow to Create Nugget

- Run the flow by clicking on the “Run” icon at the top of the workspace.

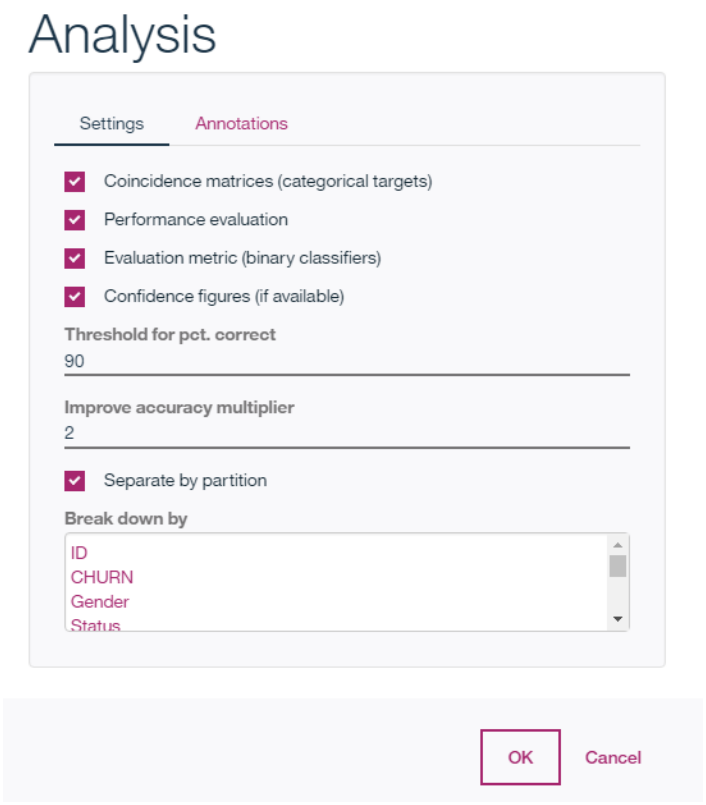
You should see a new forth object on the workspace, this is called a nugget.

- From the palette add an “**Analysis**” object to the workspace, you will find it under the “Outputs” drop down. Also, from the “**Graphs**” drop down add a “**Distribution**” object to the workspace. Connect the nugget to each of them:

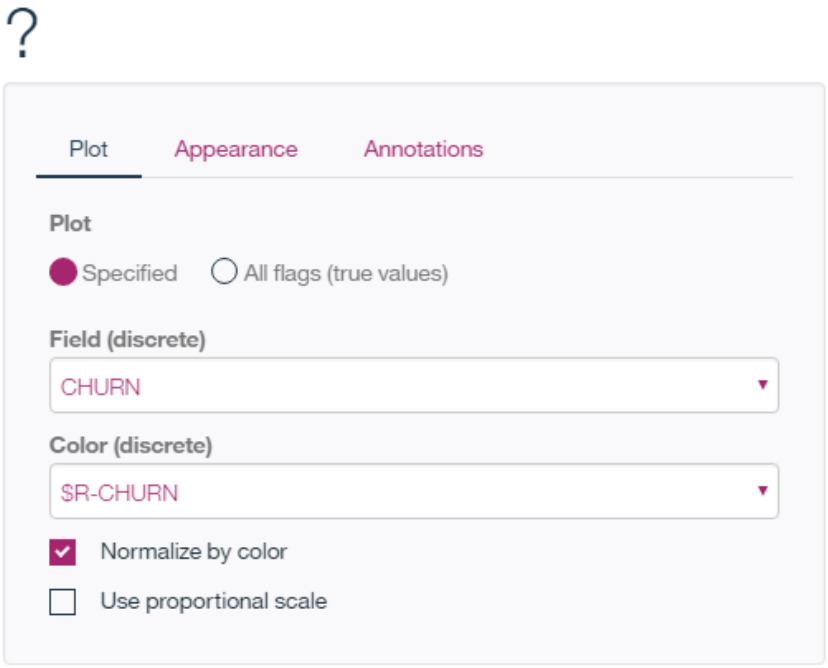


6. Add & Configure Analysis Object – Measure Model Performance

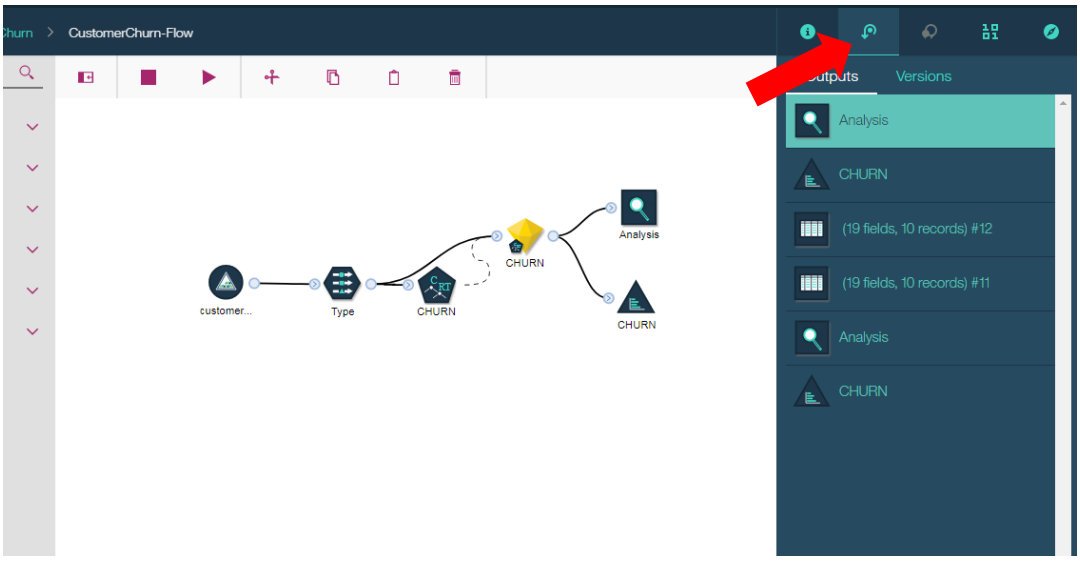
- Double click on “**Analysis**” and check off the four checkboxes, leave the rest as default:



- Click “OK” to return to the workspace.
- Double click on “Analysis” and configure it as depicted below:



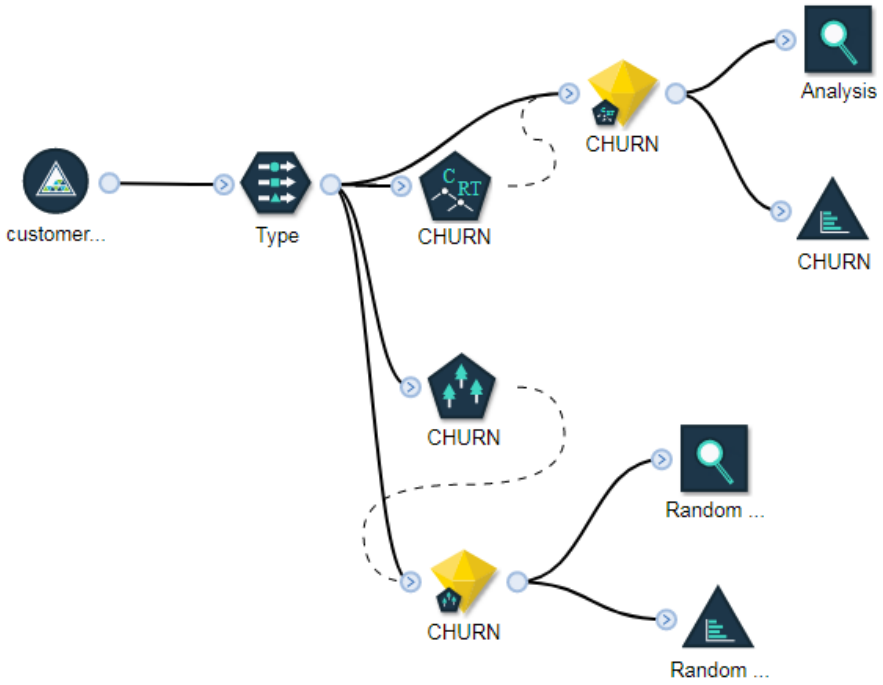
- Click on “OK” to return to the workspace.
- Run the flow again.
- On the right side of the workspace click on the “**Outputs and Versions**” icon to see the resulting analysis:



- Explore the results

7. Add Second Modeling Technique to Flow

- To the palette repeat the process for “**Random Trees**” that you did for “C&R Trees.” Your resulting workspace should look like the following:



- Explore the results.

End of Lesson 3

Lesson 4: Watson Machine Learning

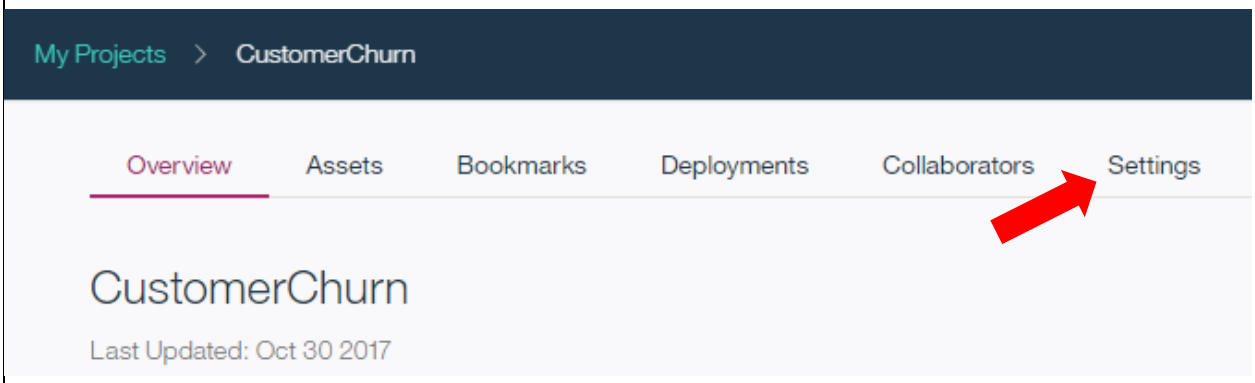
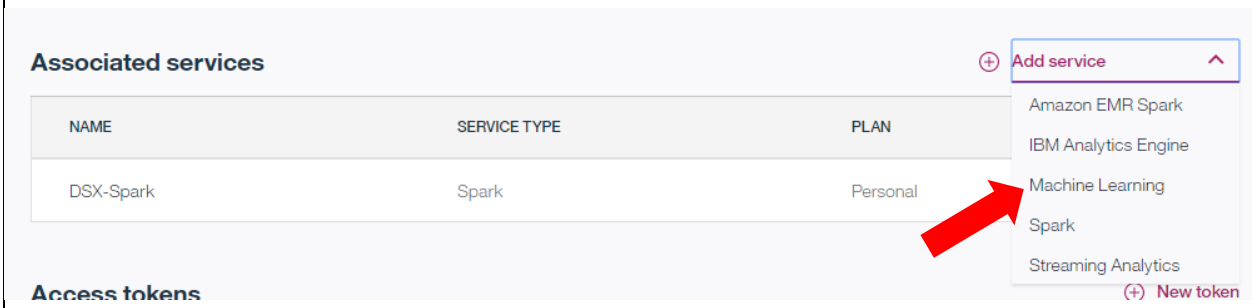
Purpose:	This lab introduces Watson Machine Learning in DSX. Watson Machine Learning makes the task of machine learning easy with as little as a few clicks of the mouse.

Tasks:	Tasks you will complete in this lab exercise include: • Creation of requisite services to support Watson Machine Learning • Creation of Watson Machine Learning Models • Model Performance Evaluation • Deployment and Prediction of Model

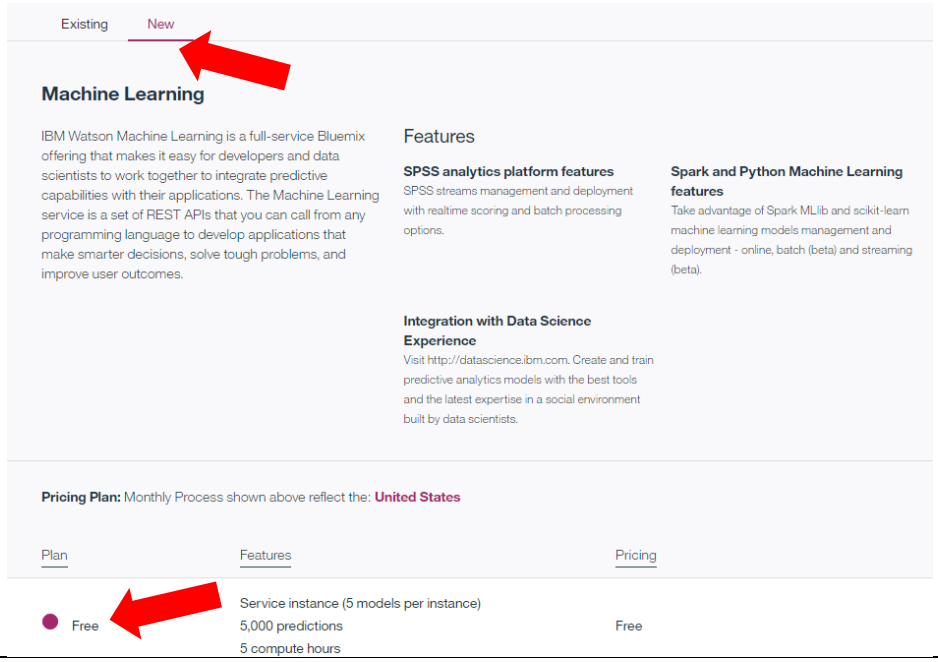
Lesson 4: Workflow Overview

1. Create Machine Learning Service
2. Create Machine Learning Model
3. Choose Modeling Technique
4. Add Estimators
5. Evaluate Models
6. Save & Deploy Model
7. Predict with Model

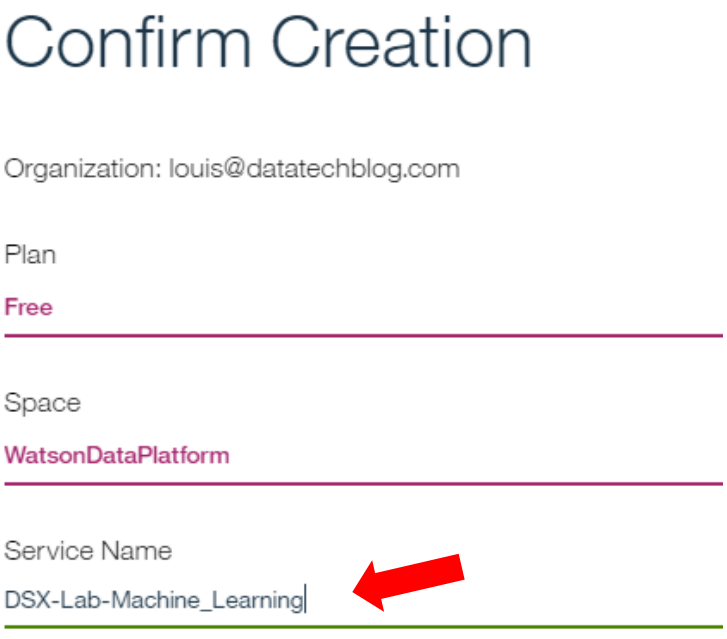
Lesson 4: Instructions

Action
1. Create Machine Learning Service • Navigate to CustomerChurn project page • At the top click on the "**Settings**" icon: 
Scroll to the middle of the page and click on "Add service" then choose "Machine Learning":  • On the Machine Learning page make sure that the tab is set to "New", for the plan choose "Free":

Action



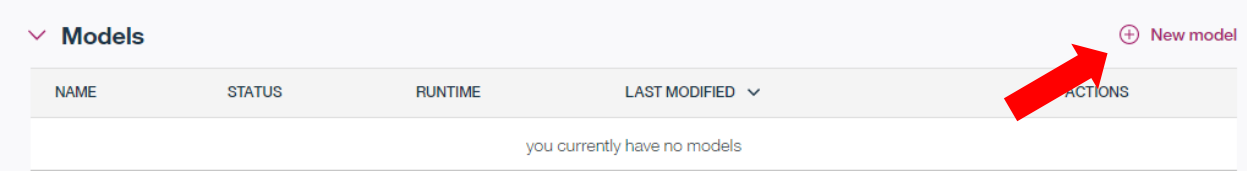
- Click on "**Create IBM Watson Machine Learning**."
- At the confirmation page you can give your service a meaningful name:



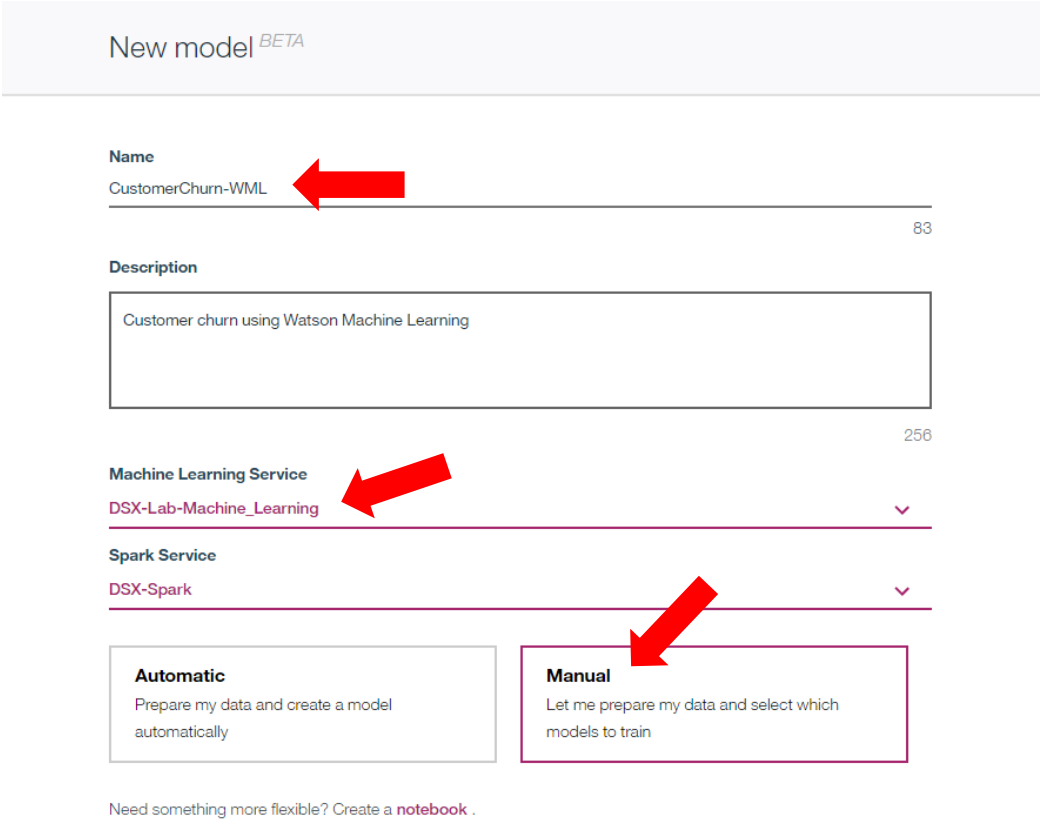
- Click “Confirm” to create Watson Machine Learning Service.

2. Create Machine Learning Model

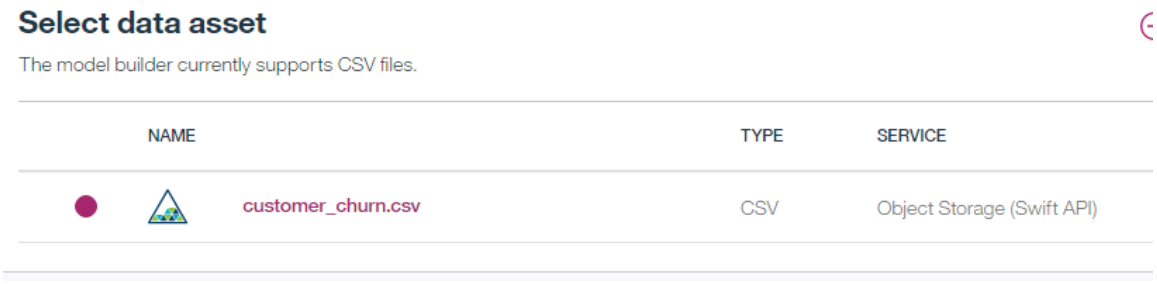
- In the Project click on “**Assets**” at the top of the window.
- In the middle of the page you will see “**Models**”, click on “New model”:



- In the “New model” window give your model a meaningful name and description, you should also see the machine learning service you just created. Click on “**Manual**” then “**Create**”:

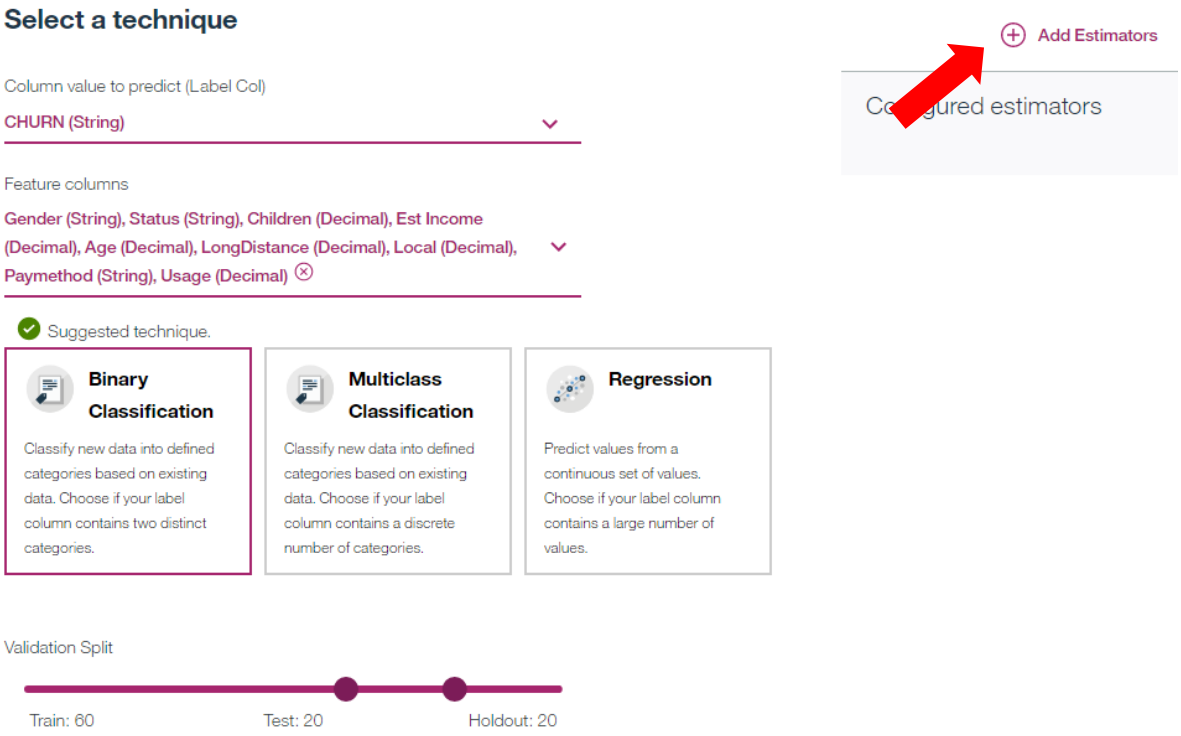


- When complete you will be prompted for a data asset, choose “customer_churn.csv”, then click “Next.”



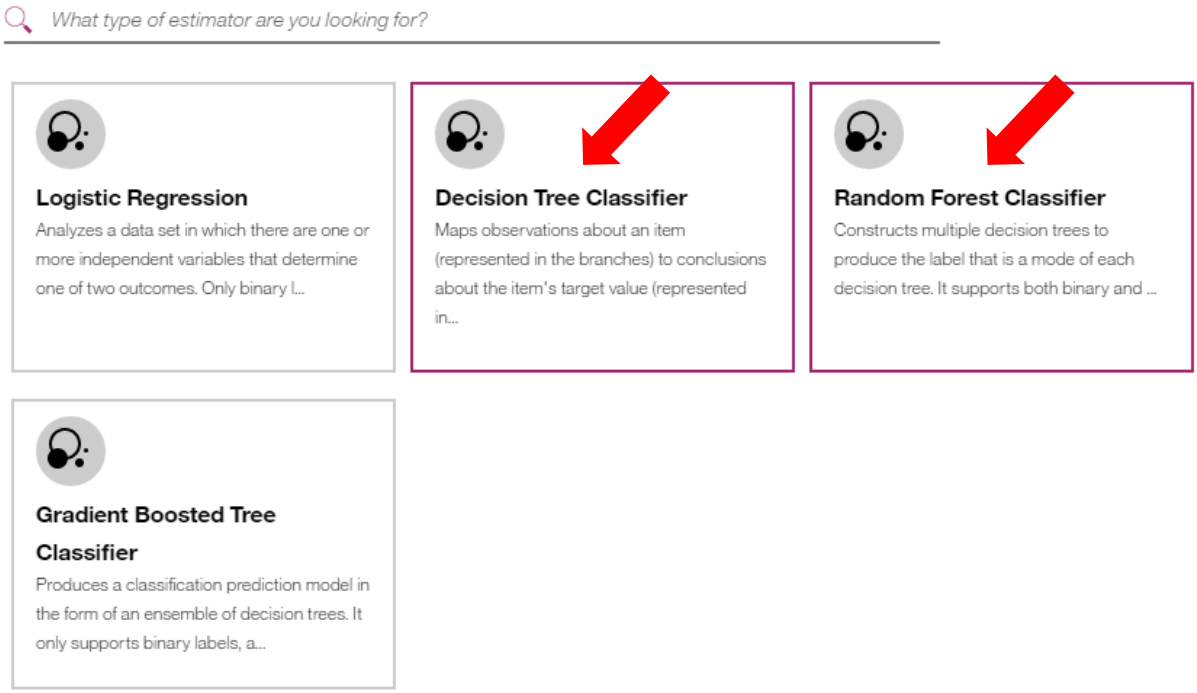
3. Choose Modeling Technique

- At the "**Select a Technique**" screen select "**CHURN**" as the "**Column value to predict**", and for the feature columns choose the top 10 features identified in the Jupyter notebook lab. Also, make sure "**Binary Classification**" is highlighted:

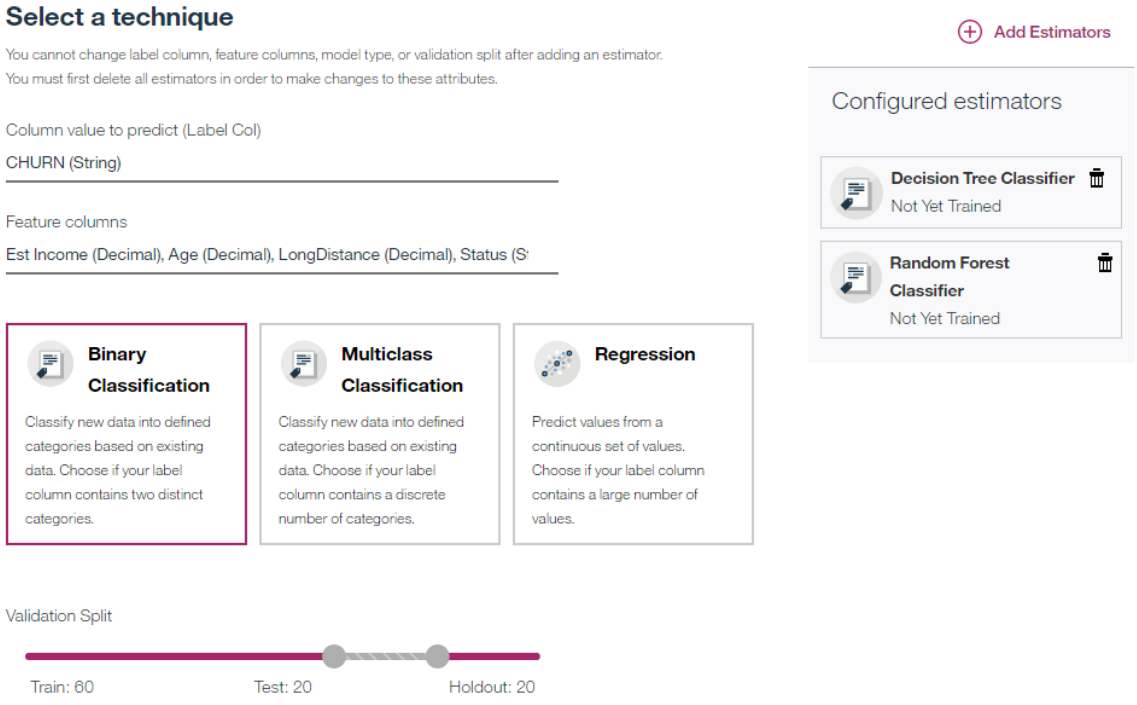


4. Add Estimators

- In the upper right-hand corner of the screen you will see "**Add Estimators**", click on the icon. In the "Select estimator(s)" screen choose **Decision Tree Classifier**, and **Random Forest Classifier**:

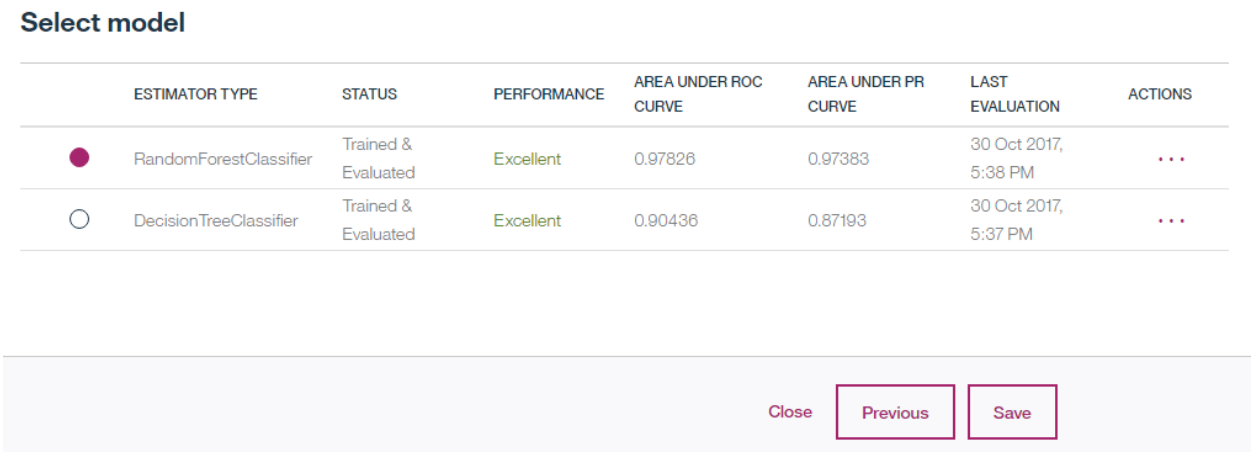


- Click “Add”:



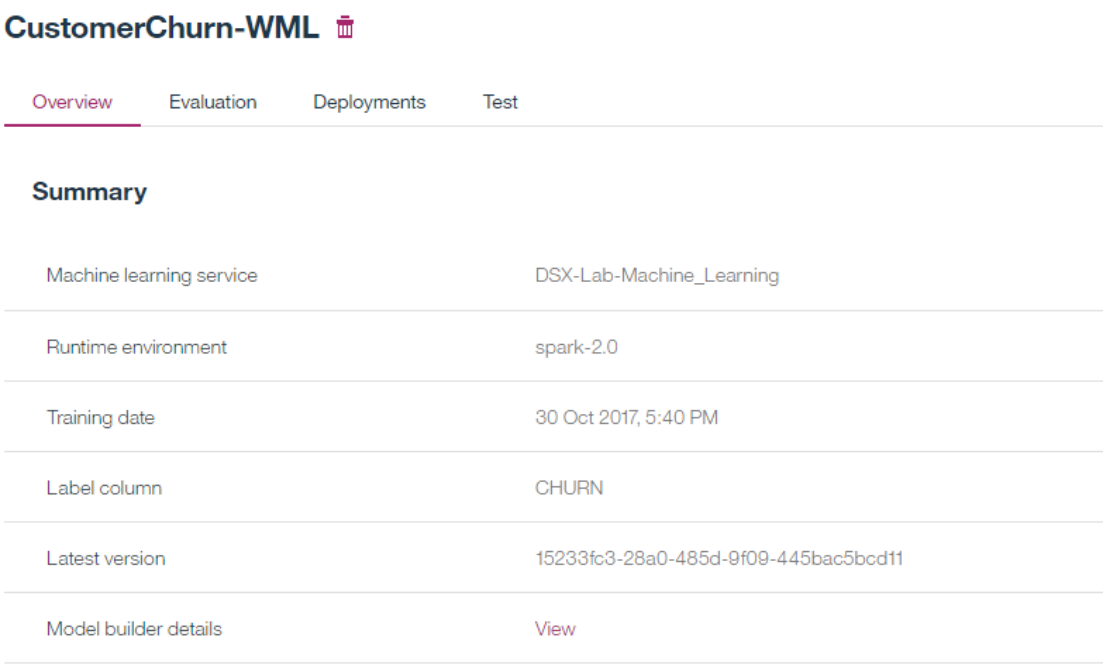
- Click “Next” to train models. This will take 1-2 minutes with the data set we are using:

5. Evaluate Models



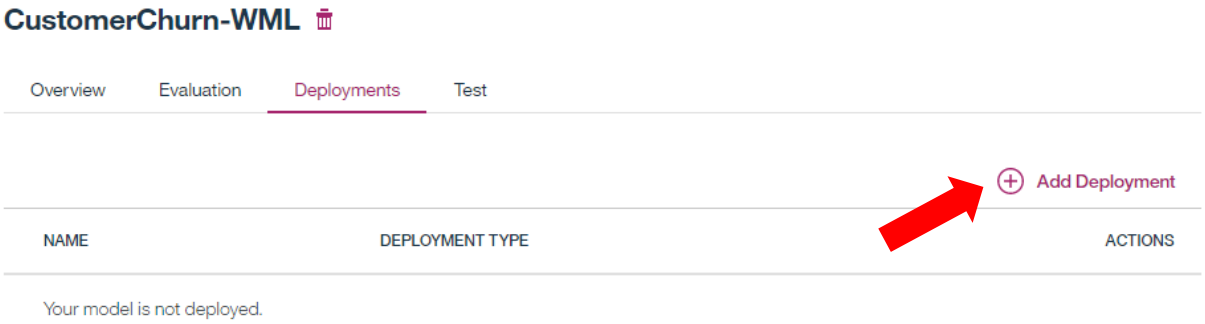
6. Save & Deploy Model

- Pick which model you want to keep then click “Save:”

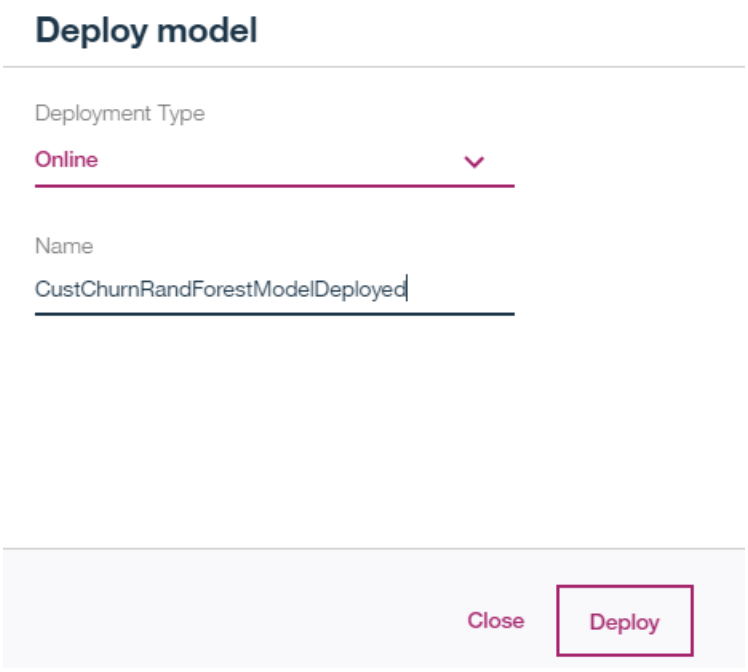


The overview page provides useful information about the model. This includes the ability to deploy and predict with the model.

- Click on “Deployments” then “**Add Deployment**”:



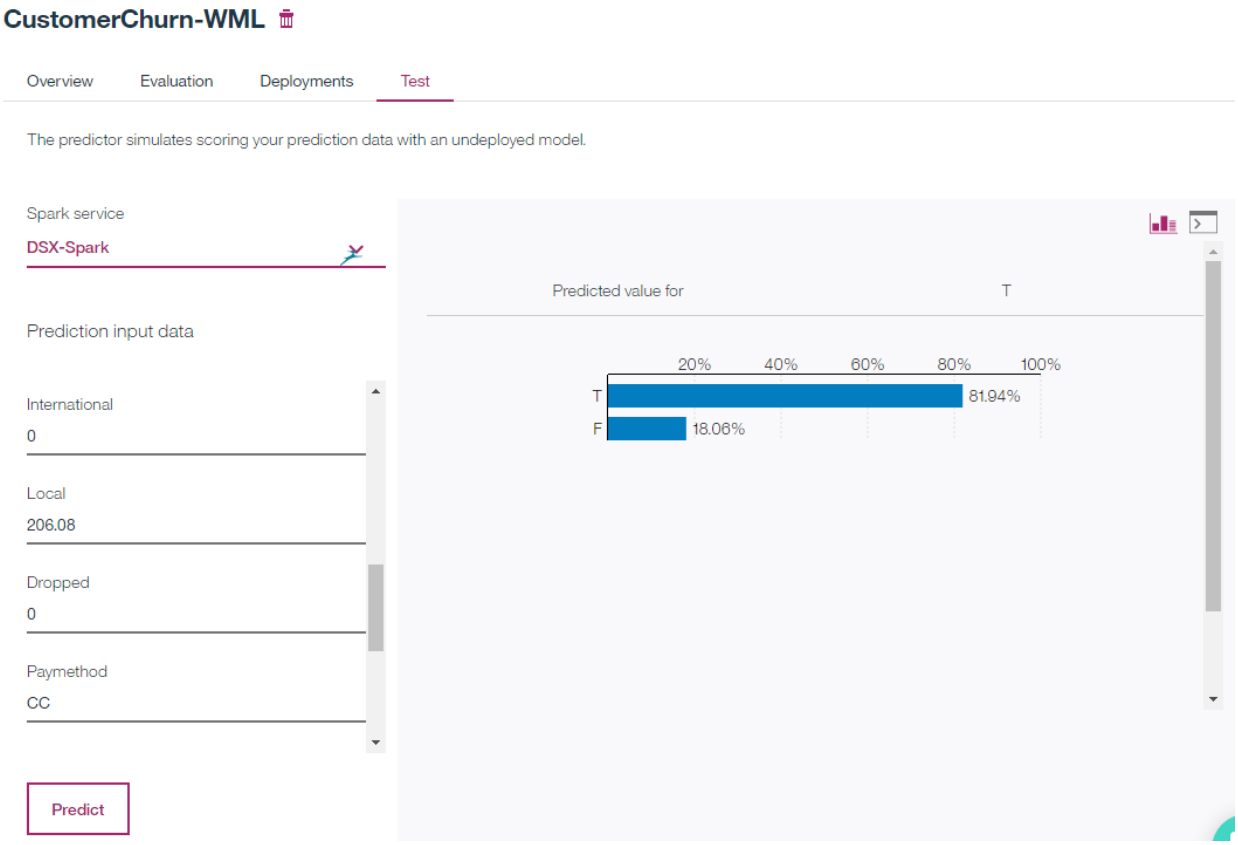
- For “**Select Deployment Type**” choose “**Online**” then give the deployment a useful name:



7. Predict with Model

- Click on “**Test**” to test the model.

The input features will be pre-populated, but you can change them to see different outcomes. Just be sure that the values you add are valid as per the data set. **See “Summary Statistics**” from the Jupyter notebook exercise:



End of Lesson 4

End of Hands-on Workshop

Thank You