Watson Studio: Machine Learning with SparkML

Introduction

In this lab, we will explore machine learning using Spark ML. We will exploit Spark ML's high-level APIs built on top of DataFrames to create and tune machine learning pipelines. We will utilize Spark ML's feature transformers to convert, modify and scale the features that will be used to develop the machine learning model. Finally, we will evaluate and cross validate our model to demonstrate the process of determining a best fit model, load the results in the database, and save the model to the model repository.

We are using machine learning to try to predict records that a human has not seen or vetted before. We will use these predictions to sort the highest priority records for a human to look at. We will use as a training set for the algorithm simulated data that has been vetted by an analyst as high, medium or low.

End-to-End Data Science

The general flow of the End to End Data Science PoT will be guided by the activities shown in Figure 1- End to End Flow. This lab spans the Prepare Data, Build Model, and Save and Deploy activities.

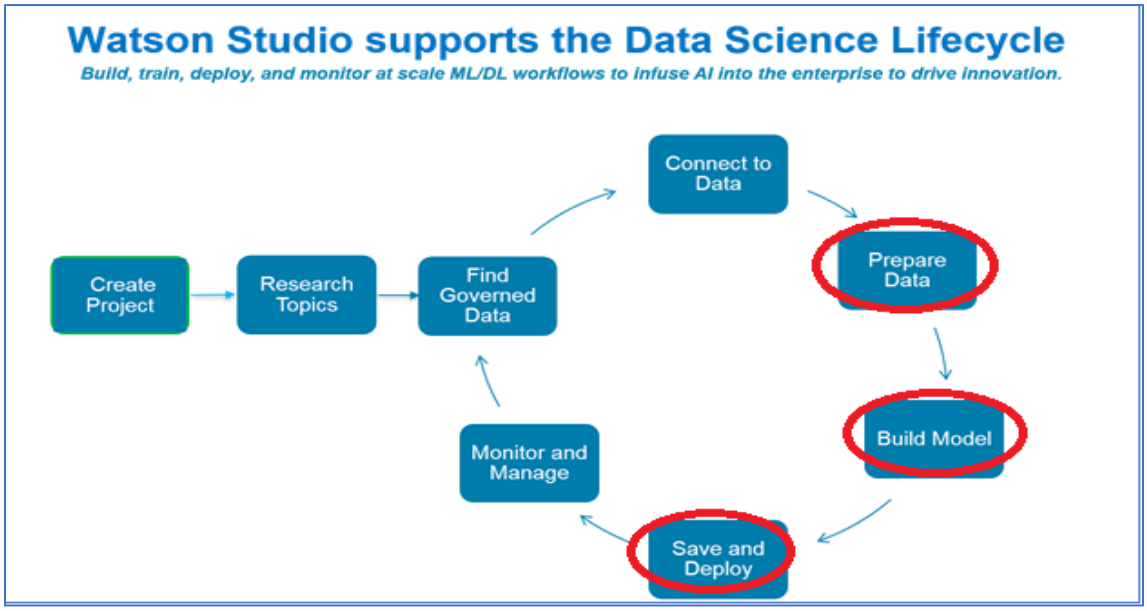


Figure 1- End to End Flow

Objectives

Upon completing the lab, you will know how to:

- Join data from three sources.
- Identify labels and transform data.
- Conduct feature engineering for algorithm data.

- Declare a machine learning model.
- Setup the Pipeline for data transforms and training.
- Train the model.
- Evaluate and show model results
- Automatically tune model
- Score data and load into a new DB2 table.
- Save the model to the model repository.

Female Human Trafficking Data

The data sets used for this lab consist of **simulated** travel itinerary data. The use case corresponds to an analyst reviewing the travel data to assign a risk of trafficking. The risk is recorded as the VETTING _LEVEL column in the dataset. Some of the records have already been analyzed and have a VETTING_LEVEL of low (value is 30), medium (value is 20), or high risk (value is 10). Others have not yet been vetted (value is 100). We will use the data that has been vetted to train a model to predict the risk for the unvetted records. This can be used to automate the process and augment the analyst. For example, one option would be to send the predicted high-risk persons to the analyst for further investigation.

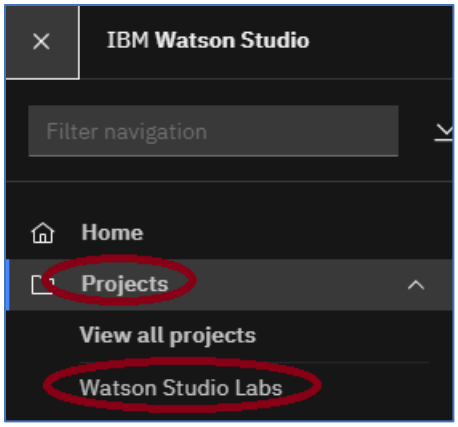
The OCCUPATION data included in the travel data is very granular. For modeling purposes, it was decided to categorize the OCCUPATION data. Two additional datasets are used for this purpose. The occupation.csv dataset maps the granular occupation data to a category code. The categories dataset maps a category code to a category description. These datasets will be joined to the main dataset to prepare the data for modeling.

Other columns in the dataset are similarly very granular and could also be categorized for modeling purposes. This lab does not include steps to accomplish this, but it would be similar to what was done for the occupation column.

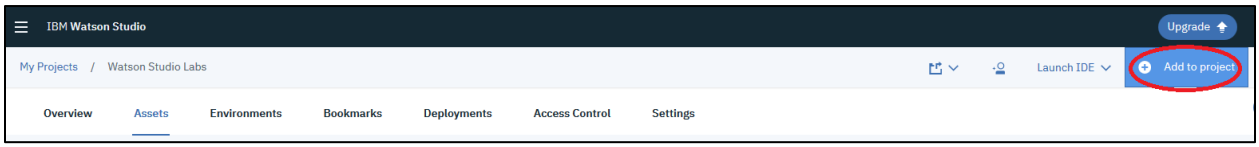
Lab Steps

Step 1 - Create a Jupyter Notebook

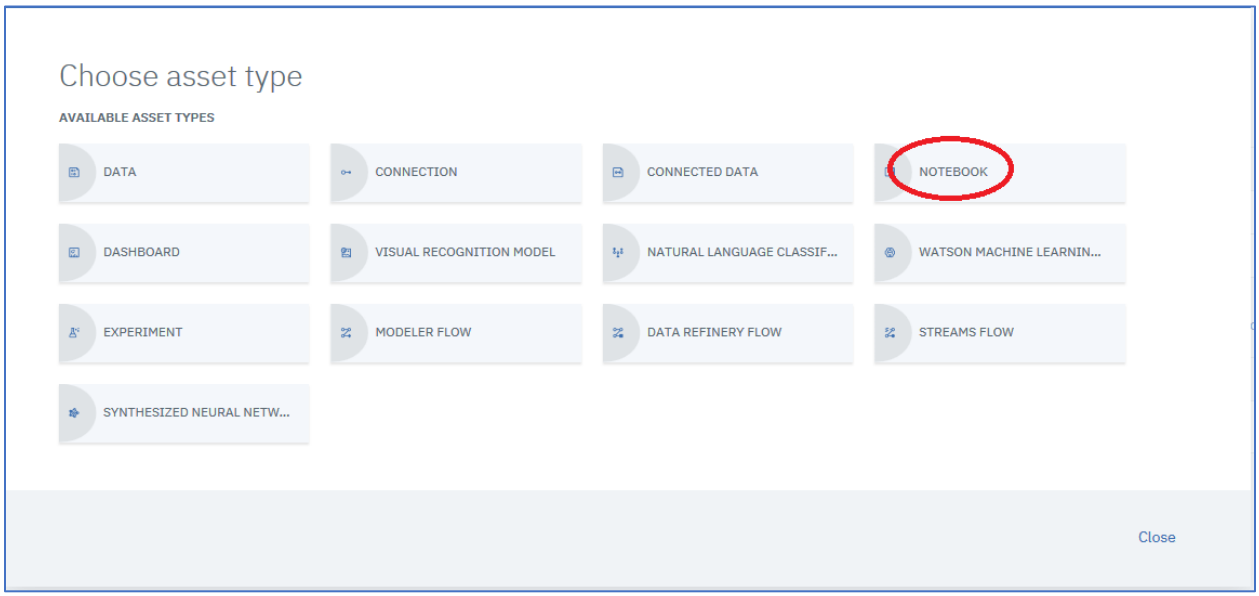
1. Click on the hamburger icon ☰, then click on **Projects**, and then **Watson Studio Labs** (or whatever you named the project)



2. We are now going to create a notebook in our project. This notebook will be created from a url that points to the Machine Learning with SparkML notebook in the github repository. Click the **Add to project** link.



3. Click on **NOTEBOOK**

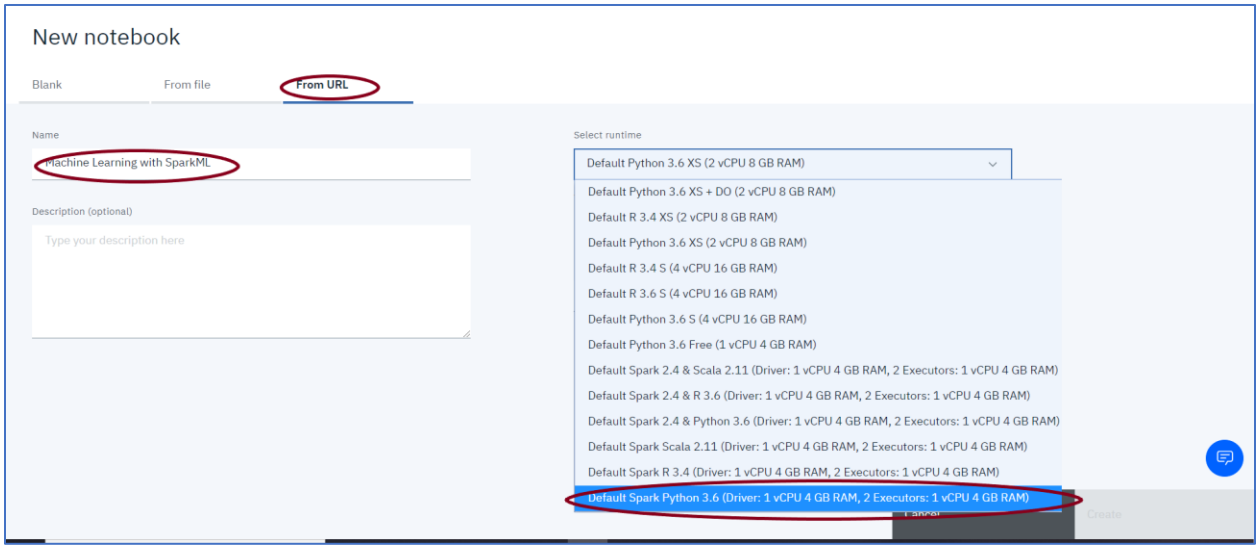


4. Click on **From URL** under New Notebook, enter **Machine Learning with SparkML** for the **Name**, and optionally enter a **Description**.

   Select the Runtime. You will need to select a runtime that includes Python and Spark. MAKE SURE TO SELECT **Default Spark 2.3 &**

Python 3.6 (Driver: 1vCPU 4GB RAM, 2 Executors: 1vCPU 4GB RAM)

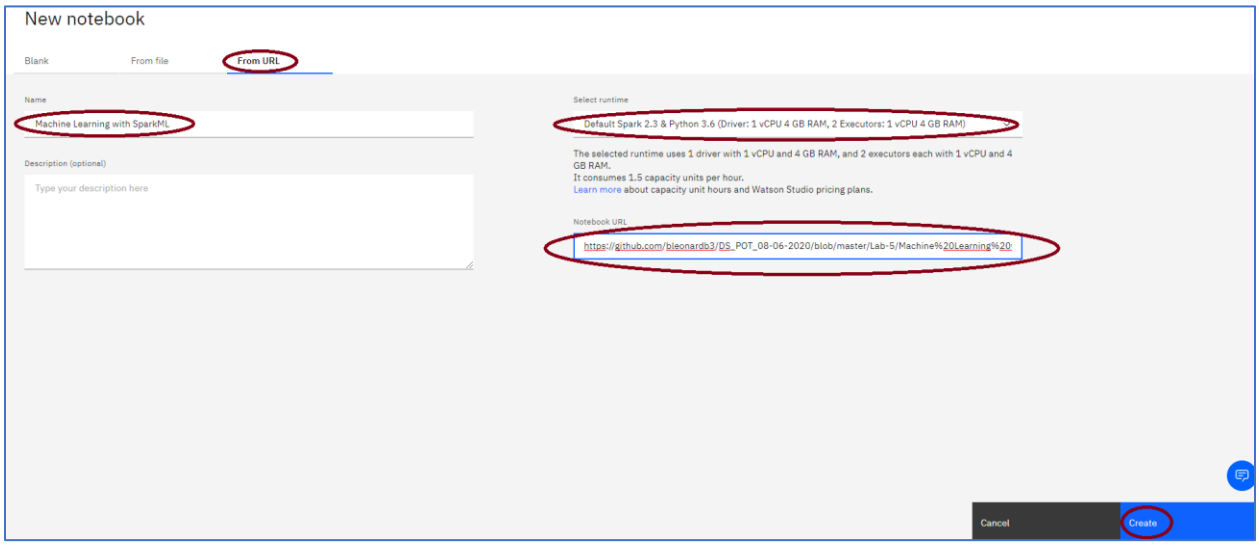
DO NOT select Default Spark 2.4 Python 3.6.



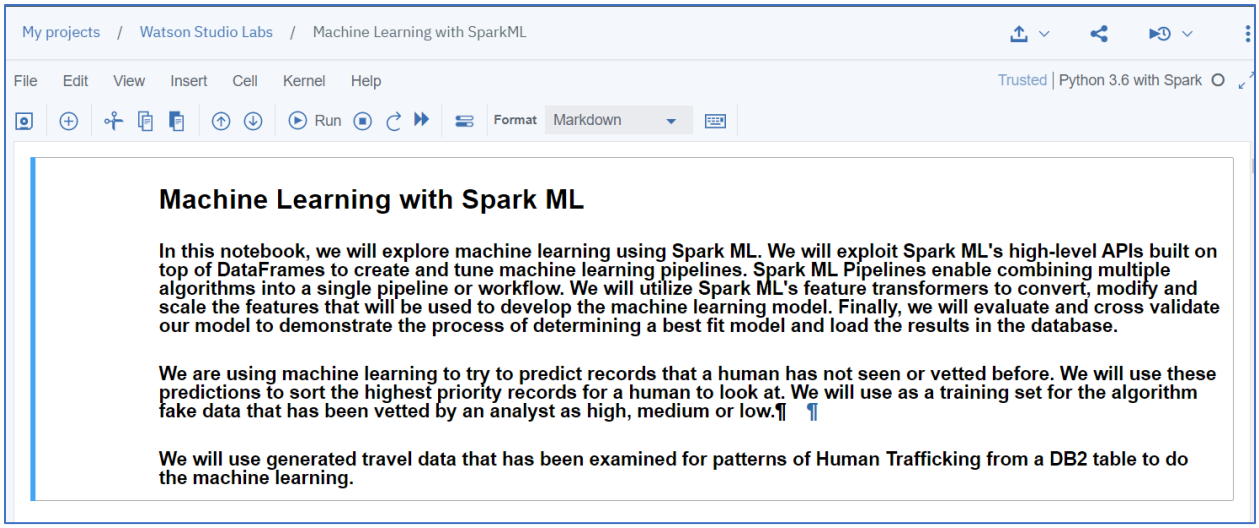
5. Cut and paste the following url into the **Notebook URL** field

https://github.com/bleonardb3/DS_POT_08-06-2020/blob/master/Lab-9/Machine%20Learning%20with%20SparkML.ipynb

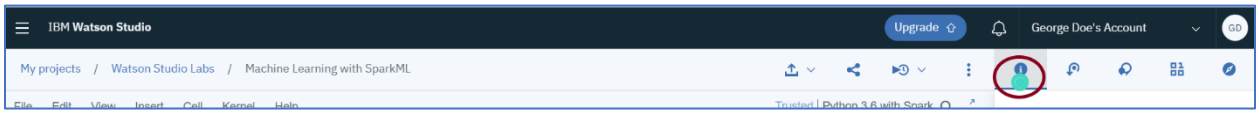
Click **Create Notebook.**



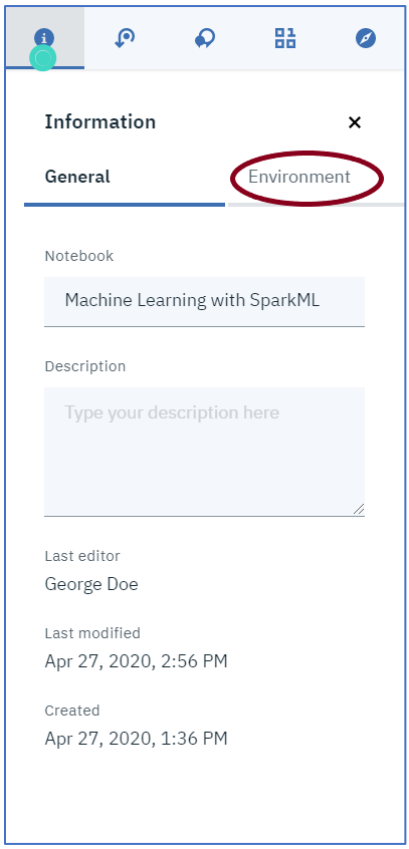
6. You should see the following notebook display.



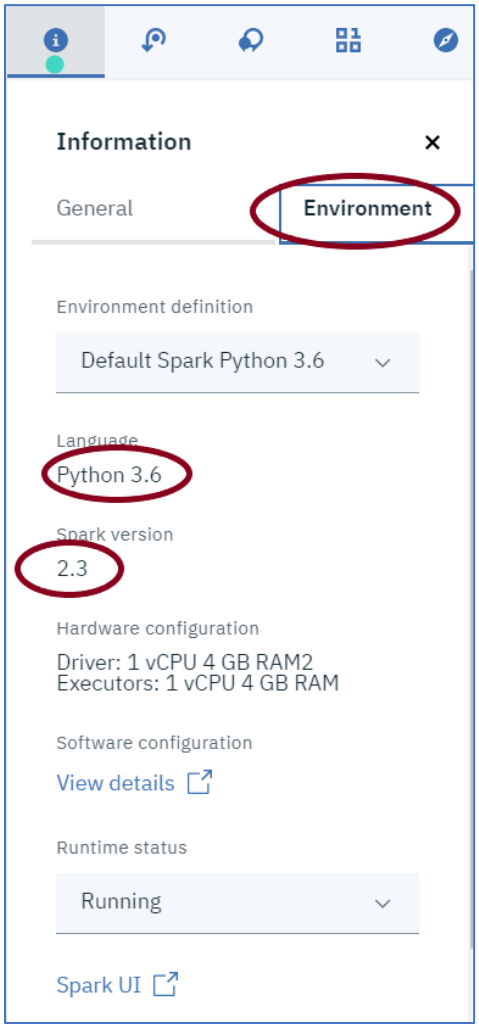
7. To verify the environment is correct, click on the notebook info icon 



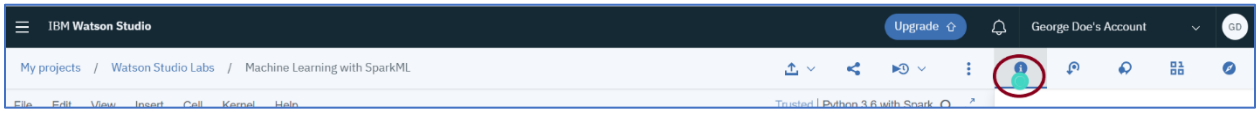
8. Click on **Environment**.



9. Verify that the language is Python 3.6 and the Spark version is 2.3.



10. Click on the notebook info icon  to remove the Info panel.

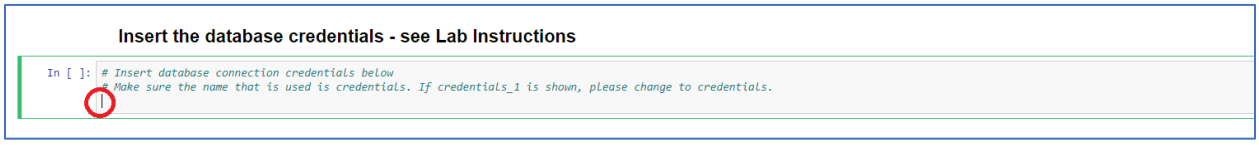


11. A Jupyter notebook consists of a series of cells. These cells are of 2 types (1) documentation cells containing markdown, and (2) code cells (denoted by a bracket on the left of the cell) where you write Python code, R, or Scala code depending on the type of notebook. Code cells can be run by putting the cursor in the code cell and pressing **<Shift><Enter>** on the keyboard. Alternatively, you can execute the cells by clicking on **Run icon** on the menu bar that will run the current cell (where the cursor is located) and then select the cell below. In this way, repeatedly clicking on **Run** executes all the cells in the notebook. When a code cell is executed the brackets on the left change to an

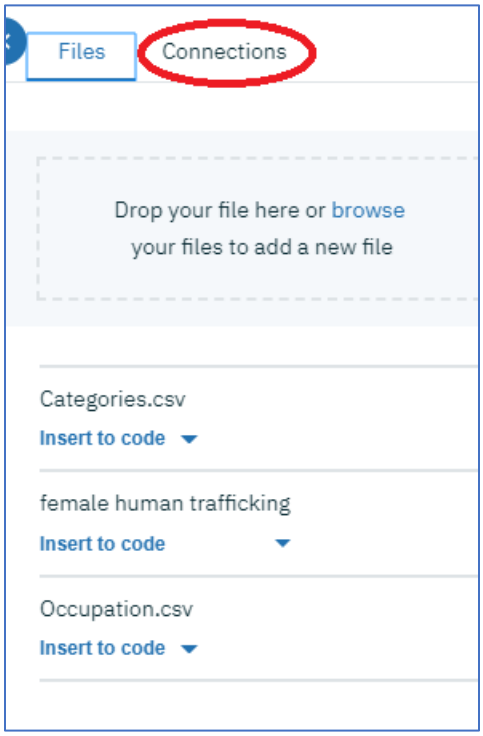
asterisk '*' to indicate the code cell is executing. When completed, a sequence number appears.

Step 2: Insert DB2 Warehouse Credentials

1. Scroll down further towards the middle of the notebook until you see **Insert the database credentials – see Lab Instructions**. Click cursor underneath the commented lines in the code cell.

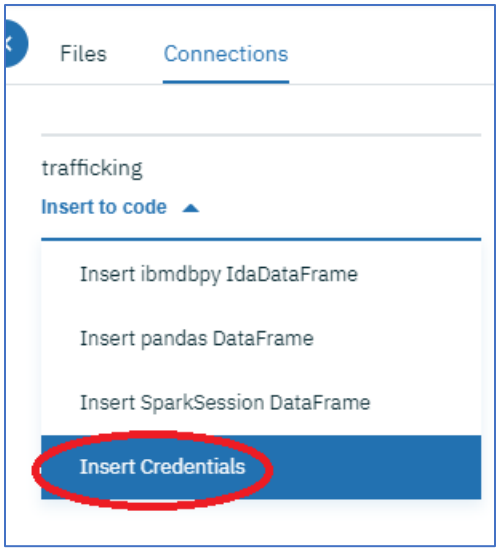


2. Click on **Connections**.

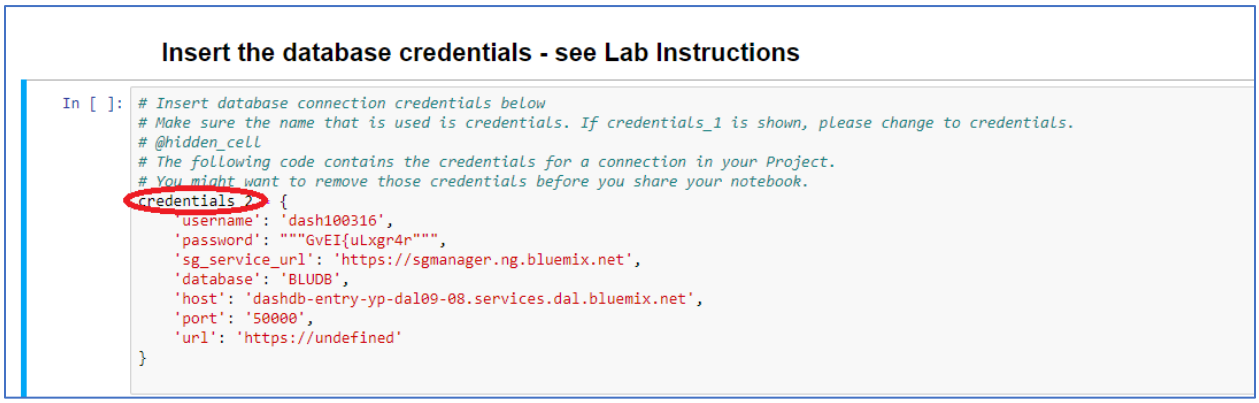


3. Click on 

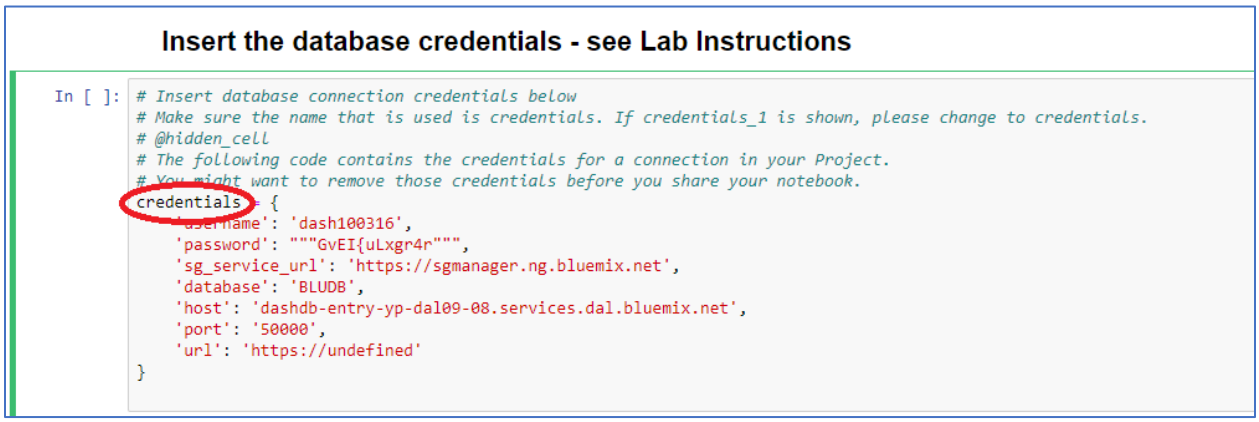
down arrow underneath **trafficking**. Click on **Insert Credentials.**



4. Locate the variable **credentials**. Make sure the variable does not have a number appended (e.g. **credentials_1** or **credentials_2** or **credentials_3**, etc). If it does, change the variable to be **credentials** (without a number).



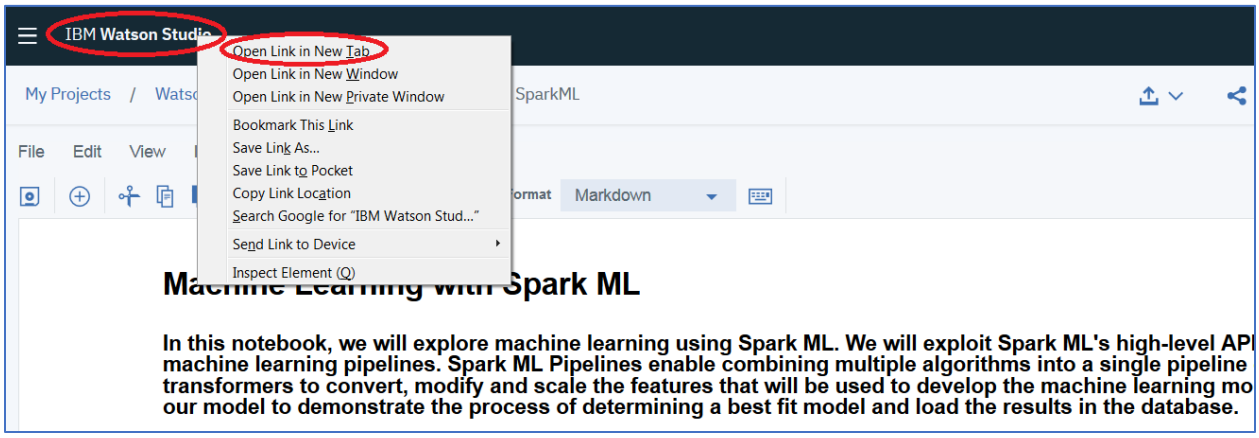
Change to:



Step 3: Copy Watson Machine Learning Credentials

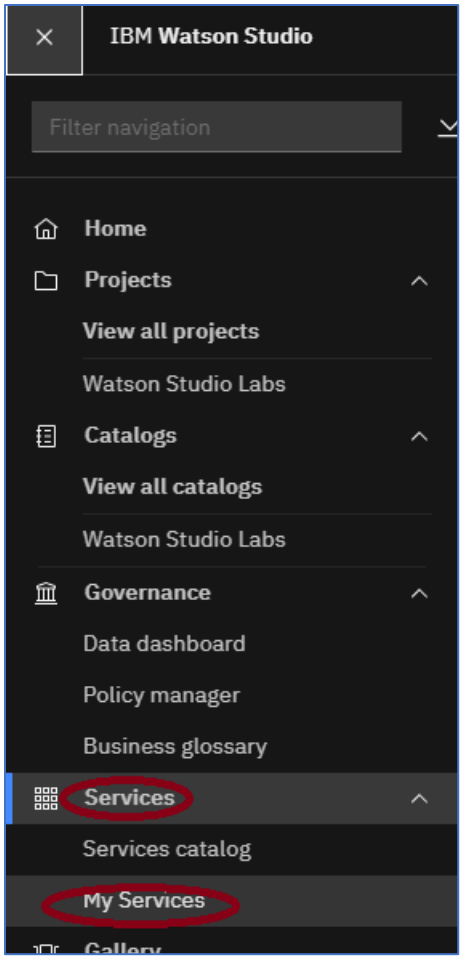
In the notebook we will save our model to the Watson Machine Learning service model repository. We need to get the credentials of the Watson Machine Learning service and will copy and paste these credentials into the appropriate notebook cell.

1. Right-click on **IBM Watson Studio**, and the click on **Open Link in New Tab**.

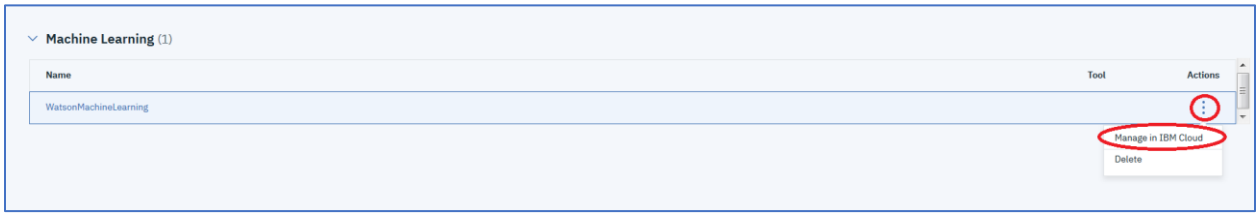


2. Click on the new **Watson Studio** browser tab.

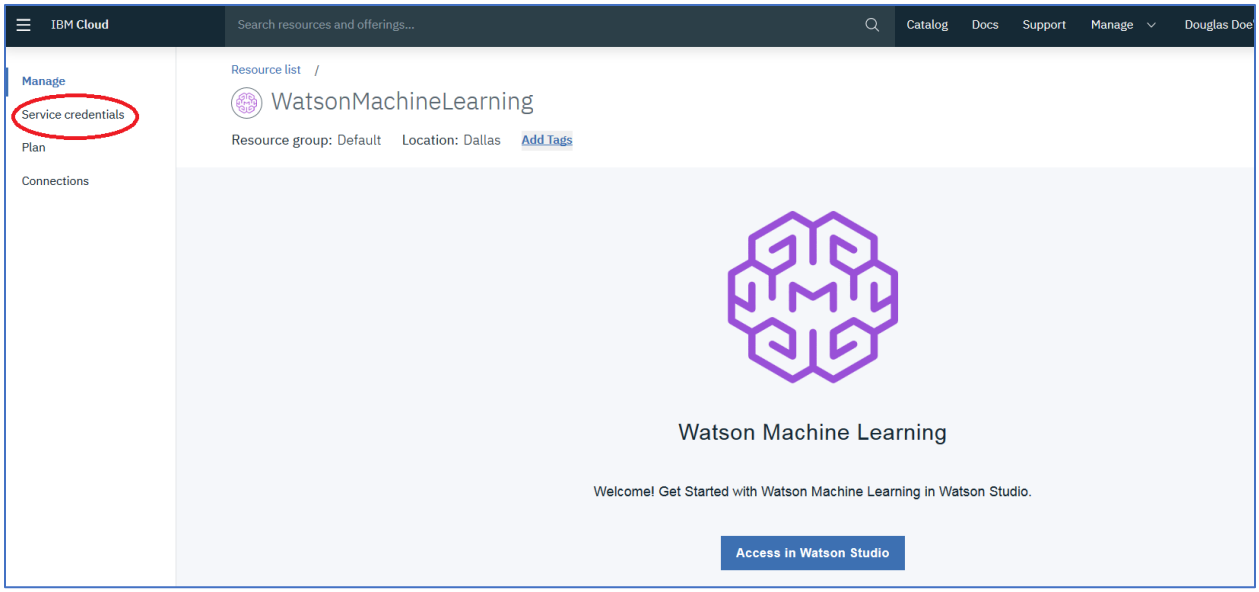
3. Click on the hamburger icon ☰, and then **Services**, and **Watson Services**.



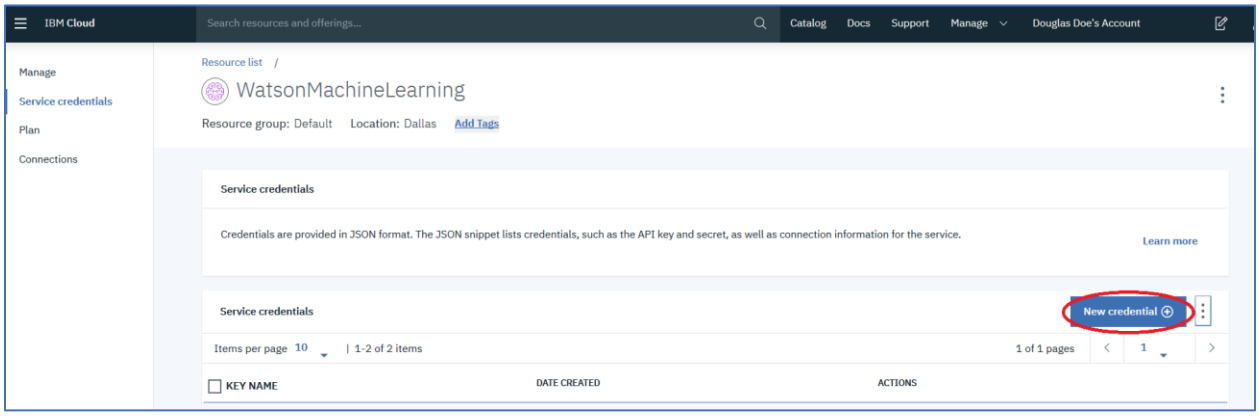
4. Click on the vertical ellipse ⋮ on the right side of the WatsonMachineLearning service, and click on **Manage in IBM Cloud**.



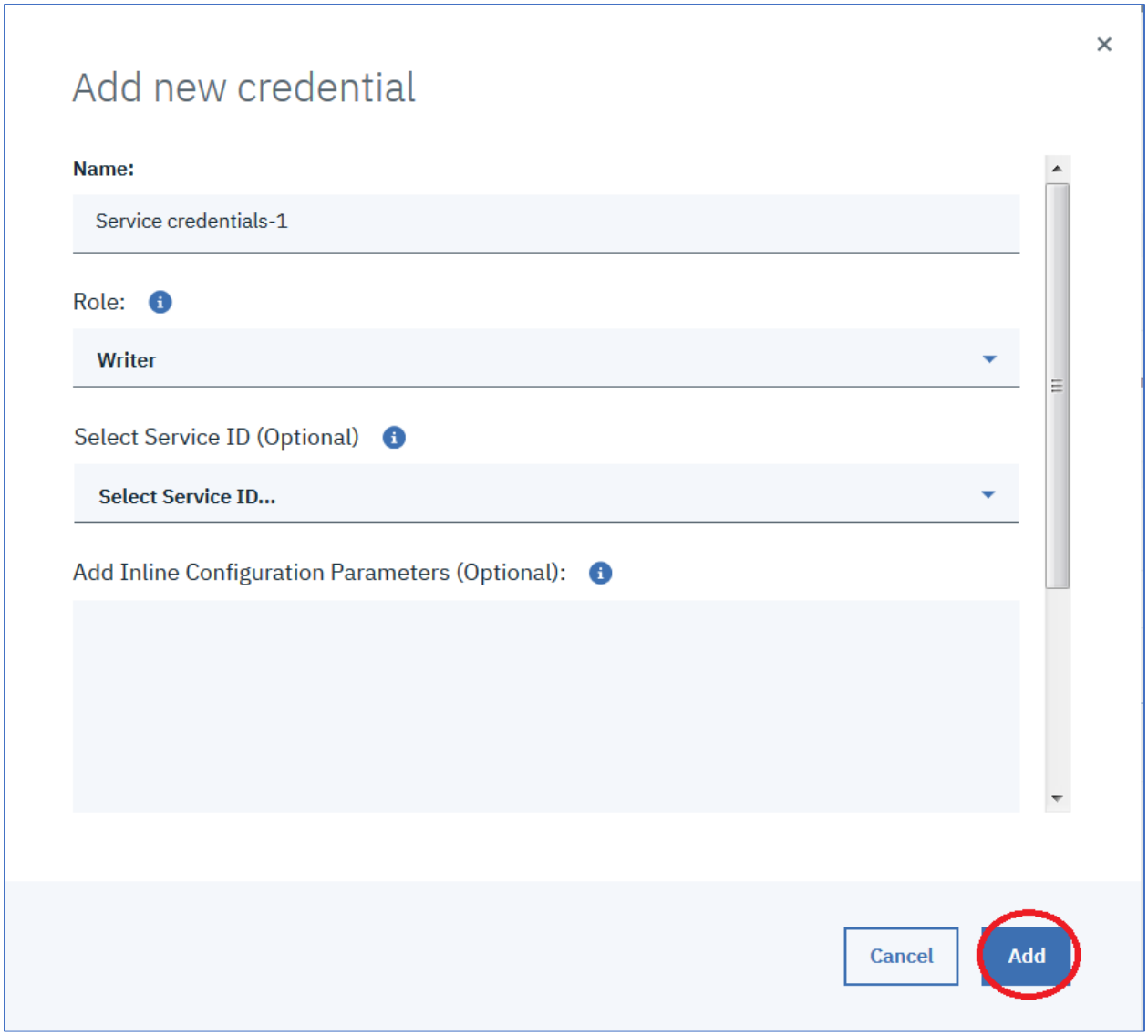
5. A new browser tab will be created titled **Service Details – IBM Cloud.** This browser tab will be interfacing with the IBM Cloud user interface. Click on **Service credentials**.



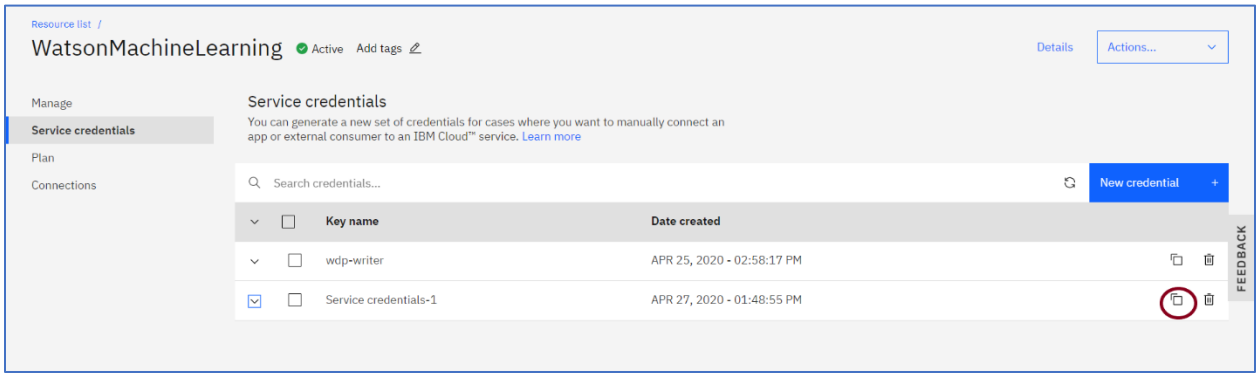
6. Click on **New Credentials**+.



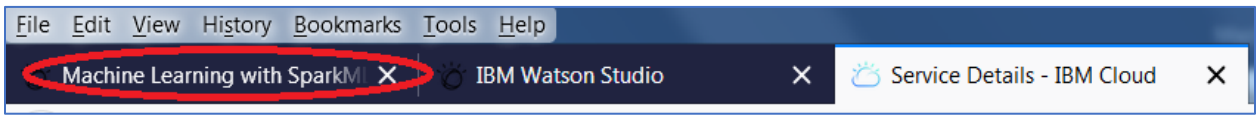
7. Click on **Add**.



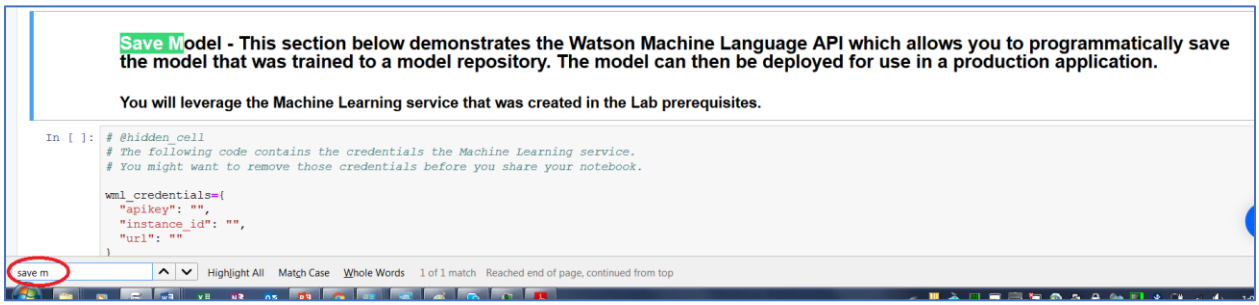
8. Click on the copy icon on the right side of the Service Credentials-1 row.



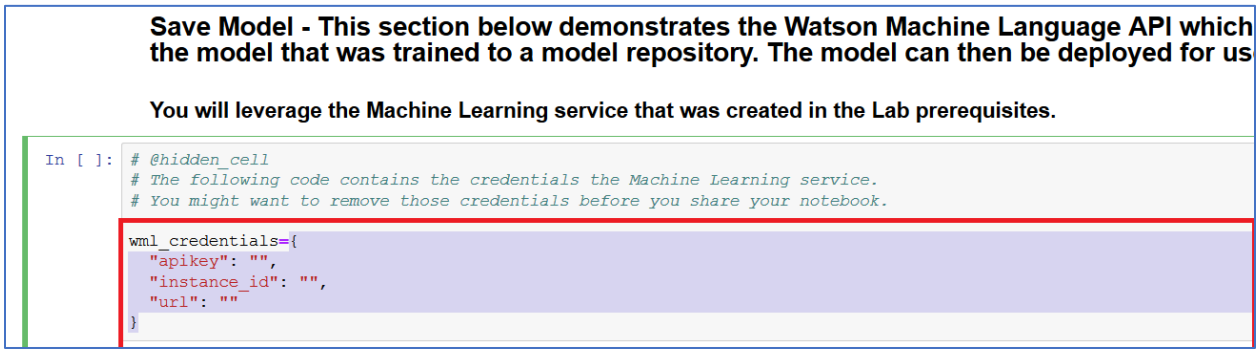
9. Click on the Machine Learning browser tab to go back to the notebook. Should be two browser tabs to the left of the **Service Details – IBM Cloud**.



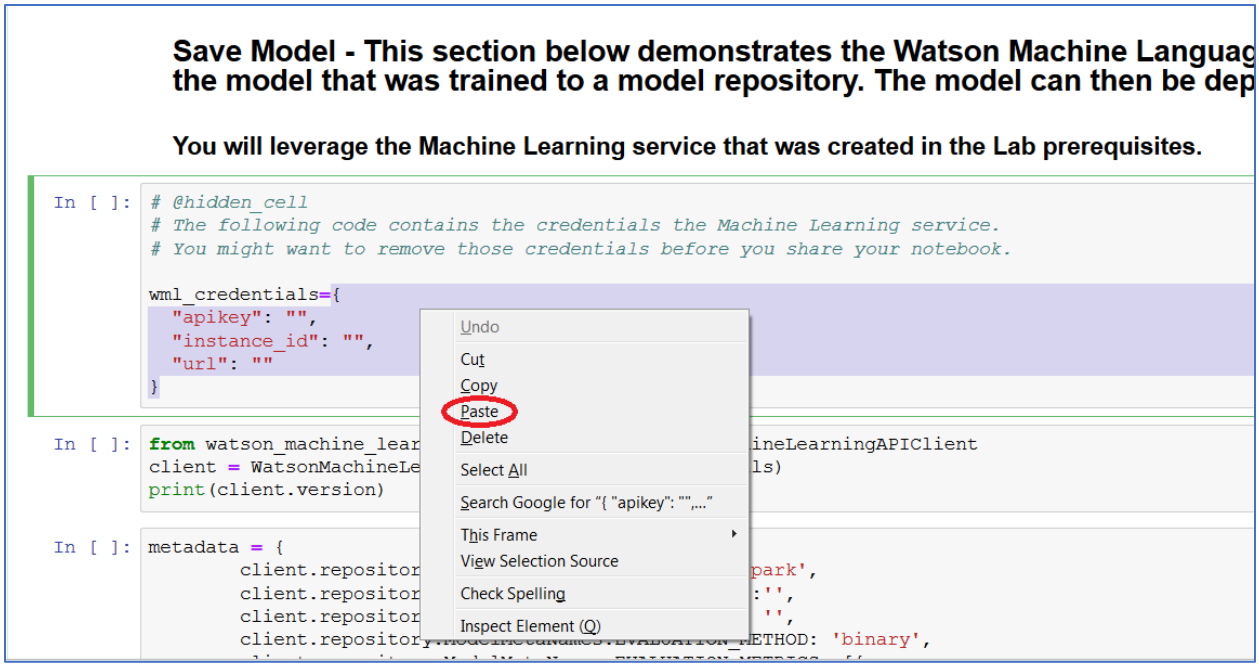
10. Type Ctrl-F and type in **save m** to find the notebook cell to paste in the credentials.



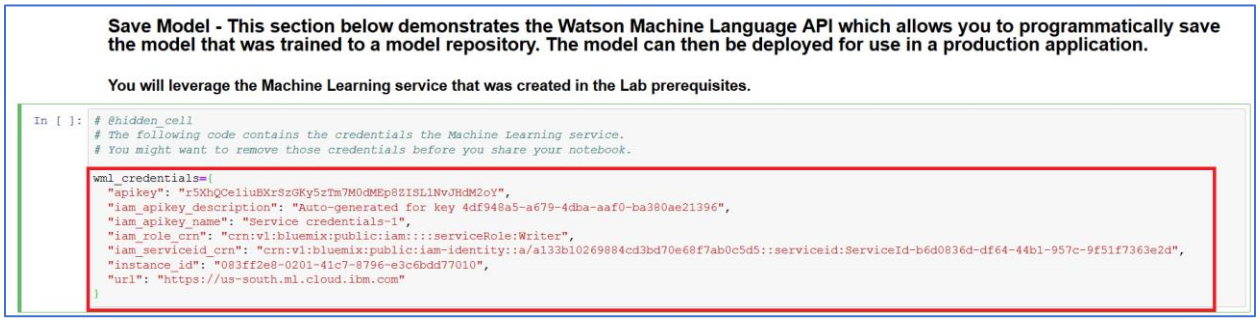
11. Highlight the text from the starting curly brace { to the ending curly brace }.



12. Right-click on the highlighted text and click **Paste**.

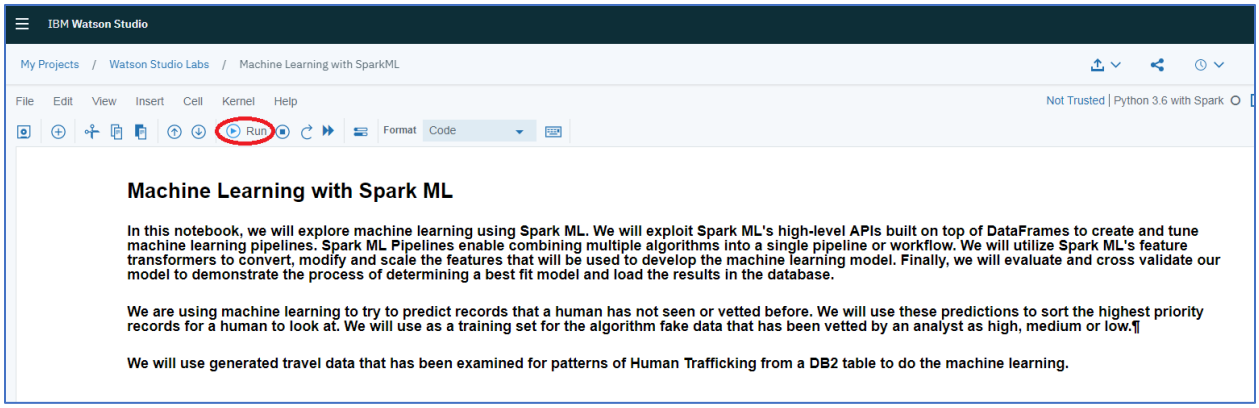


13. The credentials should appear similar to below with different values.

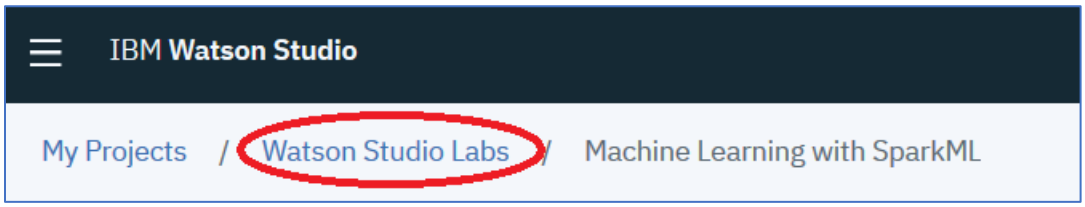


Step 4: Execute the code cells in the notebook

1. Scroll back to the top of the notebook. Click in to the first cell. Execute each of the code cells in order by clicking into each code cell starting at the top and pressing the **<Shift><Enter>** keys or by clicking into the first code cell and using the Run icon in the menu bar at the top. Read the documentation to gain an understanding of the code that is executing. **When all the cells in the notebook have been successfully executed, please return to this document, and continue with Step 2.**



2. Type **Ctrl-S** to save the notebook. Exit out of the notebook by clicking on the Watson Studio Labs in the breadcrumb area.



3. Scroll down the **Assets** page until you see the **Models** heading. The model listed was generated programmatically from the notebook using the Watson Machine Learning APIs. You should see the **FHT_Spark** model in the list of Model Assets.

You have completed the Lab-!

- ✓ Joined data from three sources.
- ✓ Identified labels and transformed data.
- ✓ Conducted feature engineering for algorithm data.
- ✓ Declared a machine learning model.
- ✓ Created the Pipeline for data transforms and training.
- ✓ Trained the model.
- ✓ Evaluated and showed model results.
- ✓ Automatically tuned model.

- ✓ Scored data and loaded into a new DB2 table.
- ✓ Saved the model to the model repository.