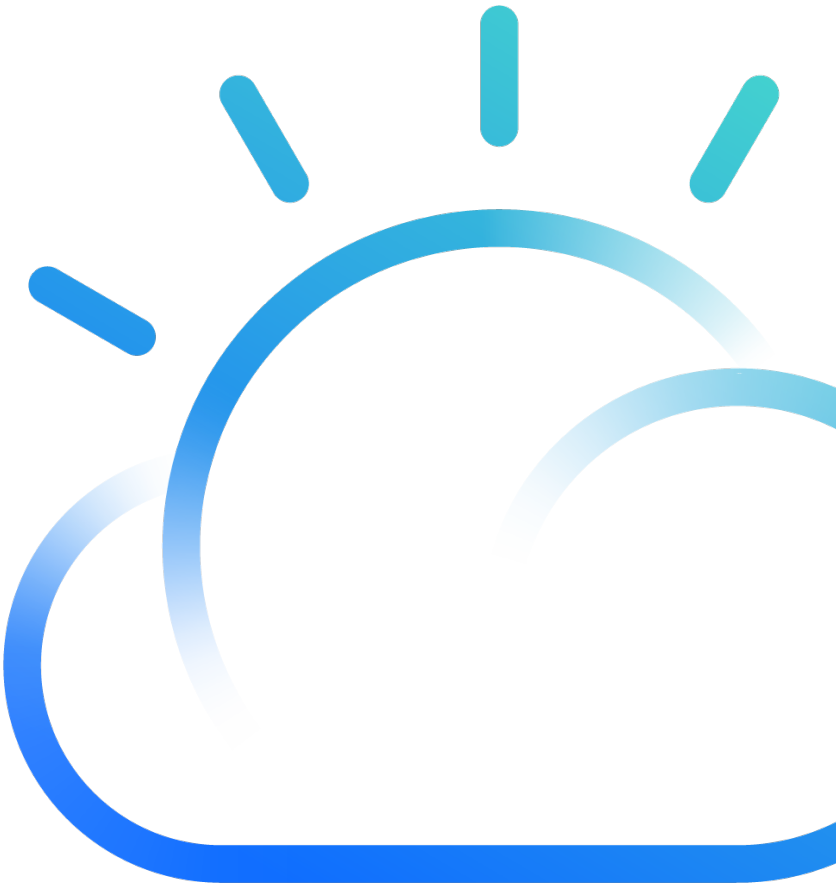
IBM Skills Academy

Lab Guide

Data Modeling and Validation



IBM

The information contained in this document has not been submitted to any formal IBM test and is distributed on an "as is" basis without any warranty either express or implied. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will result elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk.

© Copyright International Business Machines Corporation 2019.
This document may not be reproduced in whole or in part without the prior written permission of IBM.
US Government Users Restricted Rights - Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Contents

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

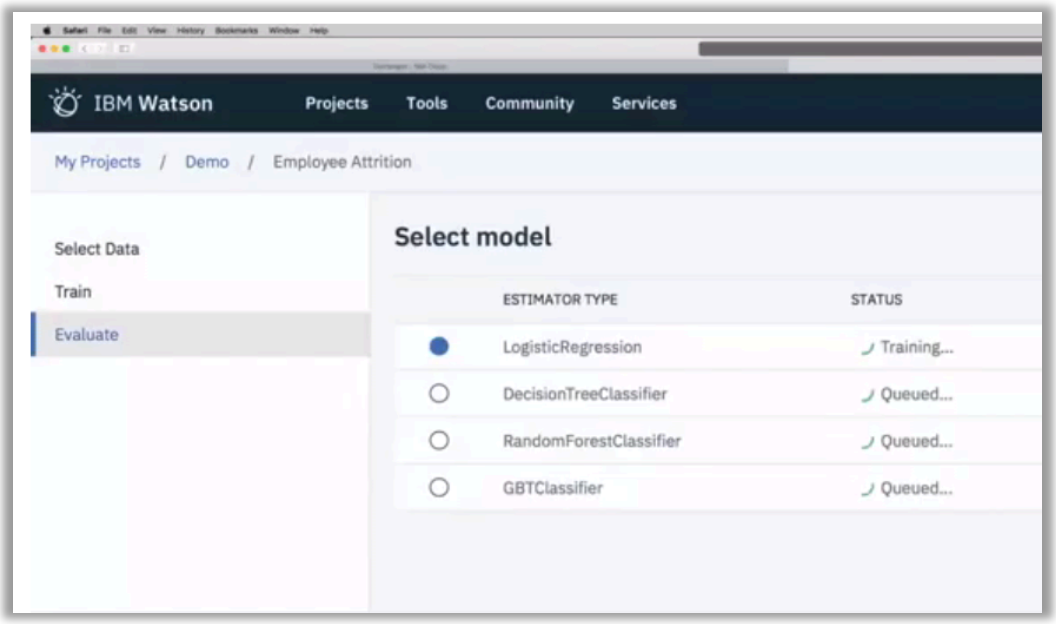
Section 1. Overview

Consider the following use case. A wonderful company with a healthy and thriving culture set in a scenic country setting has been experience alarming levels of employee attrition. It's not just that the human resources (eh, talent managers) have noticed, so have fellow employees.

Before long, the head of Human Resources, taps a Data Journalist on the shoulder and gives her a giant spread sheet hundreds of rows (employees) and dozens of columns (attributes such as age, sex, education, distance from home, you name it…. whatever can be gathered under the current GDPR guidelines).

The Analyst takes the spreadsheet and feeds it to a black box (it's a linear regression model) out comes colorful charts, scatter plots, bar carts, Pareto distribution. She applies a myriad of dependent variables to the constant of employee attrition and soon it emerges that single employees below the age of 24 who live 30+ miles from work are the first to leave. They seem to be going to firms inside the bustling cities where they can 'share' a scooter while commuting to work.

In this scenario, you will then switch caps to that of a Data Scientist, and use the appropriate machine learning model, such as Binary Classification (and yes you have others to choose from); plus, peg the results against four distinct algorithms: Logistical Regression, Decision Tree Classifier, Random Forrest Classifier and Gradient Boosted Tree Classifier.



Noteworthy of mention, that in this lab, you will be using two services from the IBM Cloud Catalog: Watson Studio and Machine Learning.

Objectives:

- Create a new Watson Studio project
- Import data set from your local drive (as you download from the Box folder)
- Perform new set of data cleansing and transformation activities
- Apply various machine learning models
- Conclude which model give the best prediction for employee attrition.

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

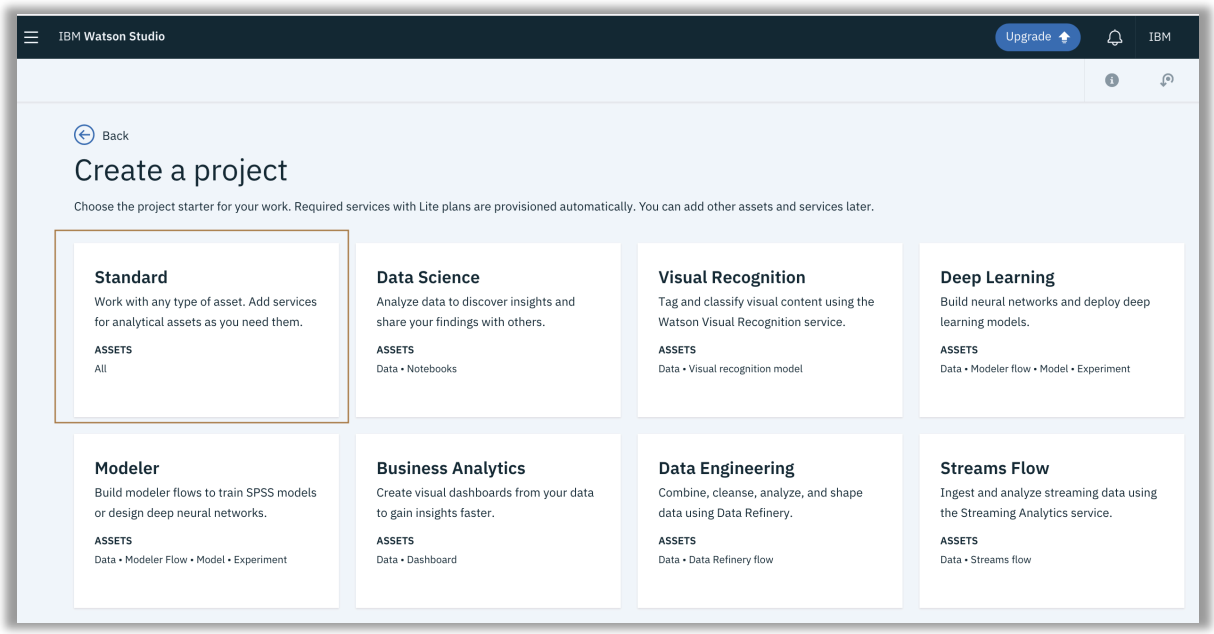
Section 2. Working with Watson Studio

By now you have already registered with IBM Cloud and applied your promocode. Let's begin our journey:

1. Login into IBM Cloud: https://console.bluemix.net/catalog/
2. Click the **Catalog** tab.
3. Search for the **Watson Studio** service and click that tile.
4. Click **Create**.
5. Click the **Get Started** button.
6. Click **Let's get started**.

After your collection has been created, you can immediately start uploading content using the upload area at the right of the screen. However, before you add your own content to the Discovery service, best practice is to configure the service to process the content the way that you want.

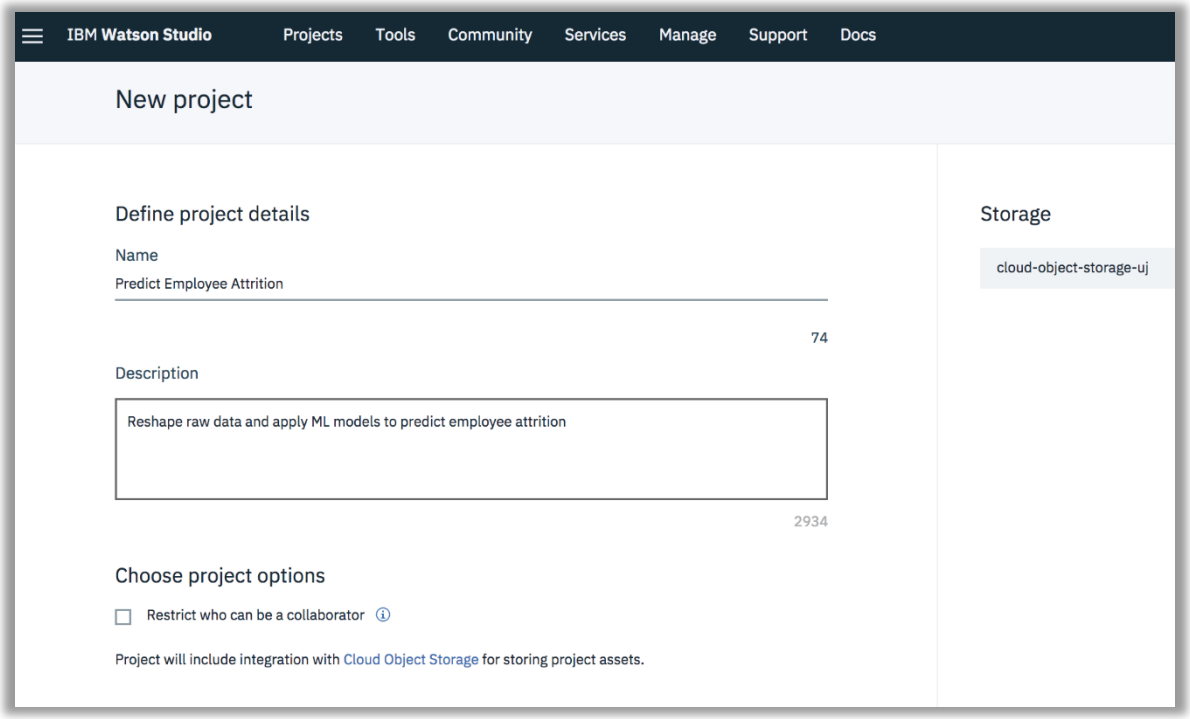
1. Click **New project**.
2. Select the **Standard** tile.
3. Click **OK**.



4. Specify a name. In this example, it is **Predict employee attrition**.

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

5. Specify a description; for example, **Reshape raw data and apply ML models to predict employee attrition**



6. Click **Create**.

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Add Data

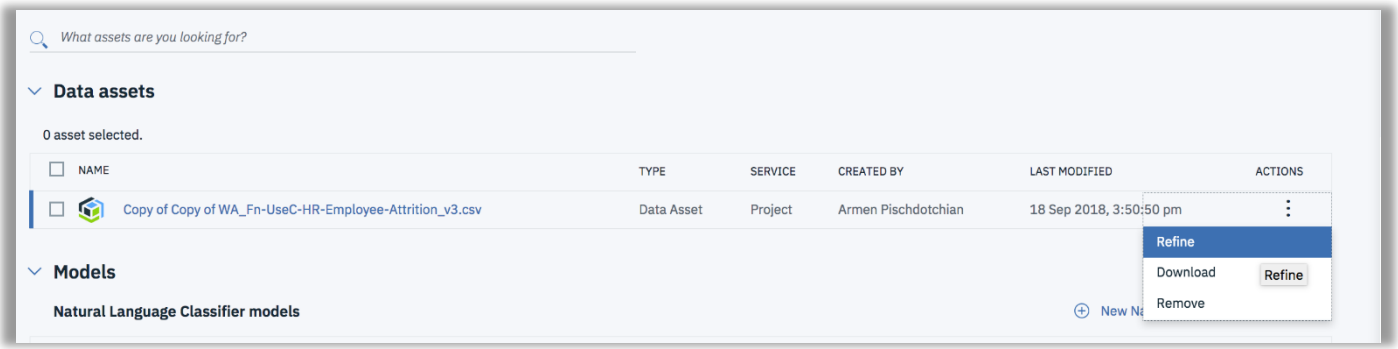
You are now ready to add data to your project. You can upload from a local drive, from a database or from the Communities.

1. In this scenario, you will upload data from this link: [WA_Fn-UseC_-HR-Employee-Attrition.xlsx]
2. Open the file, delete the second worksheet and save it as CSV.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Age	Attrition	BusinessTra	DailyRate	Department	DistanceFro	Education	EducationFie	EmployeeCo	EmployeeNu	Environment	Gender
2	41	Yes	Travel_Rarel	1102	Sales	1	2	Life Sciences	1	1	2	Female
3	49	No	Travel_Frequ	279	Research & D	8	1	Life Sciences	1	2	3	Male
4	37	Yes	Travel_Rarel	1373	Research & D	2	2	Other	1	4	4	Male
5	33	No	Travel_Frequ	1392	Research & D	3	4	Life Sciences	1	5	4	Female
6	27	No	Travel_Rarel	591	Research & D	2	1	Medical	1	7	1	Male
7	32	No	Travel_Frequ	1005	Research & D	2	2	Life Sciences	1	8	4	Male
8	59	No	Travel_Rarel	1324	Research & D	3	3	Medical	1	10	3	Female
9	30	No	Travel_Rarel	1358	Research & D	24	1	Life Sciences	1	11	4	Male
10	38	No	Travel_Frequ	216	Research & D	23	3	Life Sciences	1	12	4	Male
11	36	No	Travel_Rarel	1299	Research & D	27	3	Medical	1	13	3	Male
12	35	No	Travel_Rarel	809	Research & D	16	3	Medical	1	14	1	Male
13	29	No	Travel_Rarel	153	Research & D	15	2	Life Sciences	1	15	4	Female
14	31	No	Travel_Rarel	670	Research & D	26	1	Life Sciences	1	16	1	Male
15	34	No	Travel_Rarel	1346	Research & D	19	2	Medical	1	18	2	Male
16	28	Yes	Travel_Rarel	103	Research & D	24	3	Life Sciences	1	19	3	Male
17	29	No	Travel_Rarel	1389	Research & D	21	4	Life Sciences	1	20	2	Female

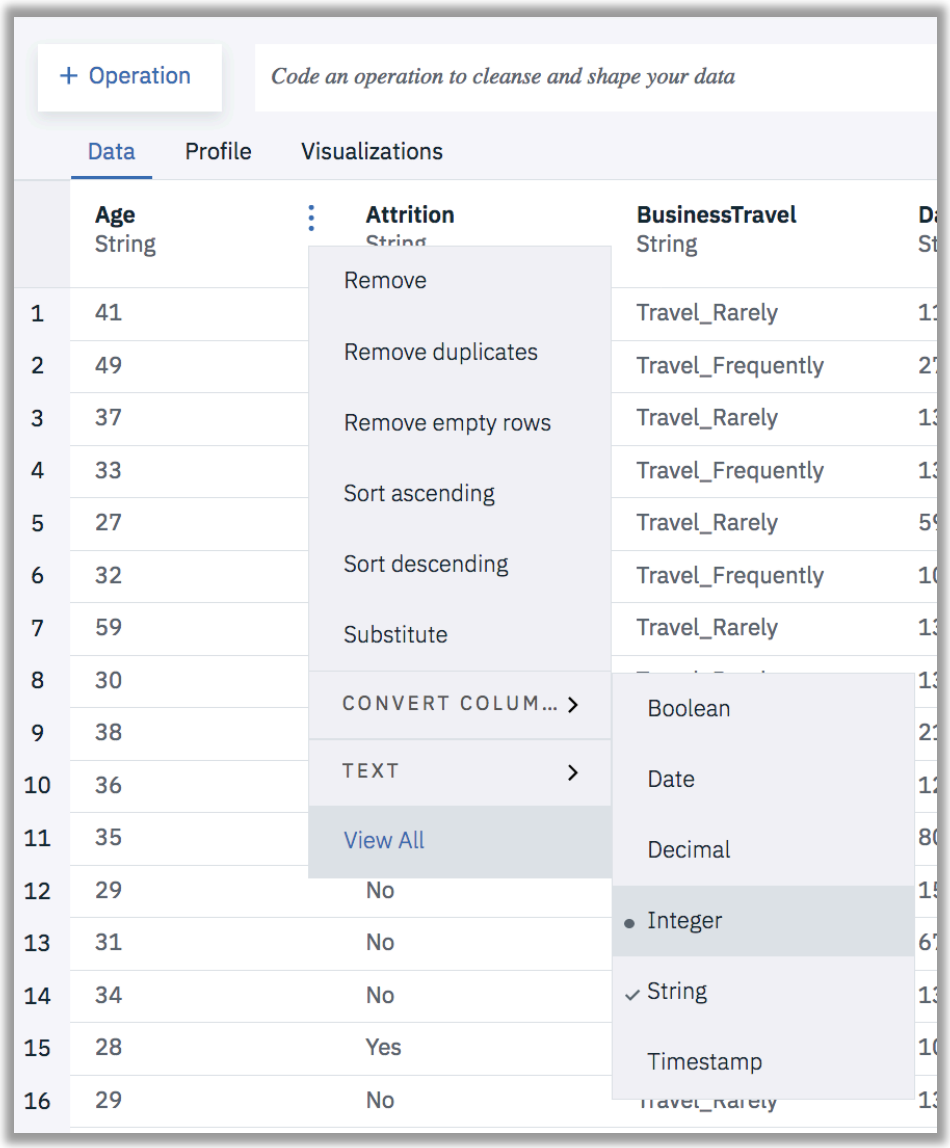
3. Click the **Assets** tab.
4. Use the **browse** link to navigate to your recently downloaded asset.
5. Select the **ACTION** three dots and click **Refine**.

   Notice that some of the columns appear as string and they should be integers.



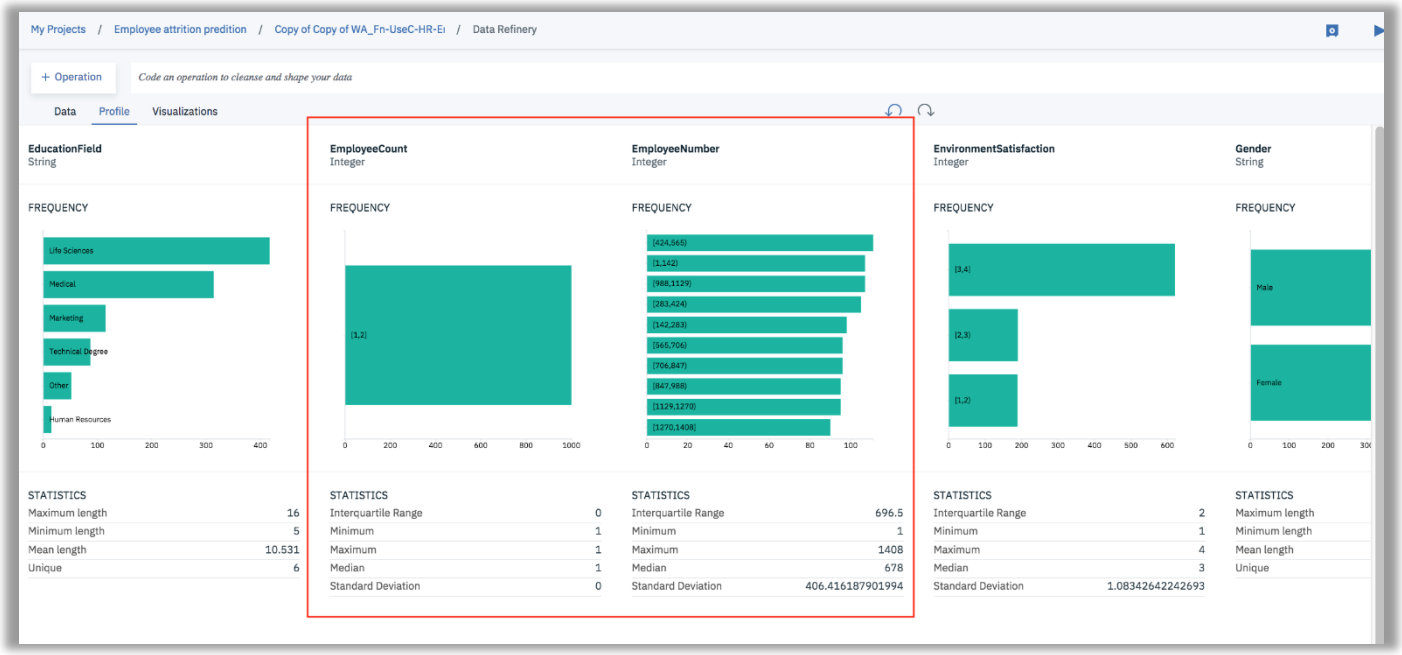
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

6. Take your time and convert the value of each column to be the 'suggested' value as deemed by a dot.
7. You may have to scroll forward each time the page gets reset back to the beginning of the data flow.
8. Save the data flow .



Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

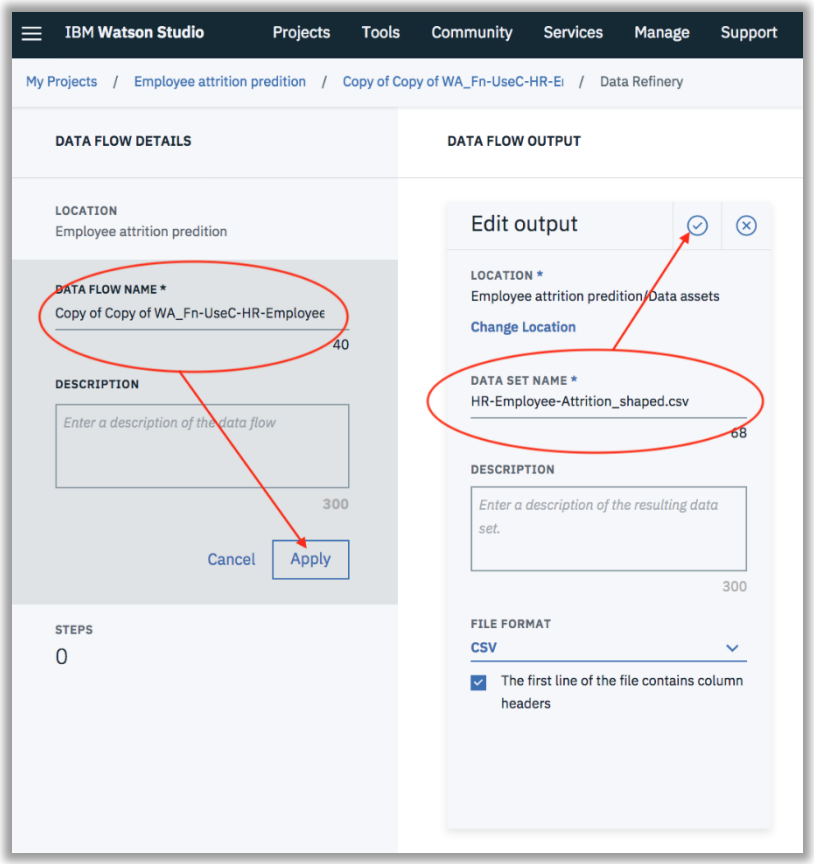
9. Click the **Profile** Tab. Notice the columns, namely the **EmployeeCount**, **EmployeeNumber**, **Over18** and the **StandardHours** columns do not lend useful values as to why one might leave.



10. Go back to the **Asset** tab.
11. **Remove** those columns.
12. Save the data flow .

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

13. Click Run the data flow. 
14. Edit and change the name of the **DATAFLOW DETAILS**; for example, change it to:
    `HR_Employee_attrition_baseline.csv_flow`
15. click **Apply**.
16. Change the name of the **DATA FLOW OUTPUT**, to reflect the same name:
    `HR_Employee_attrition_baseline.csv_flow`
17. Click the **check mark**.
    `HR_Employee_attrition_baseline.csv_shaped.csv`



18. Click **Save and run** button at the bottom right corner of the form.
19. Click **View flow** in the ensuing dialog box and allow enough time for the status to appear as **Completed**.
20. Click **Refine** to view the changes.

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

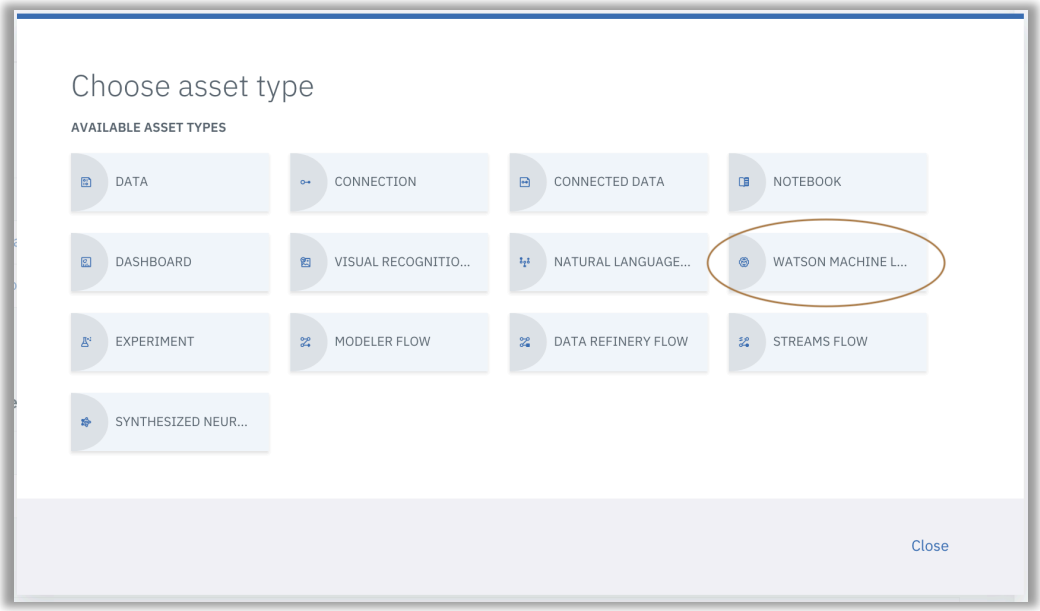
Machine Learning Models

You are now ready to use machine learning models, and you will run 4 distinct models against your data to see which on produces the best accuracy of prediction.

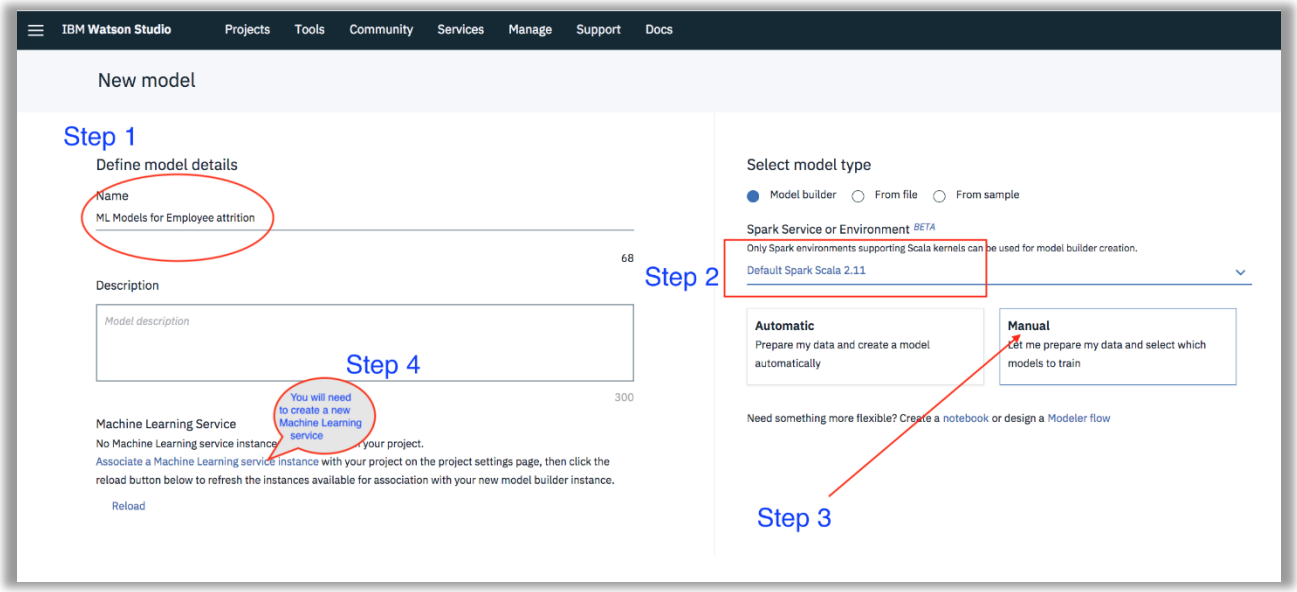
1. Click the project name from the breadcrumb link:

   `Projects/`**`Employee attrition prediction`**`/……/……`

2. From the top right, click **Add to project** and select **Watson Machine Learning.**



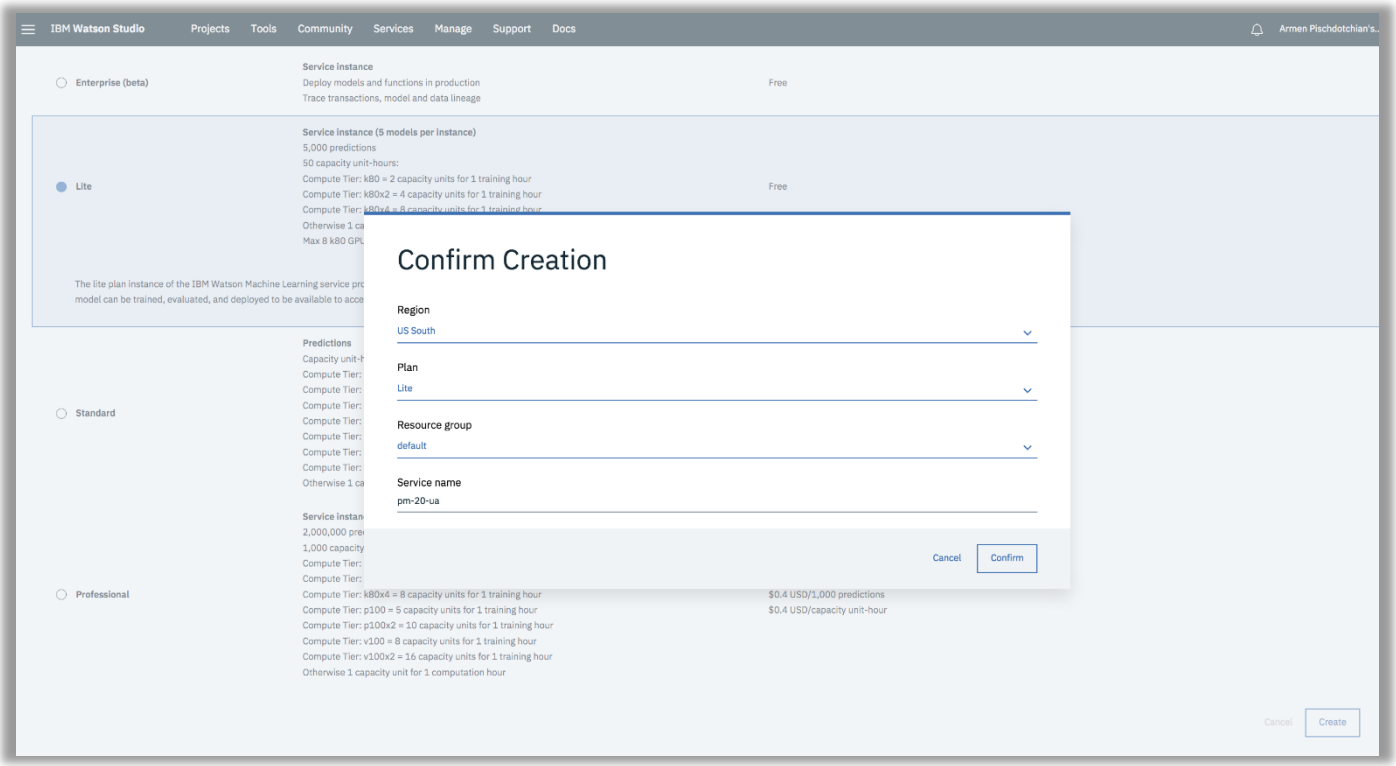
3. Define a model name. For example:
   ML models for employee attrition
4. From the Drop-down list, select the default Spark environment
5. Click the **Manual** box so you can select which models to use for your use case.
6. Click **Associate a Machine Learning service instance**.



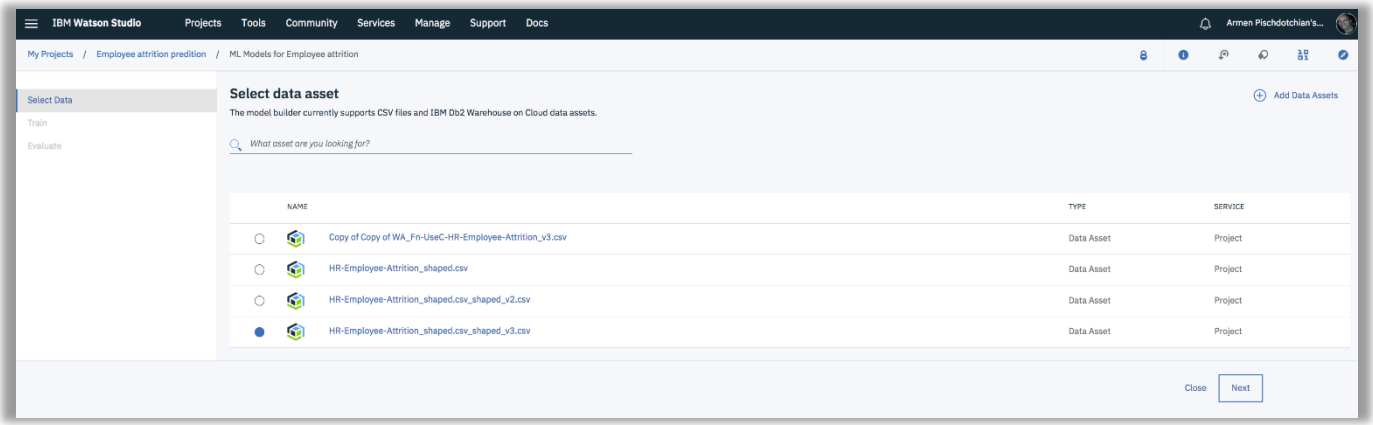
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

And now, a little about machine learning:

1. The link from the previous step takes you to the WML service. Select the **Lite plan (default)** and click **Create**. Ensure there are values in the Confirm Creation box and click **Confirm**.
2. Once the service has provisioned, it will bring you back to the Project page. Click **Reload**.
3. Notice, the service instance name appears
4. Click **Create** to proceed with your new models.

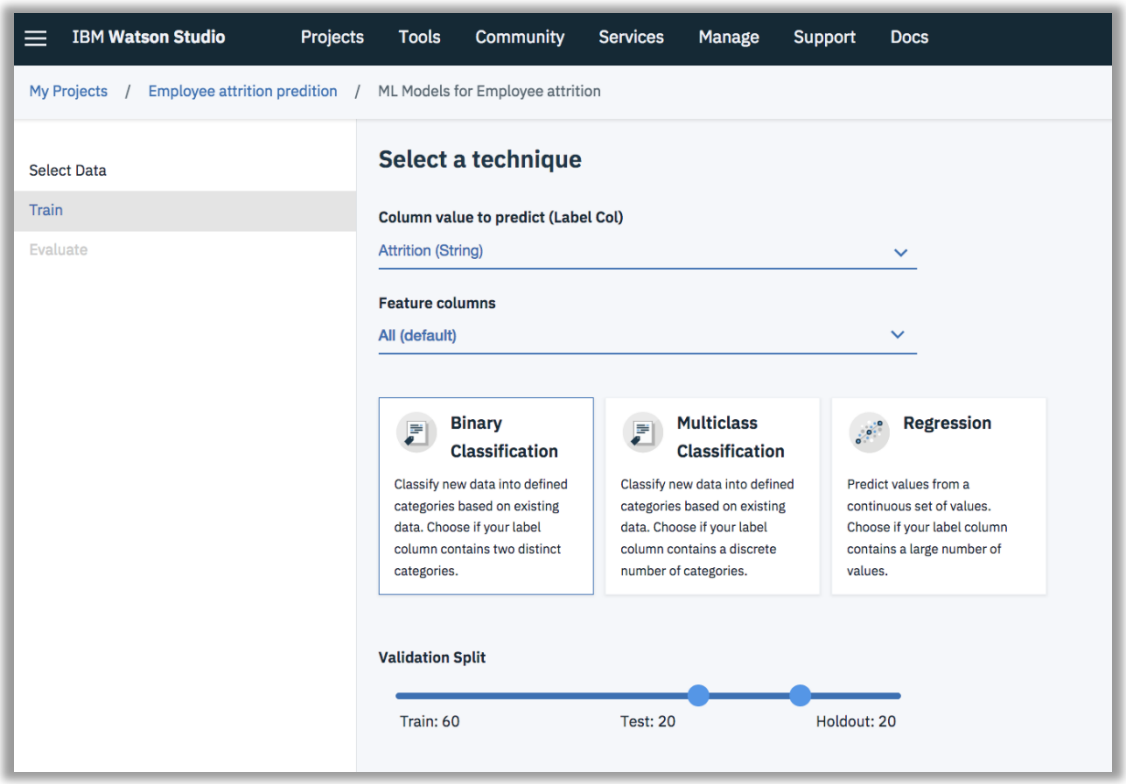


5. Select you most recent data asset HR_Employee_attrition_baseline.csv_shaped.csv
6. Click **Next**.



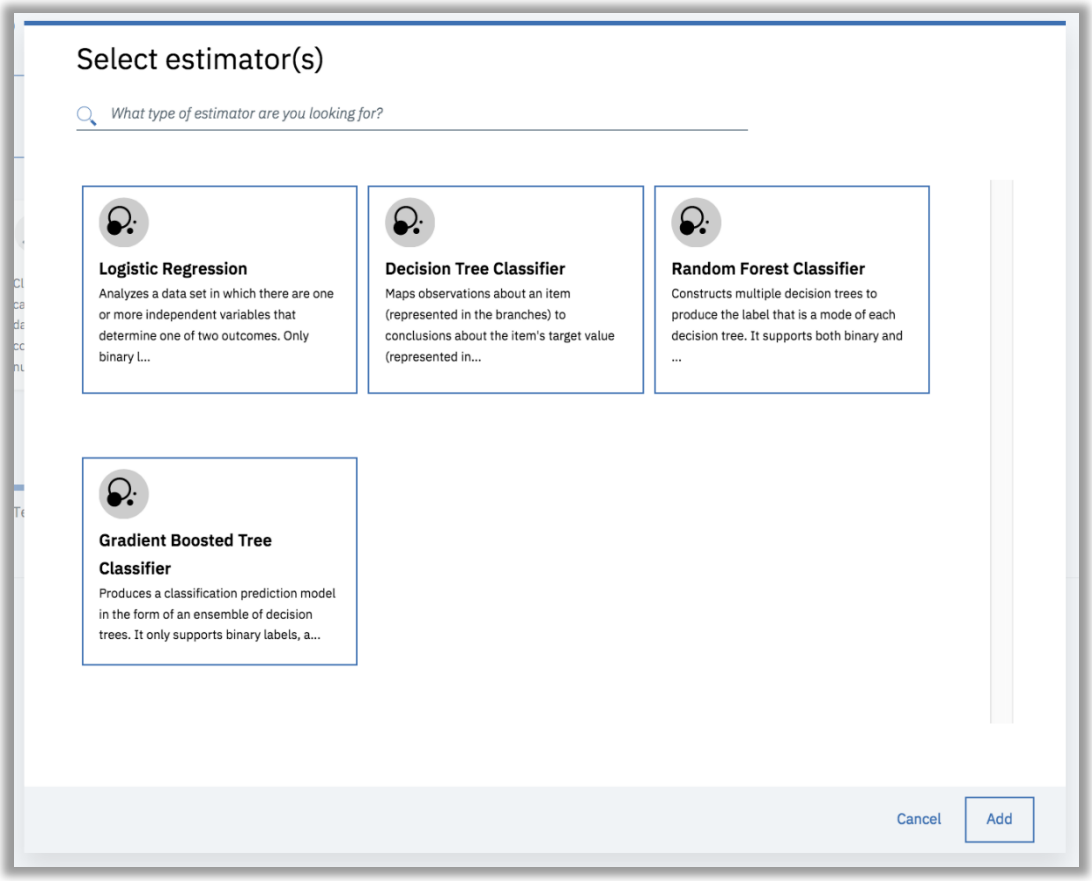
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

7. For the Column value to predict, select **Attrition** (String)
8. For Feature columns, keep the **All (default)** setting.
9. In this use case you are addressing the question of Yes, employee will stay, and No employee will leave; thus, it is a binary decision. It is also a classification problem not a regression challenge. Select **Binary Classification** as the model type.

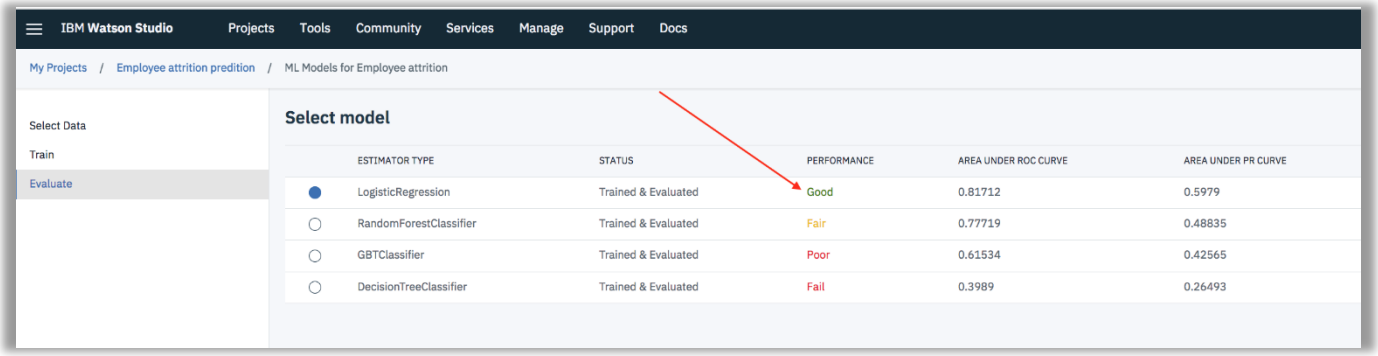


Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

10. Click **Add estimators** from the top right corner.
11. Select all four models. Let's see which model ends up with the highest accuracy.
12. Click **Add**.



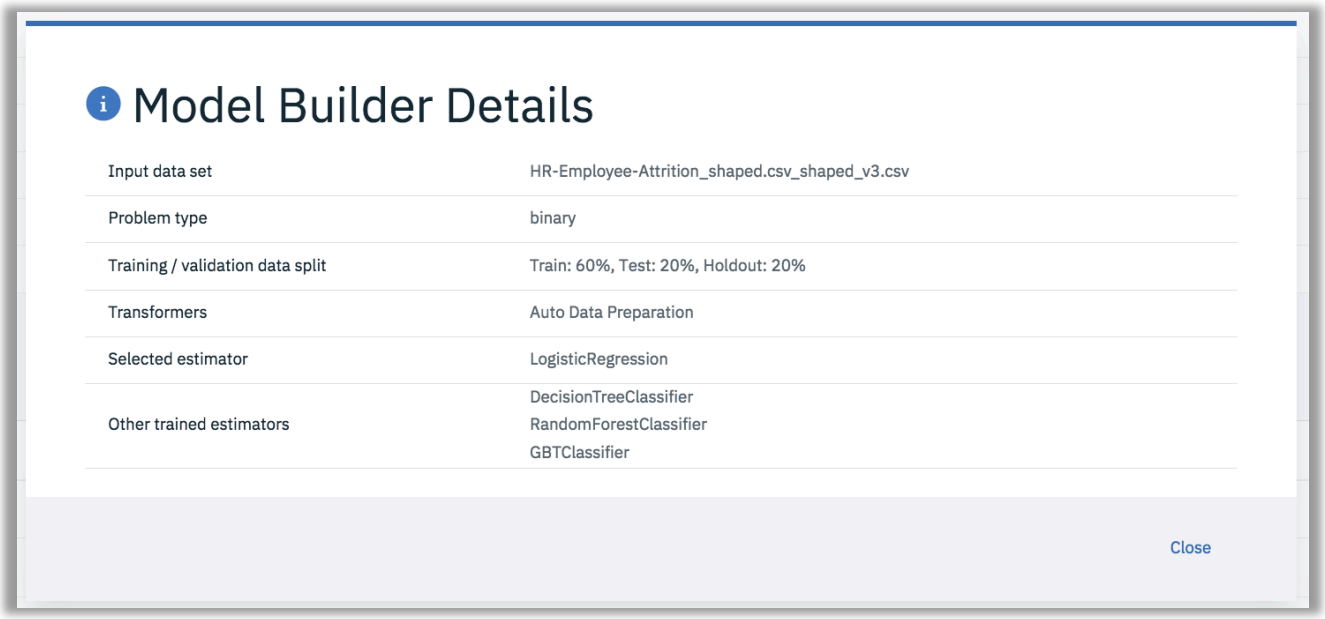
13. Click **Next**.
14. Allow enough time to pass for the models to build. Notice that Logistical Regression gave the best results.



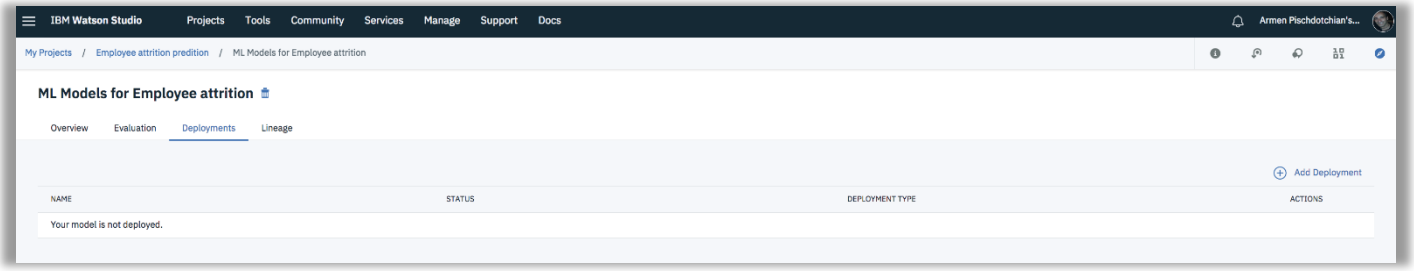
15. Go back to the **Overview** tab.
16. Click Save.

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

17. Click **View** in the Model builder details. Notice, the percentages allocated to train, test and holdout. Soon you will be testing the data.



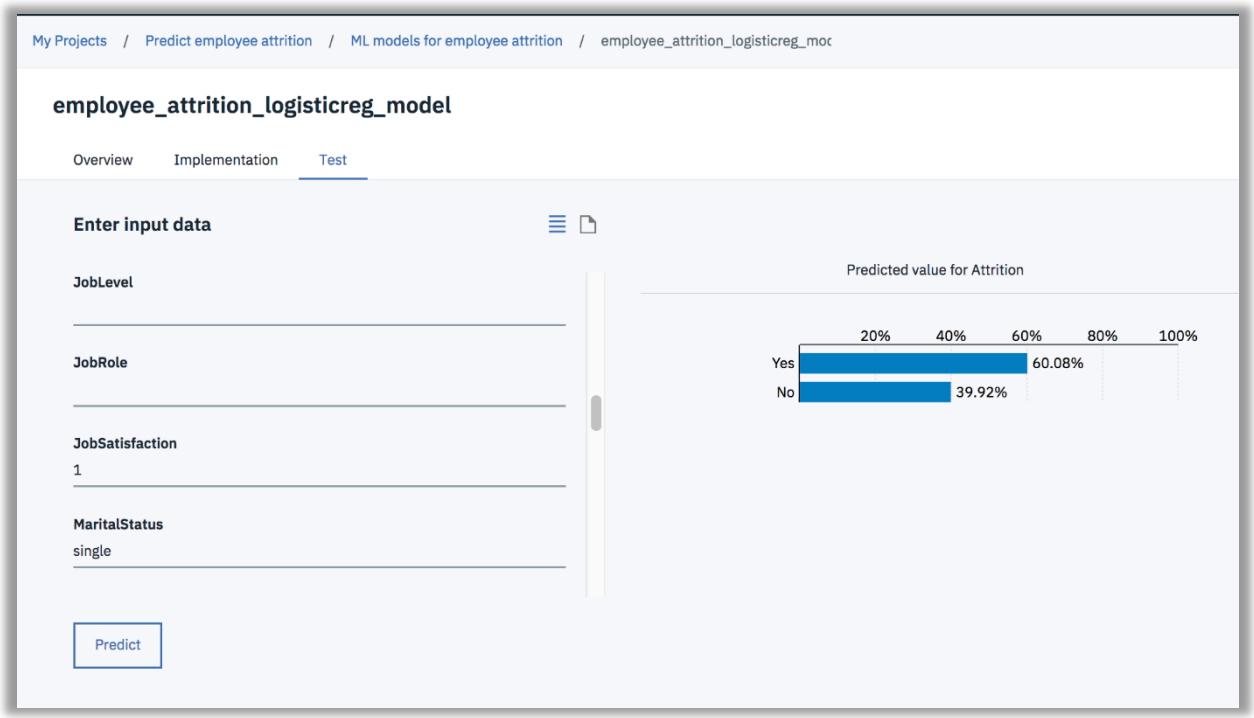
18. Click **Close**.
19. Click the **Deployment** tab.
20. Click **Add deployments**.



21. Specify a name; for example: `employee_attrition_logisticreg_model`
22. Click **Save** and allow enough time for deployment reveal success status.
23. Click the deployment model link.
24. Click the **Test** tab.
25. Use the spreadsheet to select the proper values.

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

26. Experiment with a few parameters. For example, for *Age*, enter **22**, for *JobSatisfaction*, enter **1** and **single** for *MaritalStatus*.



27. Click **Predict**.

28. Enter a few other values of your choosing and observe the newly generated prediction values.

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

© Copyright IBM Corporation 2019.

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.

IBM, the IBM logo and ibm.com are trademarks of International Business Machines Corp., 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.

Please Recycle

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.