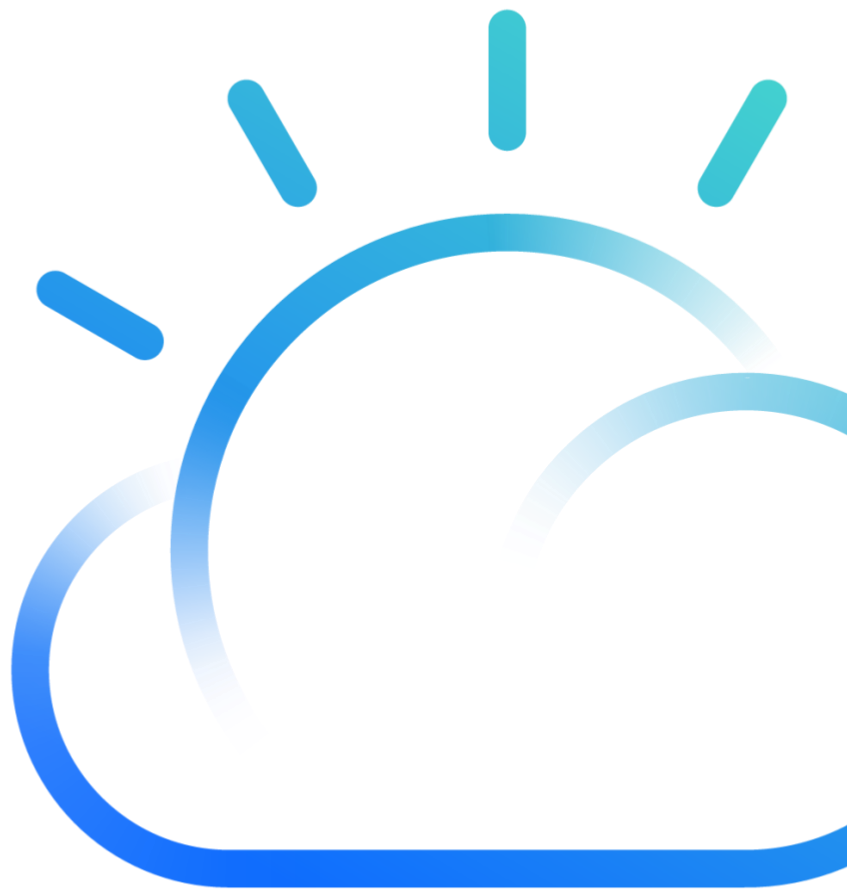


Apply ML Models to Employee Attrition

Lab 7 Guide



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.

Contents

SECTION 1.	PREFACE	4
	OVERVIEW.....	4
	AUTOAI PROCESS.....	4
	OBJECTIVES:.....	5
SECTION 2.	WORKING WITH WATSON STUDIO	6
	ADD DATA	7
	WORKING WITH AUTOAI	10
	A. DATA PRE-PROCESSING.....	12
	B. AUTOMATED MODEL SELECTION.....	13
	C. AUTOMATED FEATURE ENGINEERING.....	13
	D. HYPERPARAMETER OPTIMIZATION.....	13
	EVALUATE THE PIPELINES	14

Section 1. Preface

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.

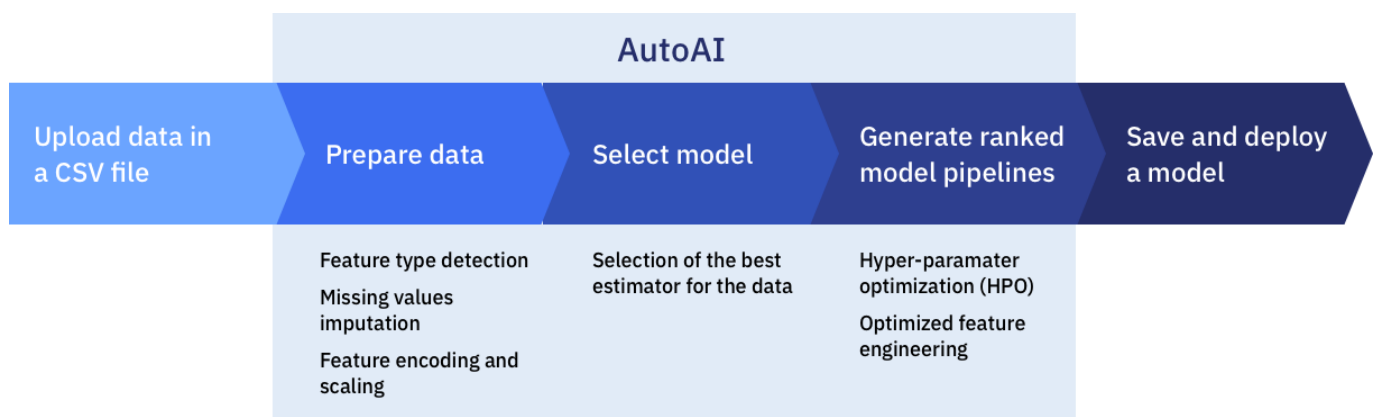
In this scenario, you will then switch caps to that of a Data Scientist, and use the appropriate machine learning model, such as Binary Classification (and yes you have others to choose from); plus, peg the results against four distinct algorithms: Logistical Regression, Decision Tree Classifier, Random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

The AutoAI graphical tool in Watson Studio automatically analyzes your data and generates candidate model pipelines customized for your predictive modeling problem. These model pipelines are created over time as AutoAI algorithms learn more about your dataset and discover data transformations, estimator algorithms, and parameter settings that work best for your problem setting. Results are displayed on a leaderboard, showing the automatically generated model pipelines ranked according to your problem optimization objective.

Note that the file type supported is a CSV file and has to be less than 100MB.

AutoAI process

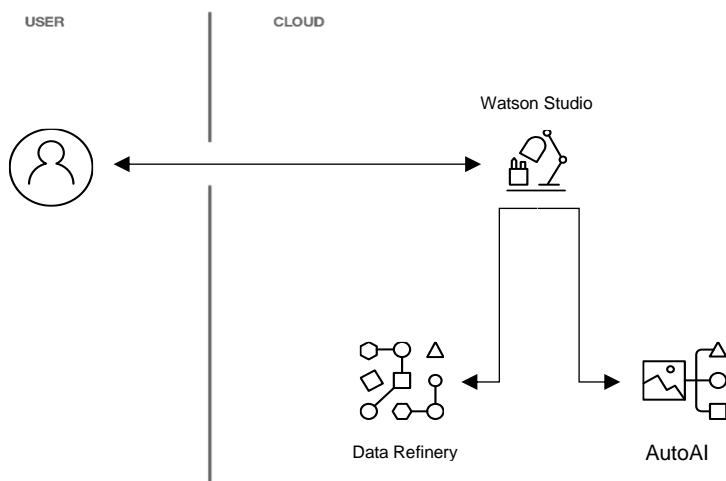
Using Auto AI, you can build and deploy a machine learning model with sophisticated training features and no coding. The tool does most of the work for you.



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.

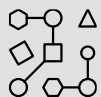
Flow



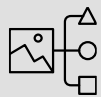
Tools



Watson Studio



Data Refinery Tool



AutoAI

You have previously worked with Watson Studio and AutoAI in prior labs. In this lab you will use AutoAI and discover the various estimators that the system pick automatically to make predictions using your data set.

Prerequisites

This is a stand-alone lab; however, having completed Lab 2 and Lab 3 will help as you work with the Data Refinery tool.

Section 2. Working with Watson Studio

Let's begin our journey:

1. Login into IBM Cloud: <https://cloud.ibm.com/registration>
2. Remove the **Label:lite** filter.
3. Click the **Catalog** tab.
4. Search for the **Watson Studio** service and click that tile.
5. Click **Create**.
6. Click the **Get Started** button.
7. Click **Get started** again.
8. Click **Create a project**.
9. Select the **Create an empty project** tile.
10. Specify a name. In this example, it is **Predict employee attrition**.
11. Specify a description; for example, **Reshape raw data and apply ML models to predict employee attrition**

IBM Watson Studio Projects Tools Community Services Manage Support Docs

New project

Define project details

Name
Predict Employee Attrition

Description
Reshape raw data and apply ML models to predict employee attrition

Choose project options

☐ Restrict who can be a collaborator ⓘ

Project will include integration with Cloud Object Storage for storing project assets.

Storage

cloud-object-storage-uj

12. You may need to add an Object Storage if the cloud-object-storage does not appear and the Create button is greyed out.
13. Click **Create**.

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:

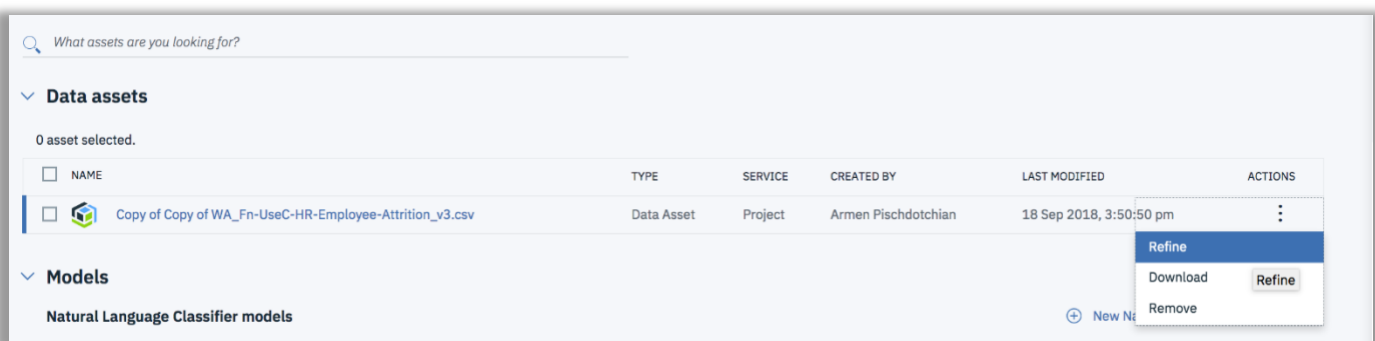
<https://github.com/apischdo/Artificial-Intelligence-and-Data-Science/blob/master/Employee-Attrition.xlsx>

2. Download the xlsx file to your local drive.


	A	B	C	D	E	F	G	H	I	J	K	L
1	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromOffice	Education	EducationField	EmployeeCount	EmployeeNumber	Environment	Gender
2	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	2	Female
3	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	3	Male
4	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	4	Male
5	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	4	Female
6	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	1	Male
7	32	No	Travel_Frequently	1005	Research & Development	2	2	Life Sciences	1	8	4	Male
8	59	No	Travel_Rarely	1324	Research & Development	3	3	Medical	1	10	3	Female
9	30	No	Travel_Rarely	1358	Research & Development	24	1	Life Sciences	1	11	4	Male
10	38	No	Travel_Frequently	216	Research & Development	23	3	Life Sciences	1	12	4	Male
11	36	No	Travel_Rarely	1299	Research & Development	27	3	Medical	1	13	3	Male
12	35	No	Travel_Rarely	809	Research & Development	16	3	Medical	1	14	1	Male
13	29	No	Travel_Rarely	153	Research & Development	15	2	Life Sciences	1	15	4	Female
14	31	No	Travel_Rarely	670	Research & Development	26	1	Life Sciences	1	16	1	Male
15	34	No	Travel_Rarely	1346	Research & Development	19	2	Medical	1	18	2	Male
16	28	Yes	Travel_Rarely	103	Research & Development	24	3	Life Sciences	1	19	3	Male
17	29	No	Travel_Rarely	1389	Research & Development	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** (the file name may appear different in the screen capture below).

Notice that some of the columns appear as string and they should be integers. This is mostly true

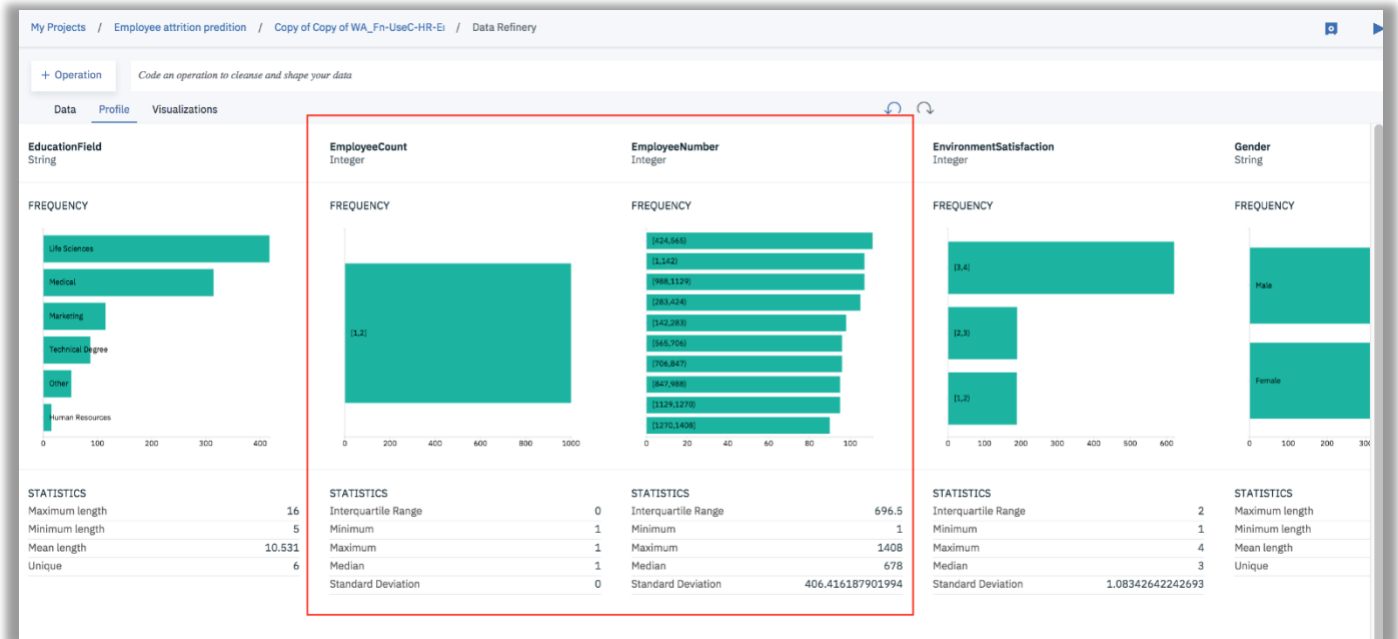



if you uploaded a csv file.

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

+ Operation <i>Code an operation to cleanse and shape your data</i>				
	Data	Profile	Visualizations	
	Age String	Attrition String	BusinessTravel String	DistanceInMiles String
1	41	Remove	Travel_Rarely	11
2	49	Remove duplicates	Travel_Frequently	27
3	37	Remove empty rows	Travel_Rarely	13
4	33	Sort ascending	Travel_Frequently	13
5	27	Sort descending	Travel_Rarely	59
6	32	Substitute	Travel_Frequently	10
7	59		Travel_Rarely	13
8	30	CONVERT COLUMN... >	Boolean	13
9	38			21
10	36	TEXT >	Date	12
11	35	View All	Decimal	80
12	29	No	Integer	19
13	31	No	String	67
14	34	No	Timestamp	13
15	28	Yes		10
16	29	No	Travel_Rarely	13

8. 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.



9. Go back to the **Data** tab.
10. **Remove** those columns.
11. Save the data flow .
12. From the Run Job icon, select the option and change both file name and flow name.

Working with AutoAI

You are now ready to use machine learning models against your data to see which model 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 **AutoAI experiment**
3. Define a model name. For example: **ML models for employee attrition**
4. Click **Associate a Machine Learning service instance**.

IBM Watson Studio

Create an AutoAI experiment

Define AutoAI experiment details

Create AutoAI experiment type

☒ From blank ☐ From sample

Asset name *

Employee Attrition using CSV

Description

Description of AutoAI experiment

Associated services

Machine Learning Service

No Machine Learning service instances associated with your project.

[Associate a Machine Learning service instance](#) with your project on the project settings page, then click the reload button below to refresh the instances available for association with your new model builder instance.

[Reload](#)

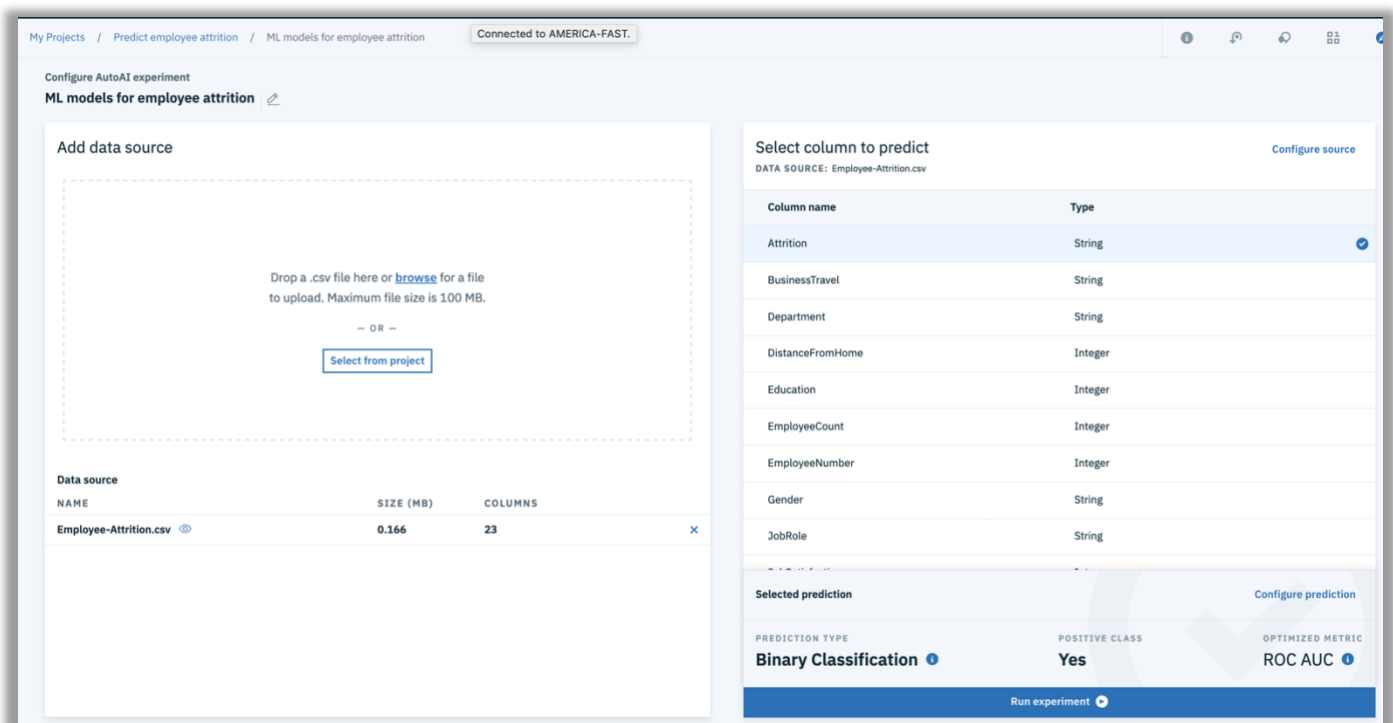
Compute configuration * ⓘ

8 vCPU and 32 GB RAM

This compute configuration consumes 20 capacity units per hour. [Learn more](#) about capacity unit hours and Watson Machine Learning pricing plans.

Cancel Create

5. Click **Create** and click **Confirm**.
6. Select an existing ML service and click **Select**.
7. Click **Reload**. Notice, the service instance name appears then click **Create**.
8. Click **Select from project**.
9. Select your most recent data asset (the newly saved one) and click **Select asset**.
10. Select the **Attrition** column name. Because this is the feature that you are basing your prediction on.
11. Click **Run Experiment**.



This will take a few minutes. Take your time and observe the process and read the following passages. A progress infographic shows you the creation of pipelines for your data. The duration of this phase depends on the size of your data set. A notification message informs you if the processing time will be brief or require more time. You can work in other parts of the product while the pipelines build.

During AutoAI training, your data set is split to a training part and a hold-out part. The training part is used by the AutoAI training stages to generate the AutoAI model pipelines and cross-validation scores used to rank them. After AutoAI training, the hold-out part is used for the resulting pipeline model evaluation and computation of performance information such as ROC curves and confusion matrices, shown in the leaderboard. The training/hold-out split ratio is 90/10.

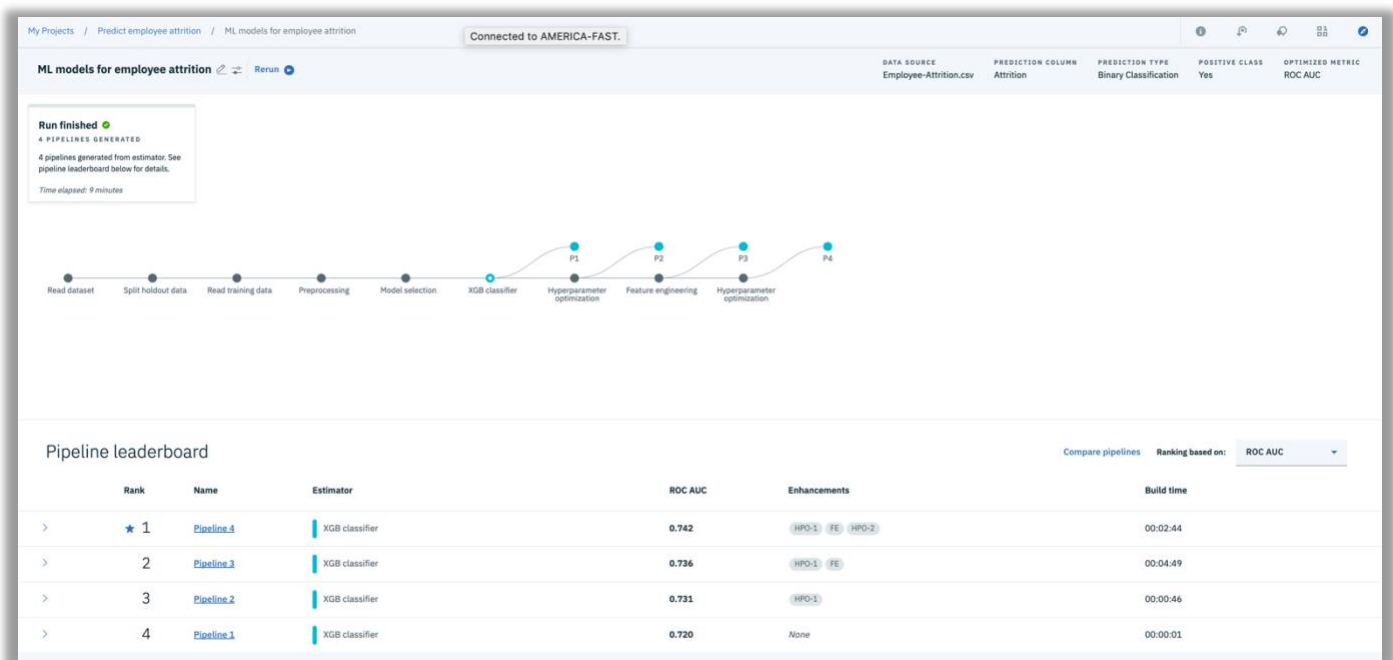
A **receiver operating characteristic curve**, or **ROC curve**, is a graphical plot that illustrates the diagnostic ability of a binary classification system as its discrimination threshold is varied.

The ROC curve is created by plotting the [true positive rate](#) (TPR) against the [false positive rate](#) (FPR) at various threshold settings.

This link and the video within, helps to better understand the details of calculating ROC and the **Area Under the Curve (AUC)**: <https://www.dataschool.io/roc-curves-and-auc-explained/>

As the training progresses, you are presented with a dynamic tree infographic and leaderboard. The tree infographic shows the sequences of AutoAI training stages (pre-processing, model selection, feature engineering, and HPO) that create the resulting model pipelines, which are shown as leaves of the tree. The leaderboard contains model pipelines ranked by cross-validation scores.

Hovering over each pipeline name on a leaf of the tree infographic displays the pipeline structure. Each AutoAI model pipeline is defined by a sequence of data transformations that transform the initial data set, ending with an estimator algorithm that generates predictions. The sequence of data transformations consists of a pre-processing transformer and a sequence of data transformers, if feature engineering was performed for this pipeline. The estimator is determined by model selection and HPO steps during AutoAI training.



So what does all this mean? Let's take a step back.

You selected a column that had values of Yes and No; that's a binary classification. If you had a Maybe in there, then the selected method would be multi-class classification.

Each model pipeline is scored for a variety of metrics and then ranked. The default ranking metric for binary classification models is the area under the ROC curve, for multi-class classification models is accuracy, and for regression models is the root mean-squared error (RMSE). The highest-ranked pipelines are displayed in a leaderboard, so you can view more information about them. The leaderboard also provides the option to save select model pipelines after reviewing them.

The AutoAI process follows this sequence to build candidate pipelines:

- Data pre-processing
- Automated model selection
- Automated feature engineering
- Hyperparameter optimization

A. Data pre-processing

Most data sets contain different data formats and missing values, but standard machine learning algorithms work with numbers and no missing values. AutoAI applies various algorithms to analyze, clean, and prepare your raw data for machine learning. It automatically detects and categorizes features based on data type, such as categorical or numerical. Depending on the categorization, it uses hyperparameter optimization to determine the best combination of strategies for missing value imputation, feature encoding, and feature scaling for your data.

B. Automated model selection

The next step is automated model selection that matches your data. AutoAI uses a novel approach that enables testing and ranking candidate estimators against small subsets of the data, gradually increasing the size of the subset for the most promising estimators to arrive at the best match. This approach saves time without sacrificing performance. It enables ranking a large number of candidate estimators and selecting the best match for the data.

C. Automated feature engineering

Feature engineering attempts to transform the raw data into the combination of features that best represents the problem to achieve the most accurate prediction. AutoAI uses a novel approach that explores various feature construction choices in a structured, non-exhaustive manner, while progressively maximizing model accuracy using reinforcement learning. This results in an optimized sequence of transformations for the data that best match the estimators of the model selection step.

D. Hyperparameter optimization

Finally, a hyper-parameter optimization step refines the best performing model pipelines. AutoAI uses a novel hyper-parameter optimization algorithm optimized for costly function evaluations such as model training and scoring that are typical in machine learning. This approach enables fast convergence to a good solution despite long evaluation times of each iteration.

What is XGBoost Classifier?

XGBoost is an open source library providing a high-performance implementation of gradient boosted decision trees. An underlying C++ codebase combined with a Python interface sitting on top make for an extremely powerful yet easy to implement package.

With a regular machine learning model, like a decision tree, we'd simply train a single model on our dataset and use that for prediction. We might play around with the parameters for a bit or augment the data, but at the end we are still using a single model. Even if we build an ensemble, all of the models are trained and applied to our data separately.

Boosting on the other hand takes a more *iterative* approach. It's still technically an ensemble technique in that many models are combined together to perform the final one, but takes a more clever approach.

Rather than training all of the models in isolation of one another, boosting trains models in succession, with each new model being trained to correct the errors made by the previous ones. Models are added sequentially until no further improvements can be made.

The advantage of this iterative approach is that the new models being added are focused on correcting the mistakes which were caused by other models. In a standard ensemble method where models are trained in isolation, all of the models might simply end up making the same mistakes!

Gradient Boosting specifically is an approach where new models are trained to predict the residuals (i.e errors) of prior models.

Evaluate the pipelines

- Click a pipeline in the leaderboard to view more detail about the metrics and performance.
- Click Compare to view how the top pipelines compare.
- Sort the leaderboard by a different metric.

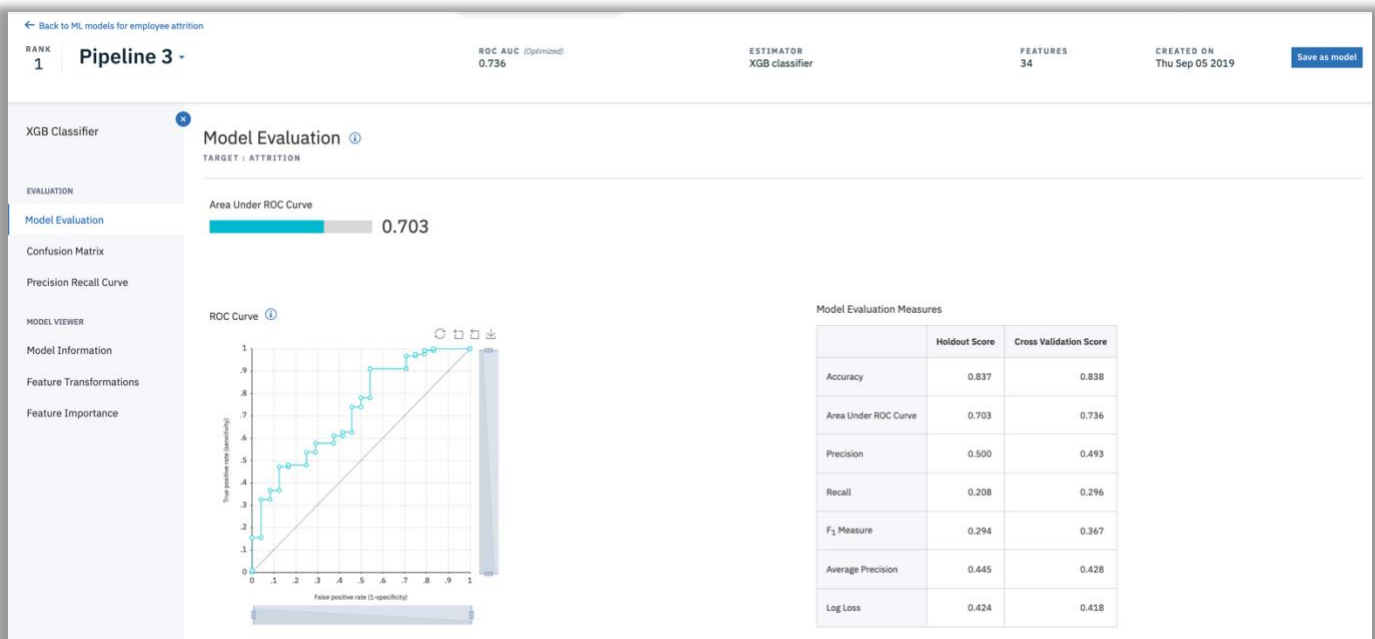
12. Change the ranking from ROC AUC (more on that later) to **Accuracy**.

Pipeline leaderboard

Compare pipelines Ranking based on: **Accuracy**

Rank	Name	Estimator	Accuracy	Enhancements	Build time
> ★ 1	Pipeline 3	XGB classifier	0.838	HPO-1 FE	00:04:49
> 2	Pipeline 1	XGB classifier	0.828	None	00:00:01
> 3	Pipeline 4	XGB classifier	0.825	HPO-1 FE HPO-2	00:02:44
> 4	Pipeline 2	XGB classifier	0.820	HPO-1	00:00:46

13. Click the pipeline with the star next to it. In this example, it is pipeline 1.



14. When you are satisfied with a pipeline, click **Save as model** to save the candidate as a model to your project so you can test and deploy it. A notification confirms that you saved the model to the space associated with the project. Click the space to configure, train, test, and deploy the model.

15. Once the model is saved, click **View in Project**.

16. Click the **Deployment** tab and click **Add Deployment**.

17. Give it a name and click **Save**.

18. Wait until the status is **Ready**.

19. Click the Deployment name.

20. Click the **Test** tab.

21. Use the spreadsheet to select the proper values.
 22. Experiment with a few parameters. For example, for *Age*, enter **22**, for *JobSatisfaction*, enter **1** and **single** for *MaritalStatus*.
 23. Click **Predict**.
 24. Enter a few other values of your choosing and observe the newly generated prediction values.
- Congratulations, you have just deployed a machine learning model that based on the column that you selected.



© 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
