PREDICT ATTRITION OF YOUR VALUABLE EMPLOYEES

(With the help of AUTOAI)

Business Problem

Attrition is a problem that impacts all businesses, irrespective of geography, industry and size of the company. Employee attrition leads to significant costs for a business, including the cost of business disruption, hiring new staff and training new staff.

This data set presents an employee survey from IBM, indicating if there is attrition or not. The data set contains “1470” entries. Given the limited size of the data set, the model should only be expected to provide modest improvement in identification of attrition vs a random allocation of probability of attrition.

The Data

#	Column	Non-Null Count	Dtype
0	Age	1470 non-null	int64
1	Attrition	1470 non-null	object
2	BusinessTravel	1470 non-null	object
3	DailyRate	1470 non-null	int64
4	Department	1470 non-null	object
5	DistanceFromHome	1470 non-null	int64
6	Education	1470 non-null	int64
7	EducationField	1470 non-null	object
8	EmployeeCount	1470 non-null	int64
9	EmployeeNumber	1470 non-null	int64
10	EnvironmentSatisfaction	1470 non-null	int64
11	Gender	1470 non-null	object
12	HourlyRate	1470 non-null	int64
13	JobInvolvement	1470 non-null	int64
14	JobLevel	1470 non-null	int64
15	JobRole	1470 non-null	object
16	JobSatisfaction	1470 non-null	int64
17	MaritalStatus	1470 non-null	object
18	MonthlyIncome	1470 non-null	int64
19	MonthlyRate	1470 non-null	int64
20	NumCompaniesWorked	1470 non-null	int64
21	Over18	1470 non-null	object
22	OverTime	1470 non-null	object
23	PercentSalaryHike	1470 non-null	int64
24	PerformanceRating	1470 non-null	int64
25	RelationshipSatisfaction	1470 non-null	int64
26	StandardHours	1470 non-null	int64
27	StockOptionLevel	1470 non-null	int64
28	TotalWorkingYears	1470 non-null	int64
29	TrainingTimesLastYear	1470 non-null	int64
30	WorkLifeBalance	1470 non-null	int64
31	YearsAtCompany	1470 non-null	int64
32	YearsInCurrentRole	1470 non-null	int64
33	YearsSinceLastPromotion	1470 non-null	int64
34	YearsWithCurrManager	1470 non-null	int64

The Solution - Methodology

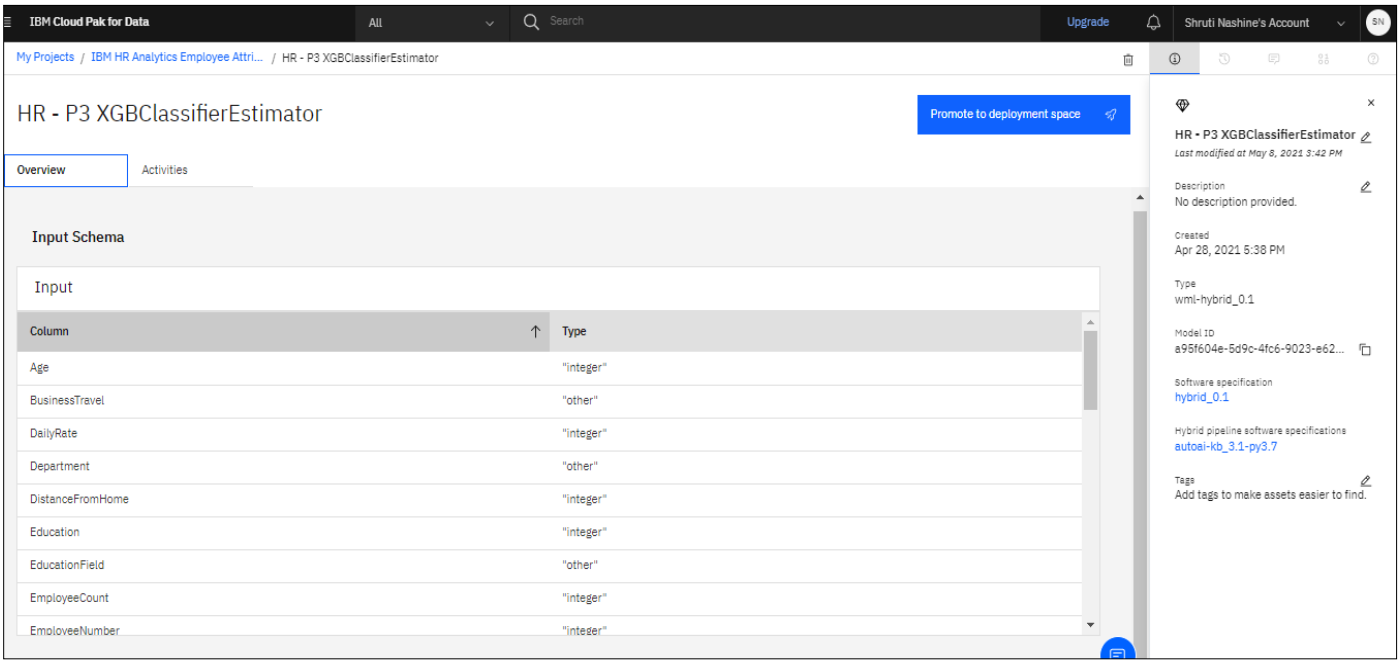
We will then test different parameters and probability threshold using confusion Matrixes, Area under the Curve and Decision matrix to determine which of the three models are the best.

The Process for Classification

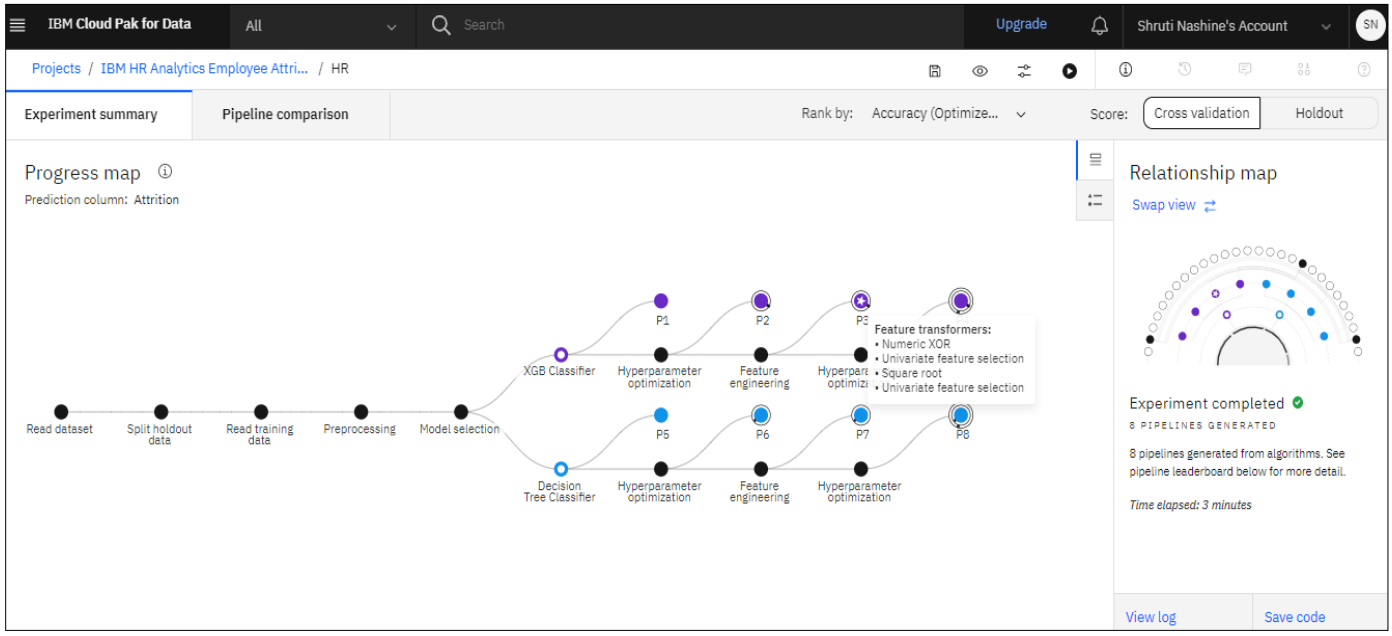
1. Create an estimation sample and two validation samples by splitting the data into three groups.
2. Set up the dependent variable, employee attrition as a categorical 0-1 variable.
3. Estimate the classification model using the estimation data, and interpret the results.
4. Assess the accuracy of classification in the first validation sample, possibly repeating steps 2-5 a few times changing the classifier in different ways to increase performance.
5. Finally, assess the accuracy of classification in the second validation sample. You should eventually use and report all relevant performance measures and plots on this second validation sample only.

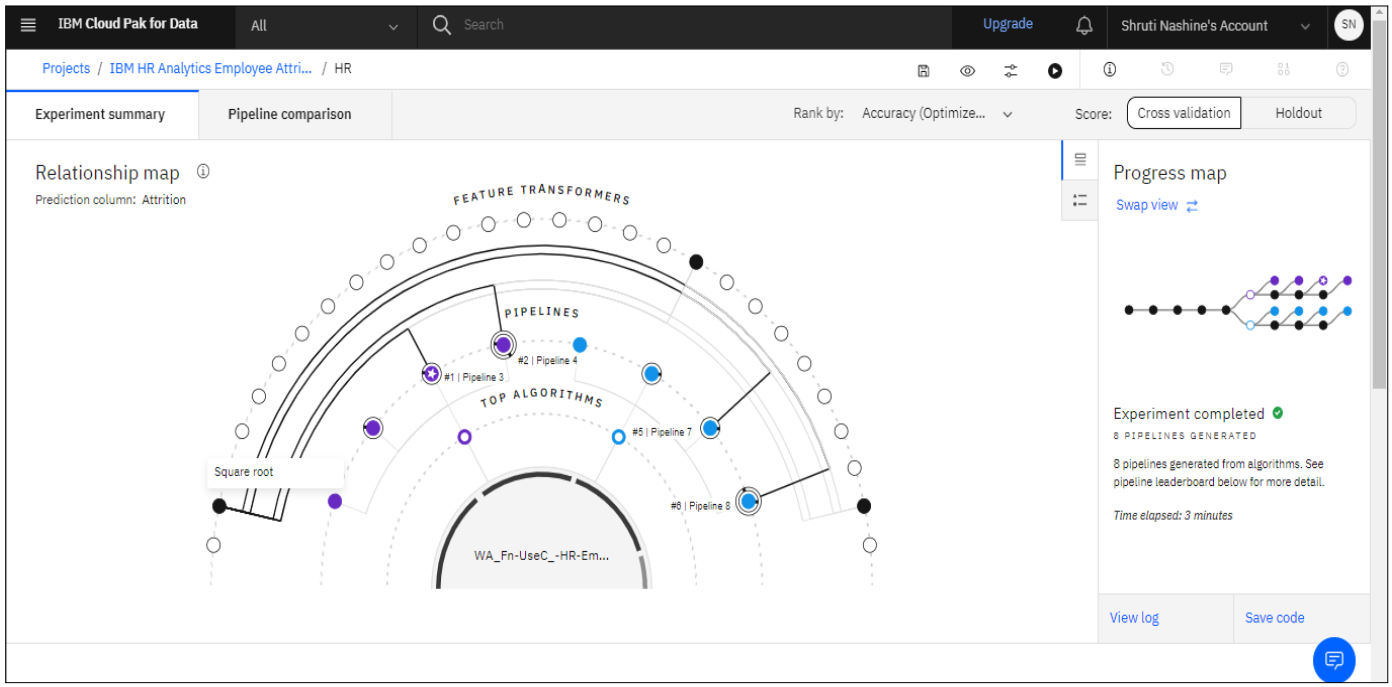
Step 1: Set up the dependent & Independent variable

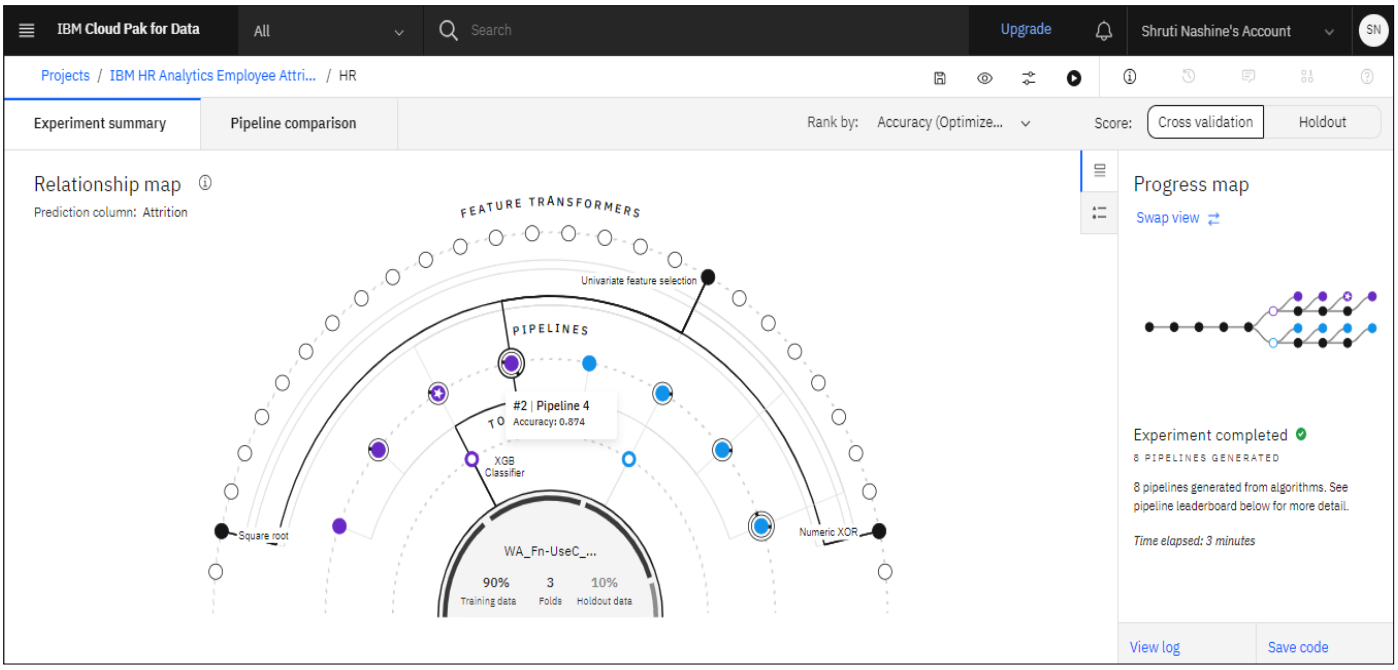
Only Attrition will be the dependent variable rest all of them are independent variable.



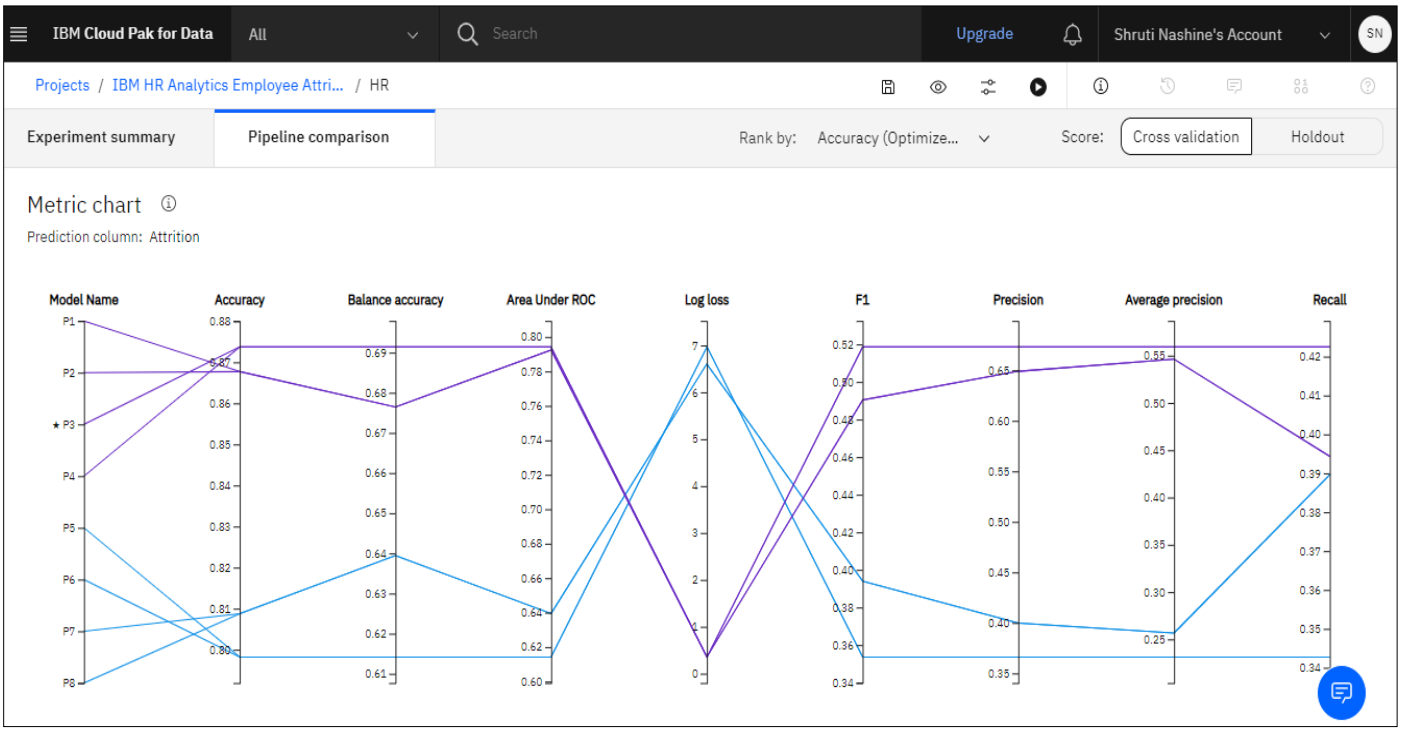
Step 2: Simple Analysis





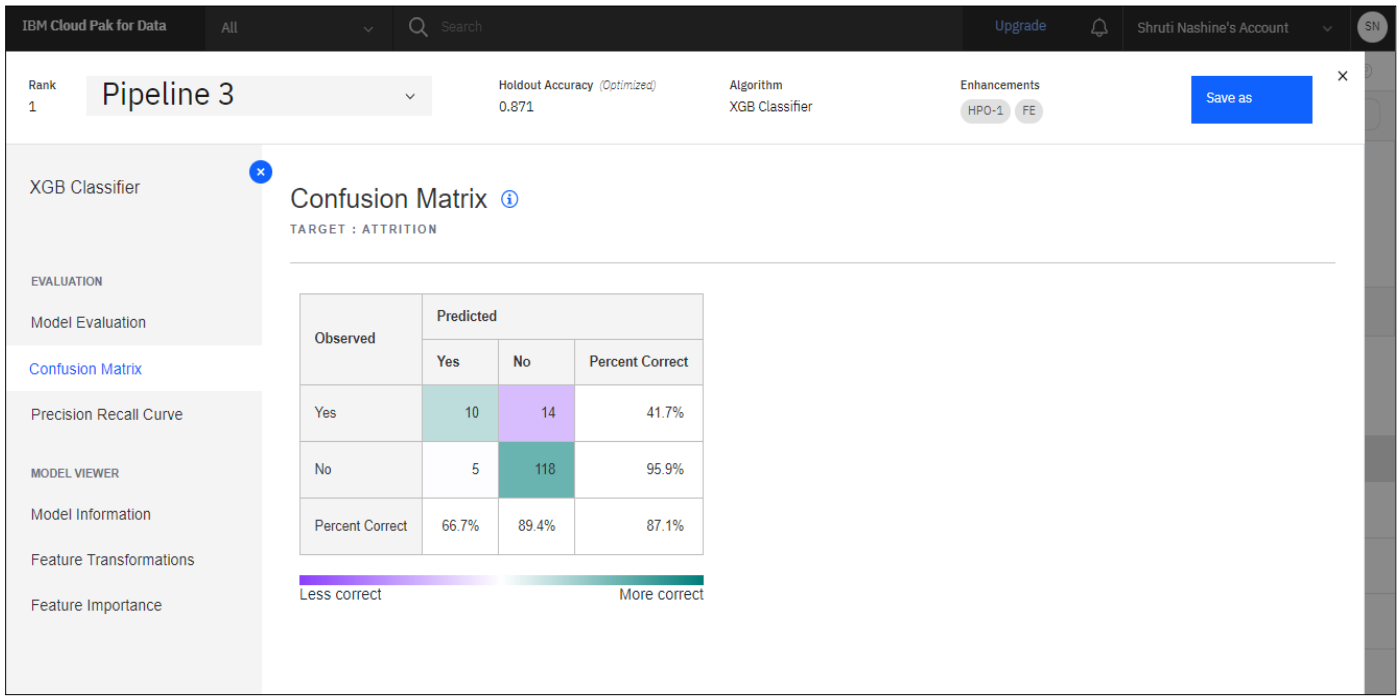


Step 3: Comparison & Interpretation

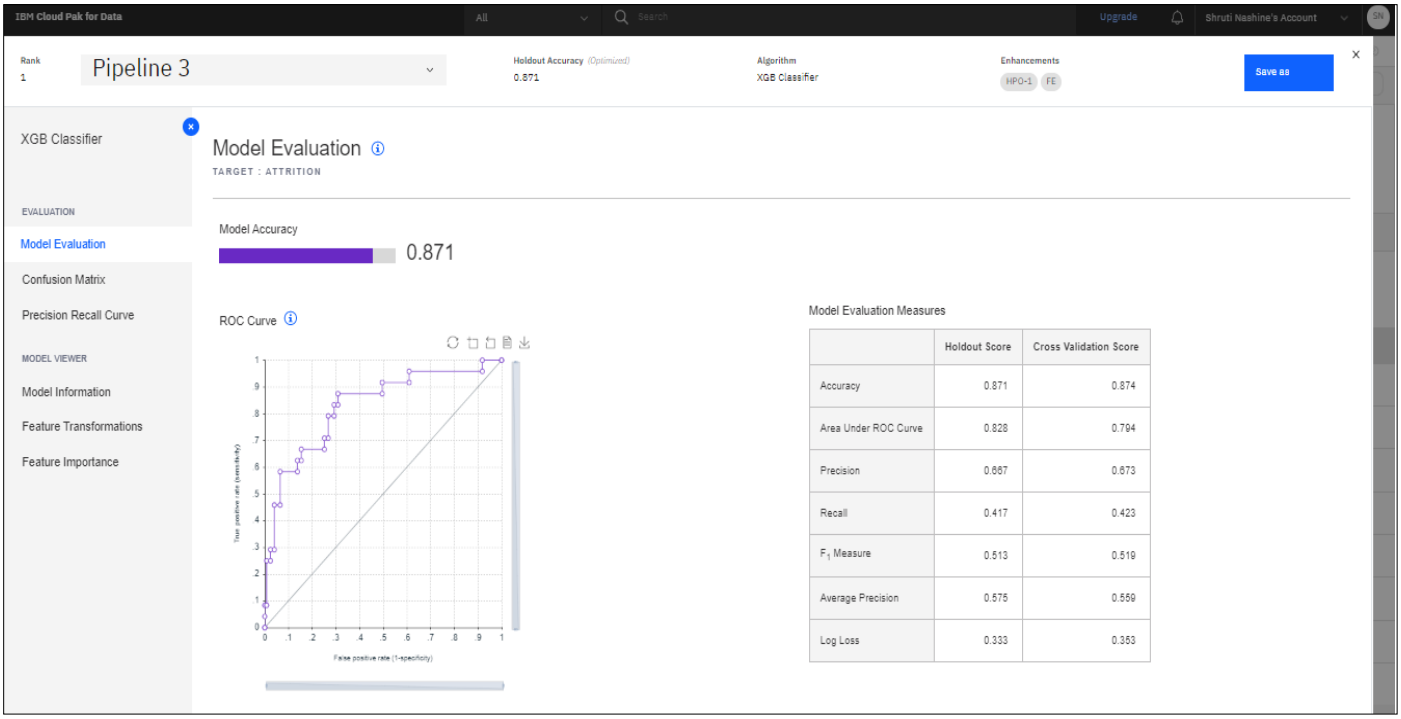


Step 4: Validation accuracy

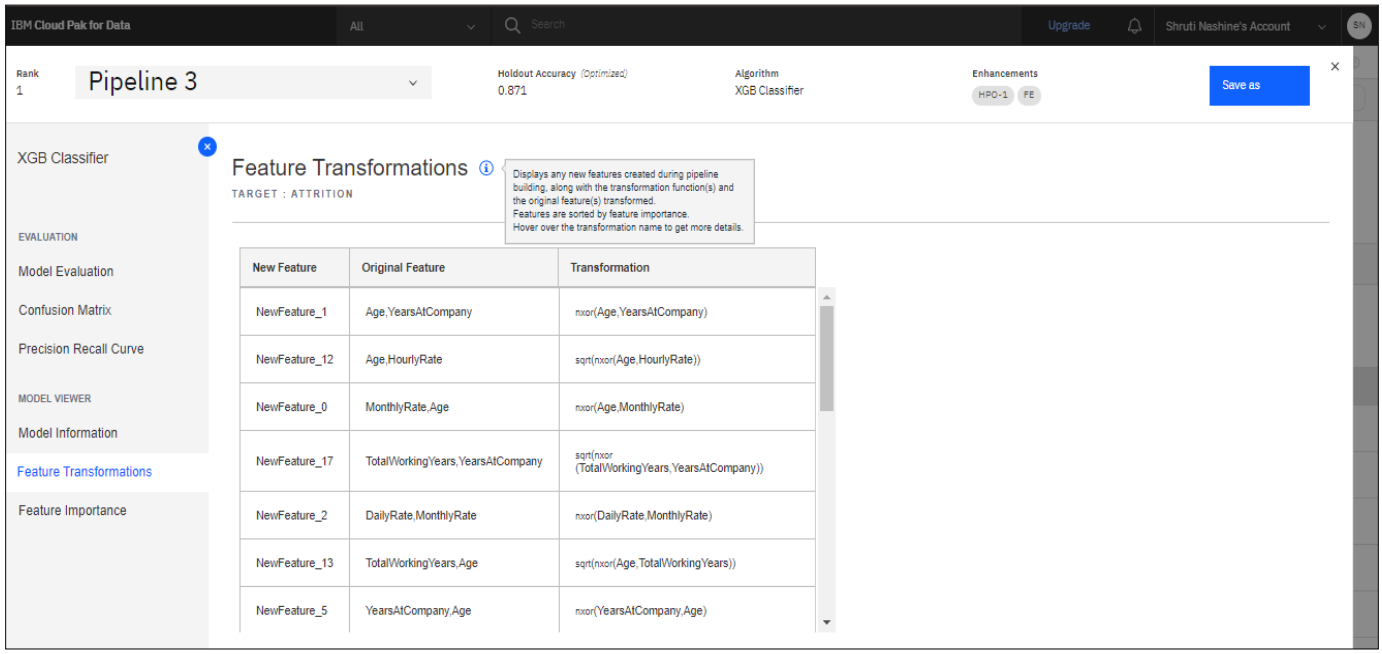
1. Confusion matrix



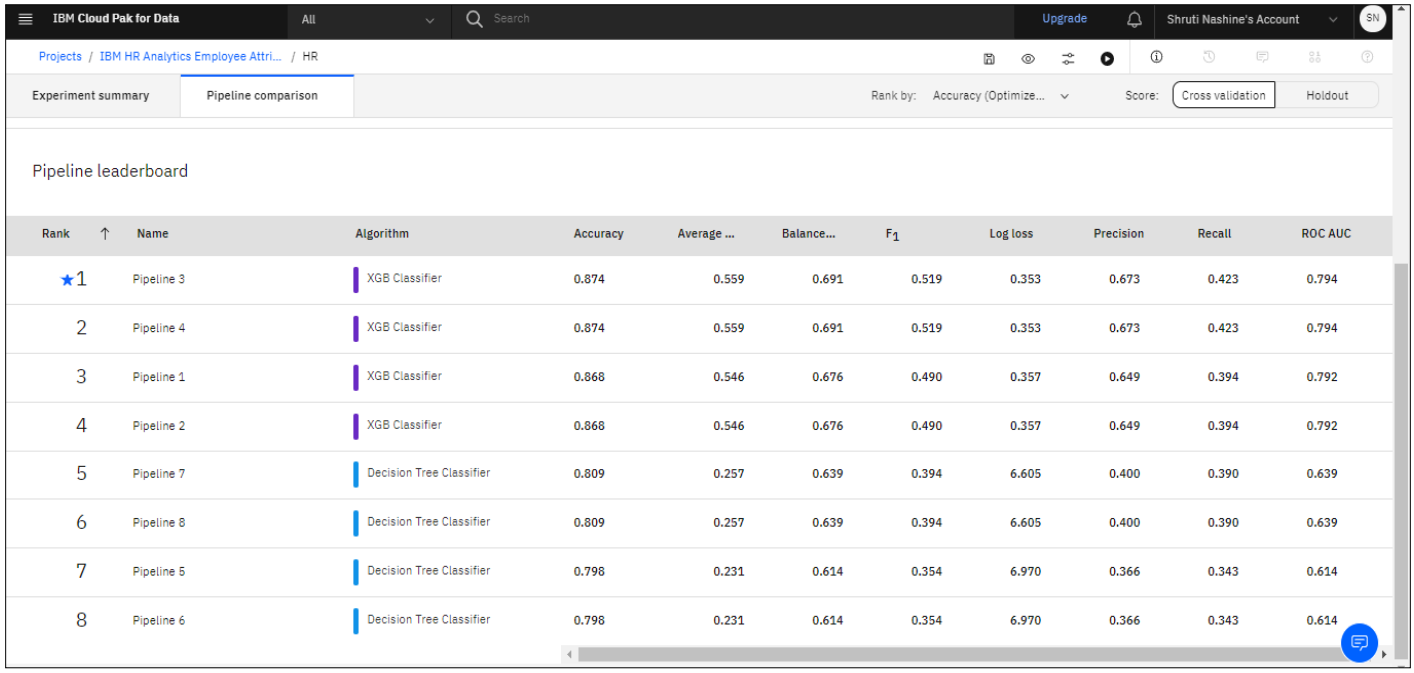
2. ROC curve



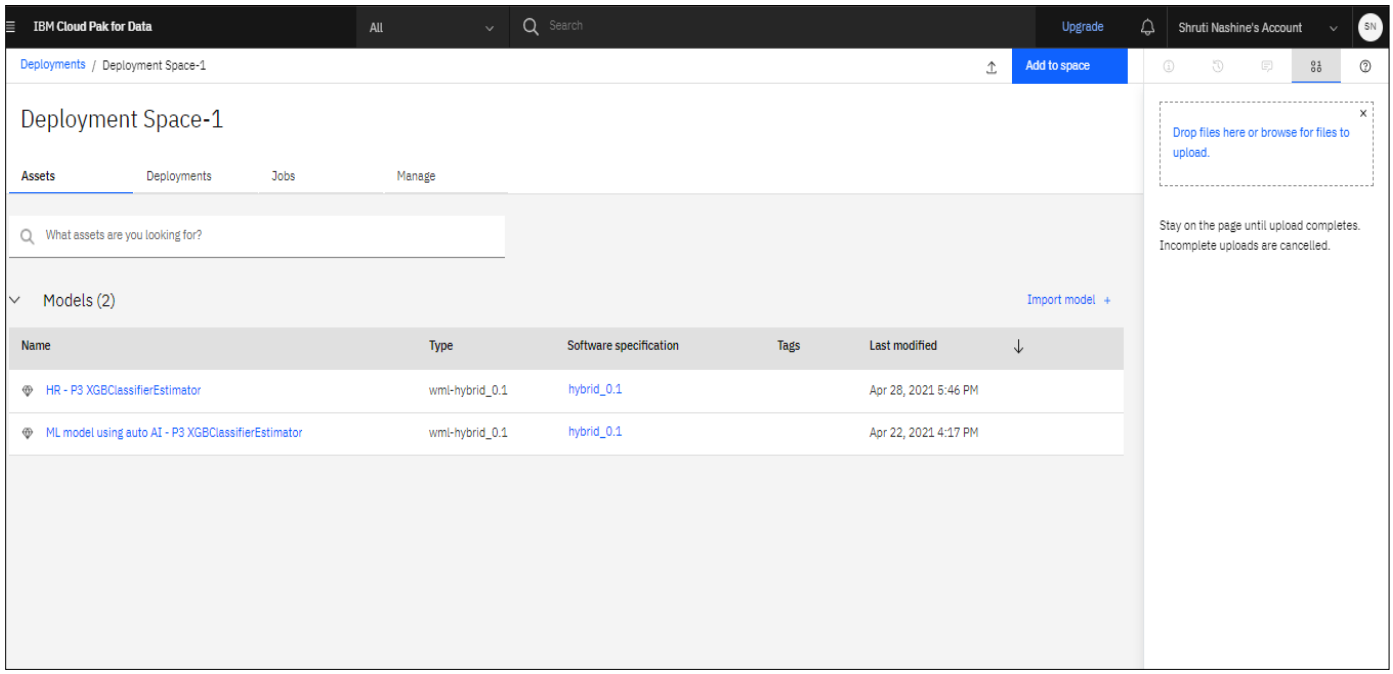
3. Feature Transformation

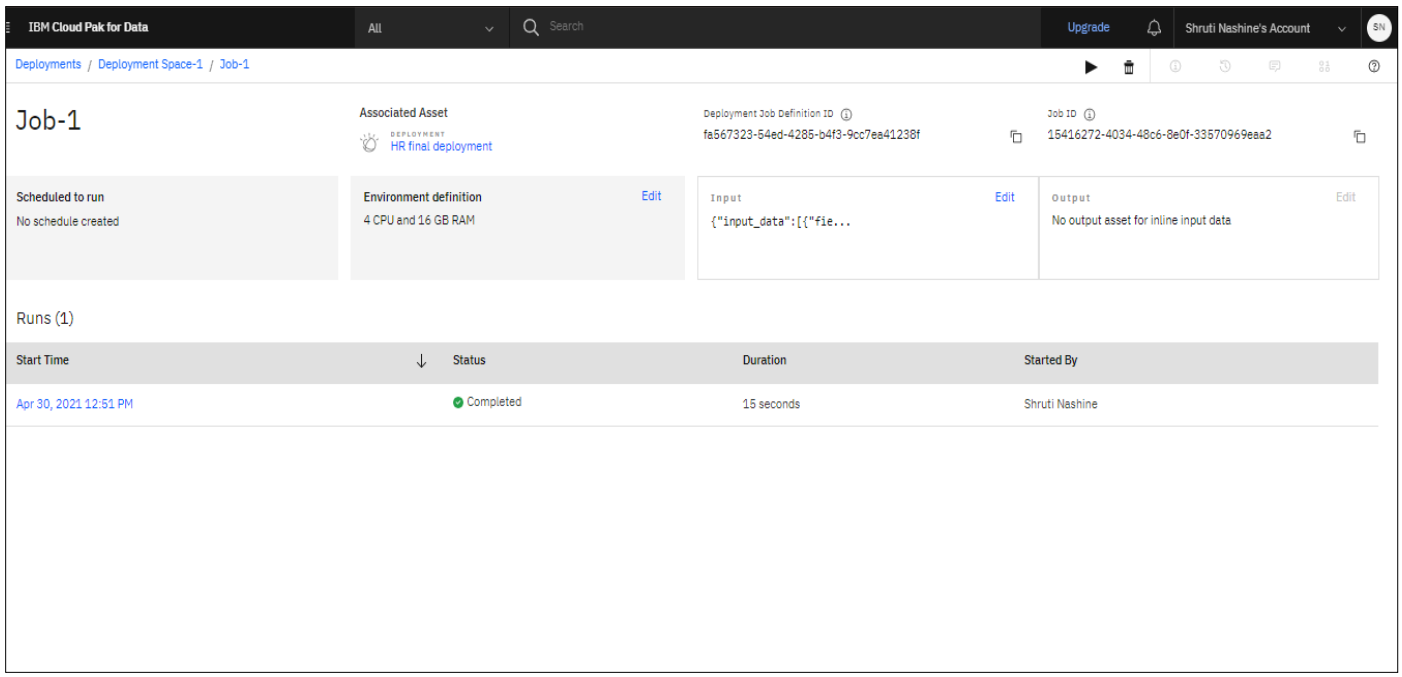


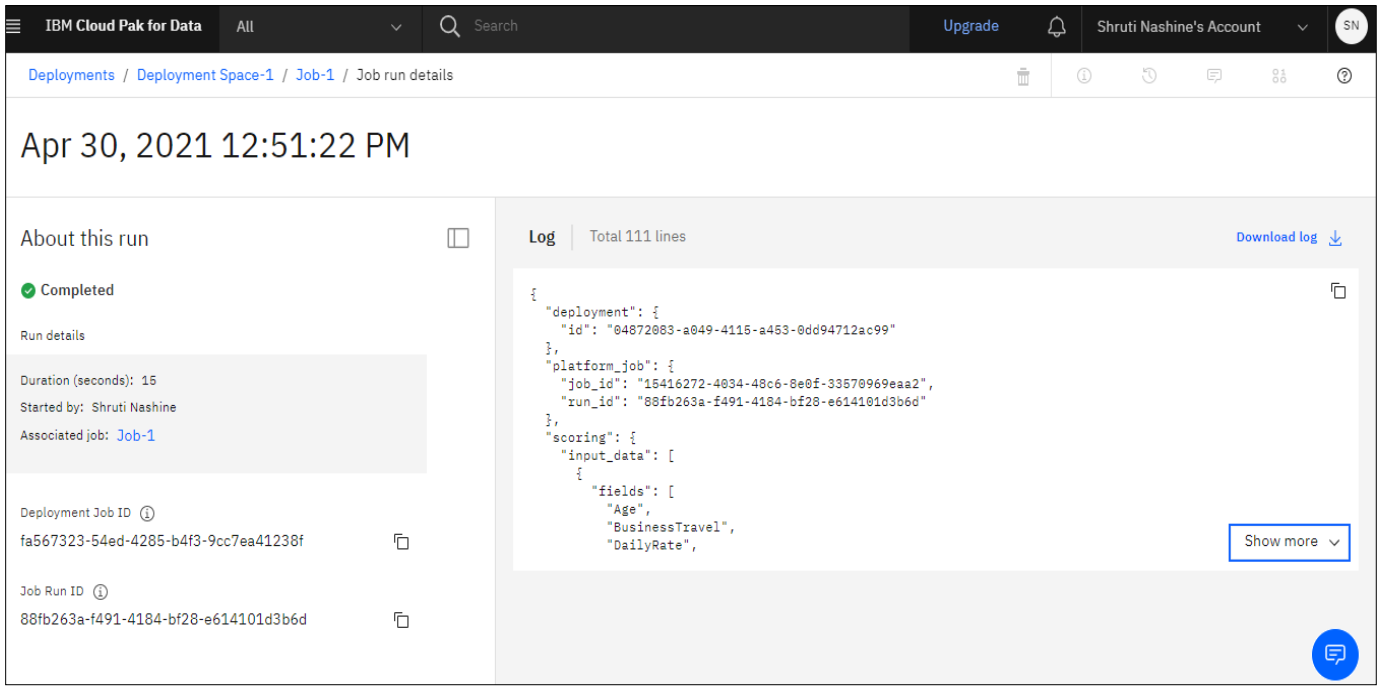
Step 5: Test Accuracy

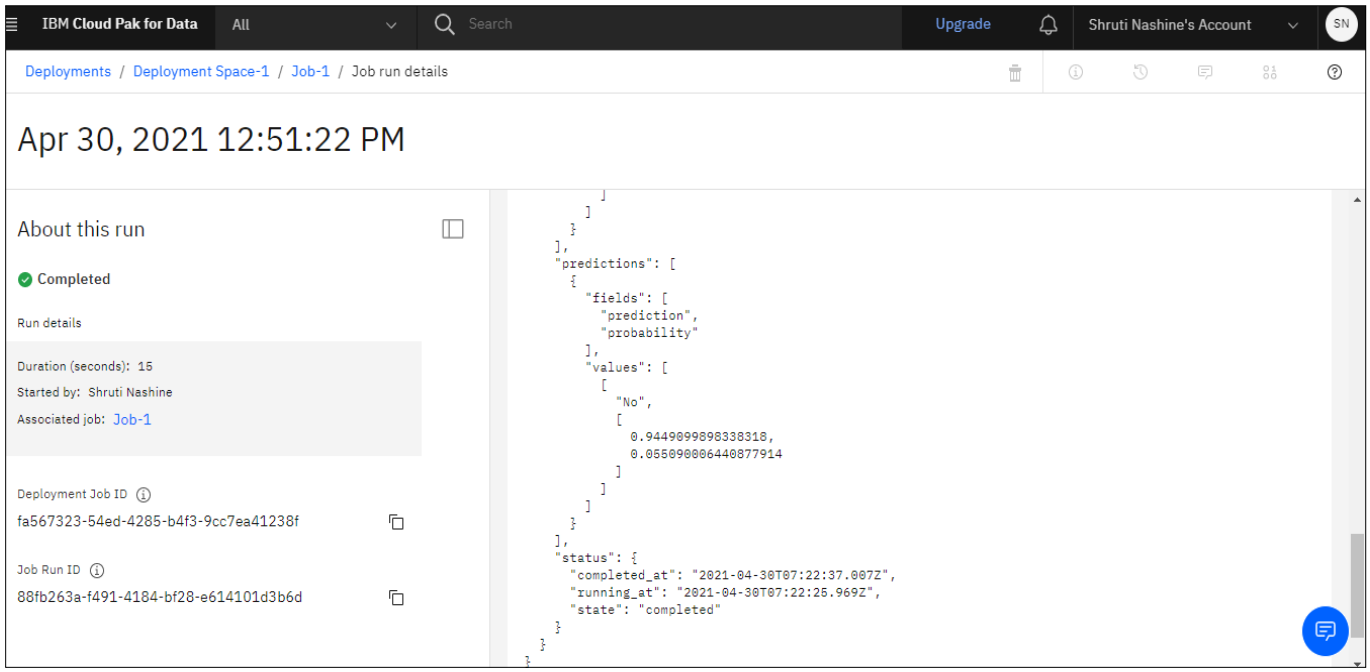


Step 6: Job Deployment






IBM Cloud Pak for Data
All
Search
Upgrade
Shruti Nashine's Account
SN
Deployments / Deployment Space-1 / Job-1 / Job run details
Apr 30, 2021 12:51:22 PM
About this run
Completed
Run details
Duration (seconds): 15
Started by: Shruti Nashine
Associated job: Job-1
Deployment Job ID
fa567323-54ed-4285-b4f3-9cc7ea41238f
Job Run ID
88fb263a-f491-4184-bf28-e614101d3b6d
Log
Total 111 lines
Download log
{
"deployment": {
"id": "04872083-a049-4115-a453-0dd94712ac99"
},
"platform_job": {
"job_id": "15416272-4034-48c6-8e0f-33570969eaa2",
"run_id": "88fb263a-f491-4184-bf28-e614101d3b6d"
},
"scoring": {
"input_data": [
{
"fields": [
"Age",
"BusinessTravel",
"DailyRate",
Show more


IBM Cloud Pak for Data
All
Search
Upgrade
Shruti Nashine's Account
SN
Deployments / Deployment Space-1 / Job-1 / Job run details
Apr 30, 2021 12:51:22 PM
About this run
Completed
Run details
Duration (seconds): 15
Started by: Shruti Nashine
Associated job: Job-1
Deployment Job ID
fa567323-54ed-4285-b4f3-9cc7ea41238f
Job Run ID
88fb263a-f491-4184-bf28-e614101d3b6d
]
]
}
],
"predictions": [
{
"fields": [
"prediction",
"probability"
],
"values": [
[
"No",
[
0.9449099898338318,
0.05509000644087914
]
]
]
}
],
"status": {
"completed_at": "2021-04-30T07:22:37.007Z",
"running_at": "2021-04-30T07:22:25.969Z",
"state": "completed"
}
}
}

Step 7: Data Analysis

After we ran the model multiple times and iterate to find the best value, we came with some conclusions:

- XGB Classifier is the best model, as it always predicts a higher area under the curve and a better confusion matrix.
- Model is biased towards predicting non attrition.
- Pipeline 3 found as best with accuracy as 0.874, ROC AUC as 0.794, precision as 0.673 and recall as 0.423
- Model Accuracy was found as 87.1%
- Deployment job was completed and found that prediction was 0.944 and probability of attrition was 0.055 i.e. very less attrition.