AutoAI walkthrough using Historical Wildfires combined with Historical Weather data

Useful Links :
The Contest landing page http://ibm.biz/cfcsc-wildfires
The Contest GitHub https://github.com/Call-for-Code/Spot-Challenge-Wildfires

NOTE: Only the NSW – New South Wales Region from the Historical Wildfires data and the Temperature from the Historical Weather data have been combined for this walkthrough

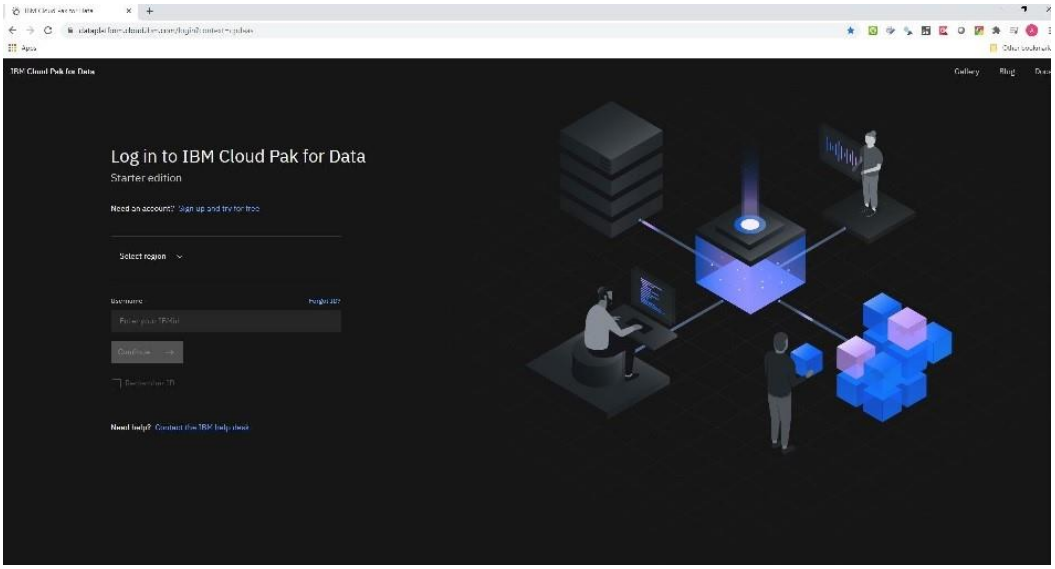
Please see the following datasets, accompanying readme and slides, which are available via GitHub https://github.com/Call-for-Code/Spot-Challenge-Wildfires together with a starter notebook.

The short dataset was created by reading the Historical Wildfires data into a Pandas dataframe, only the NSW region was selected from the dataframe, the Date column was converted to a datetime format. Next, the Historical Weather data was read into another Pandas dataframe, the NSW region and the Temperature parameter were selected from the dataframe, the Date column in this dataframe was also converted to datetime format.

These two selected dataframes were then merged using the NSW Historical Wildfires dataframe as the anchor, since the wildfires measurements were not available for all corresponding days in the weather data.

NOTE: This raw merged data was used in this AutoAI experiment walkthrough without any further processing or wrangling.

Log in to: https://dataplatform.cloud.ibm.com/login?context=cpdaas



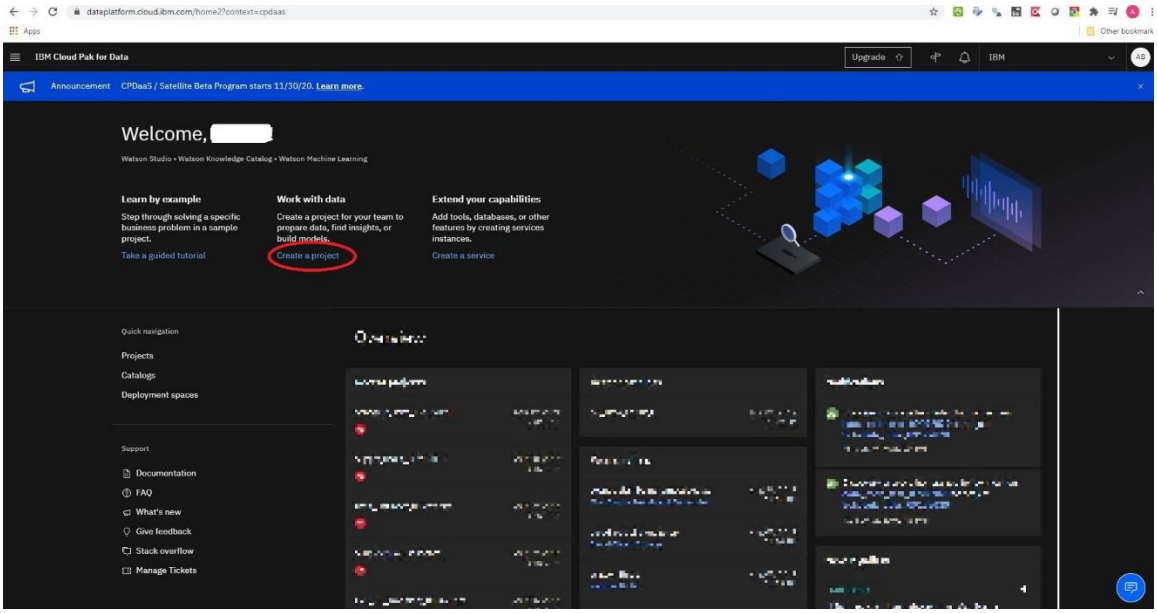
Select Region, closest to your location (Options: Dallas, London, Frankfurt, Tokyo)

If previously registered, provide username, then hit “Continue” and follow the instructions.

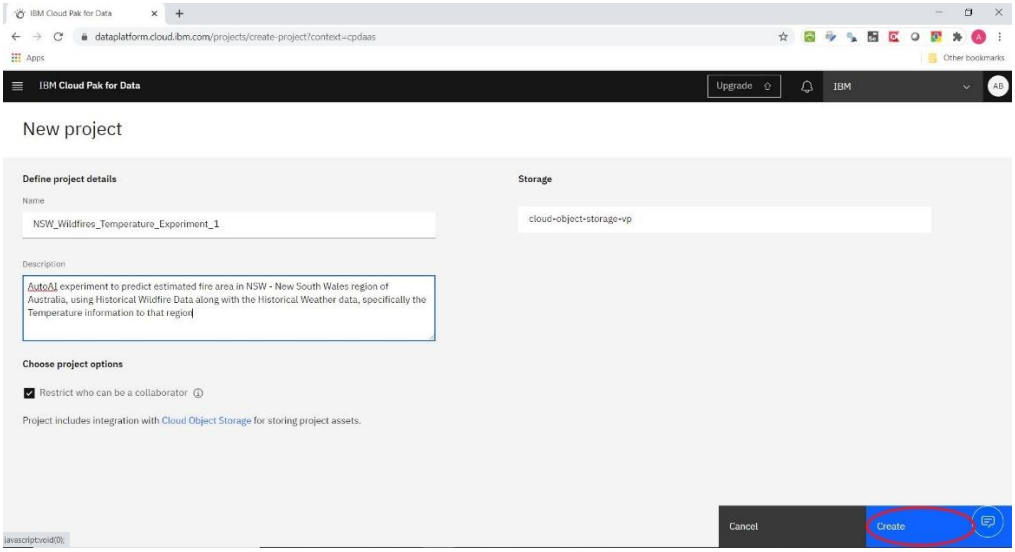
Else please use the link Sign up and try for free as shown in the screen above and follow the steps to register.

If you have used the IBM Cloud Pack for Data previously, you will be presented with a welcome screen as shown below, listing past projects, Notifications etc..

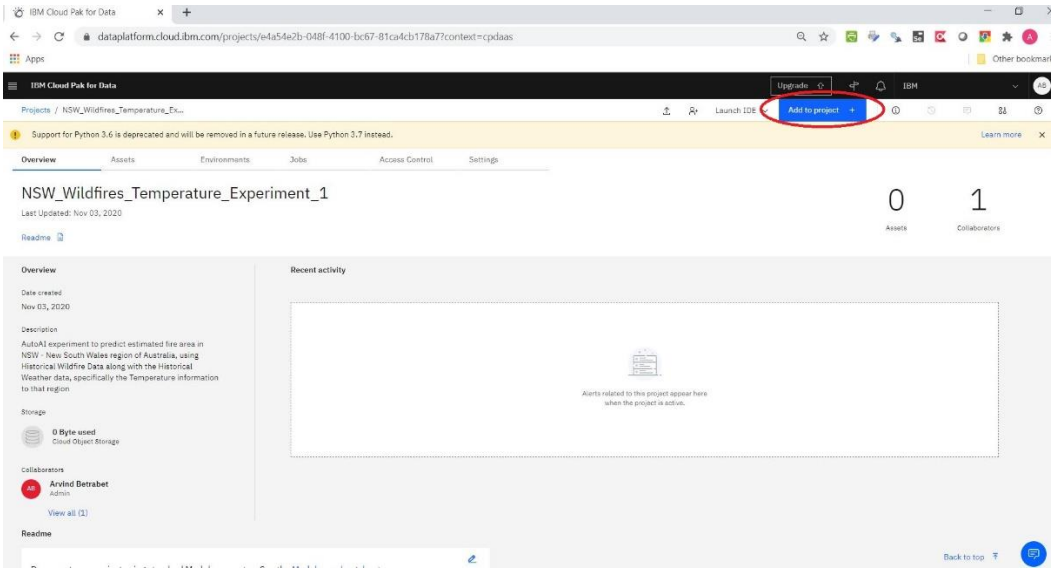
Click the “Create a project” link, as shown below, to create a new project.



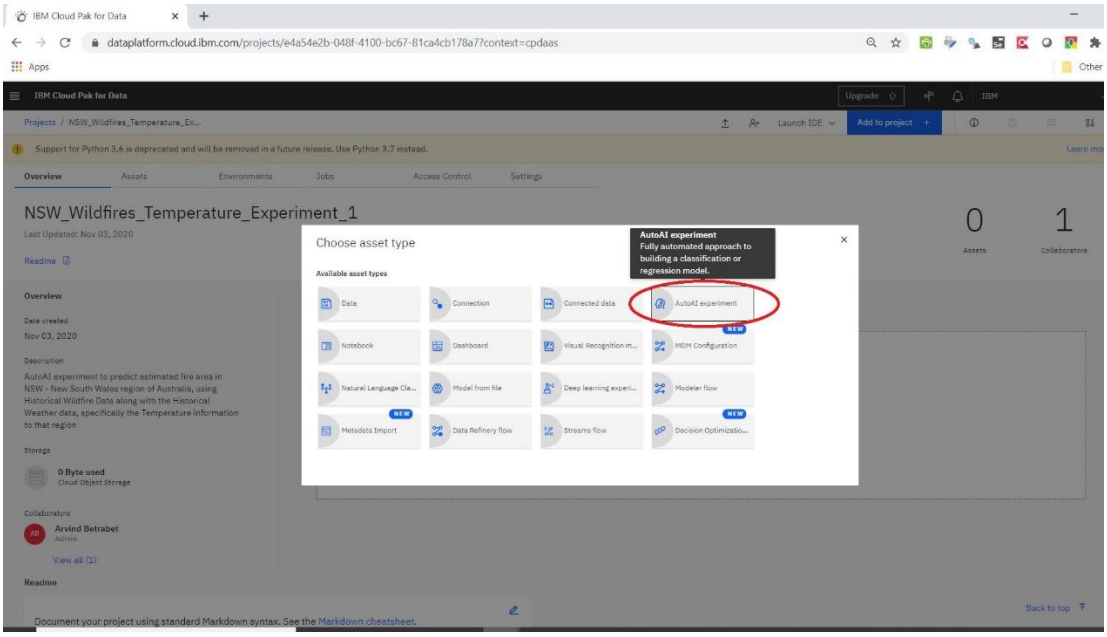
Fill in the appropriate information for the project (Name and Description), as shown below and click “Create”



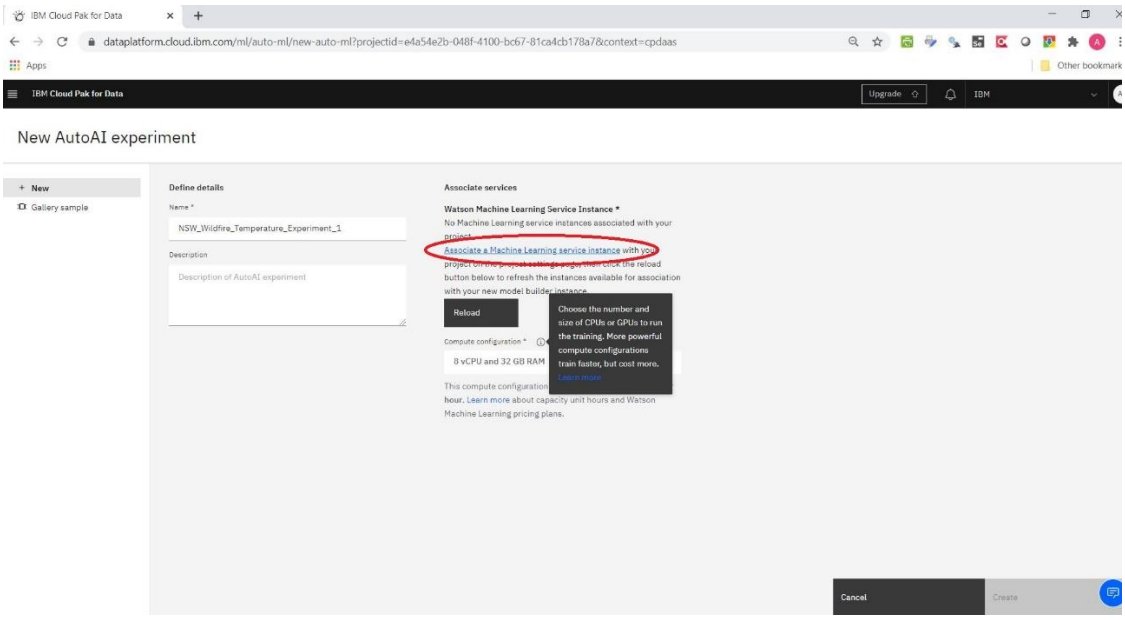
Takes a few seconds to create an environment.



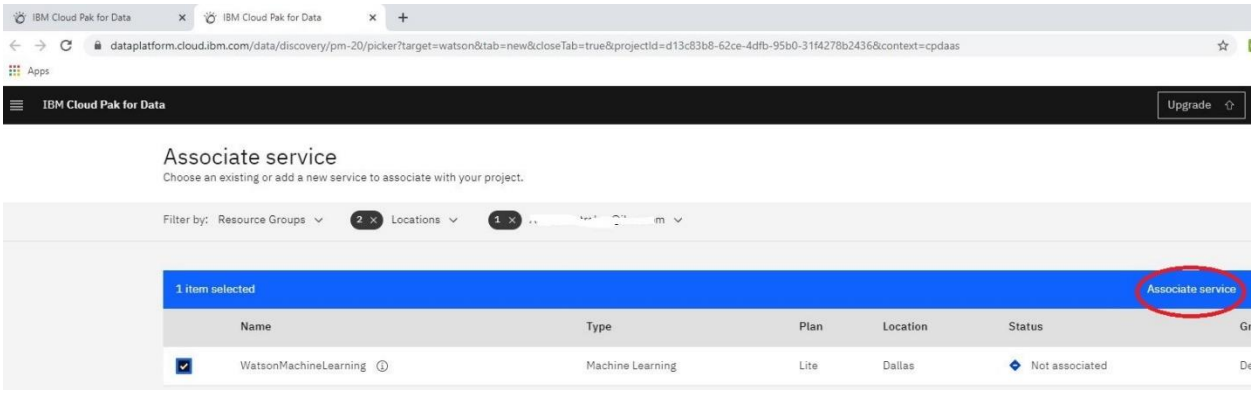
Click on Add to project on the top right, then select “AutoAI experiment” as shown:



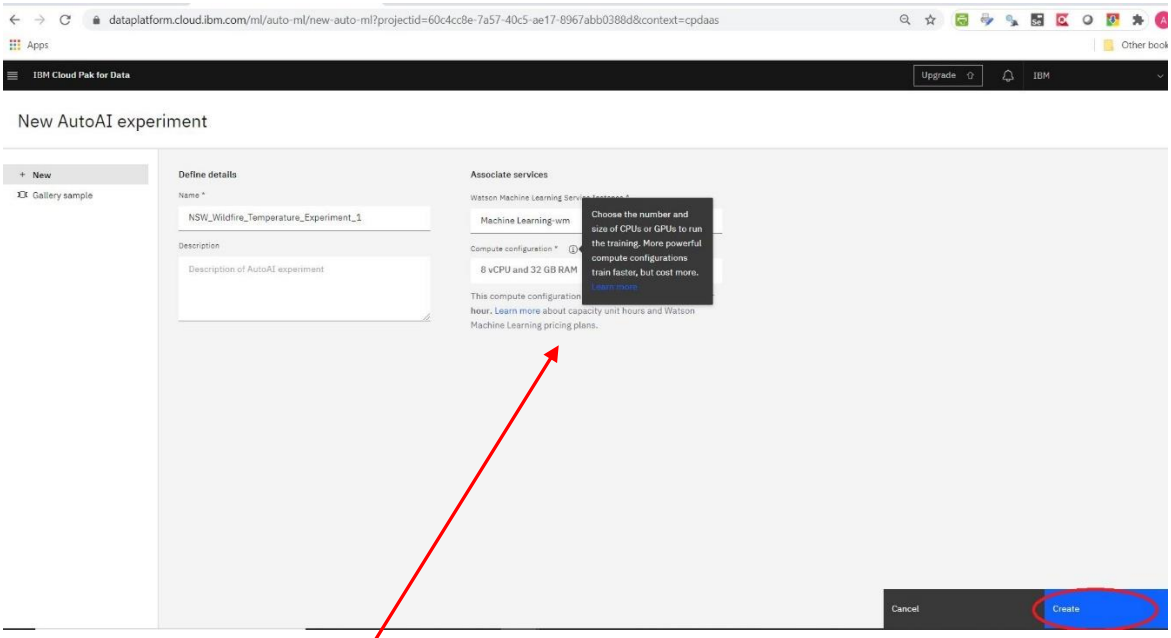
In the next screen, provide a name for the experiment, as well as read the information icon on the screen, as shown

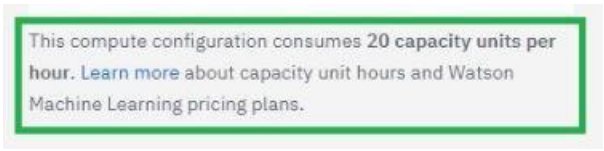


Click on “Associate a Machine Learning service instance” link to select the resource required for your experiment (a new tab is opened). Select the “WatsonMachineLearning” to associate a service, click on the “Associate Service” Button.



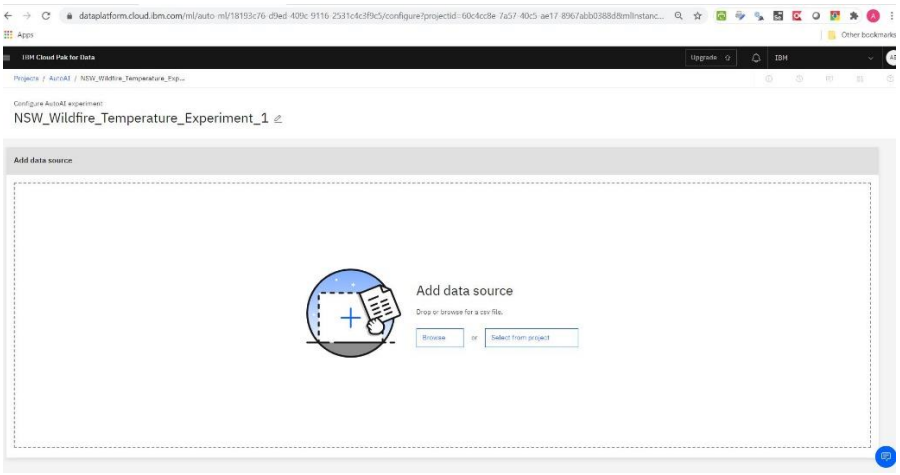
That window will close and take you back to the previous window, Click on “Reload” button to complete the association. Next click the “Create” button.



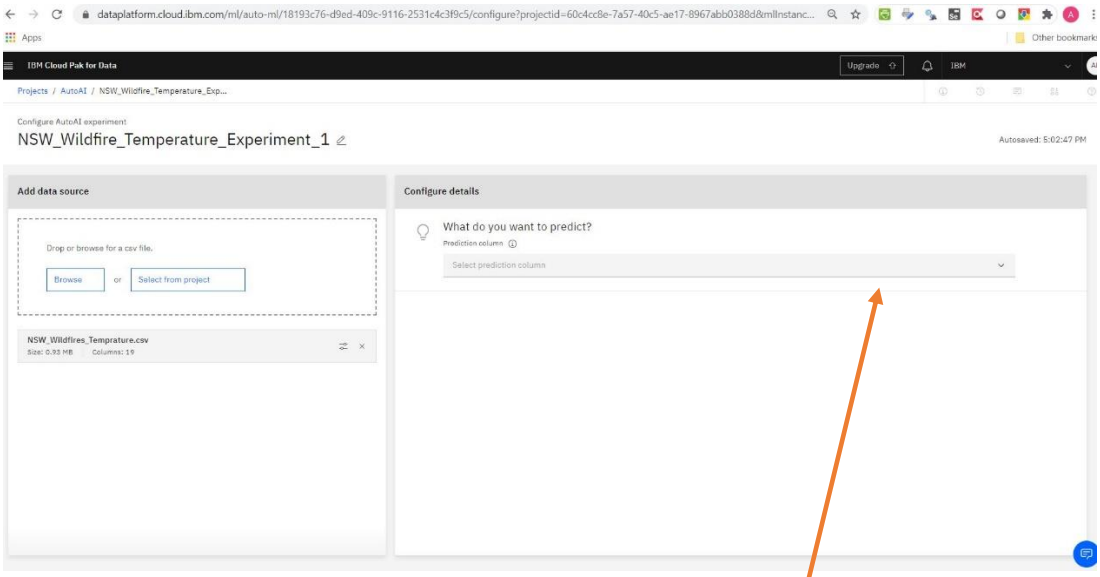


Please NOTE: the green box above informs how many CUH – Capacity Units per Hour will be consumed by this service/instance that has 8-Virtual CPUs and 32GB of RAM

Now add the data source

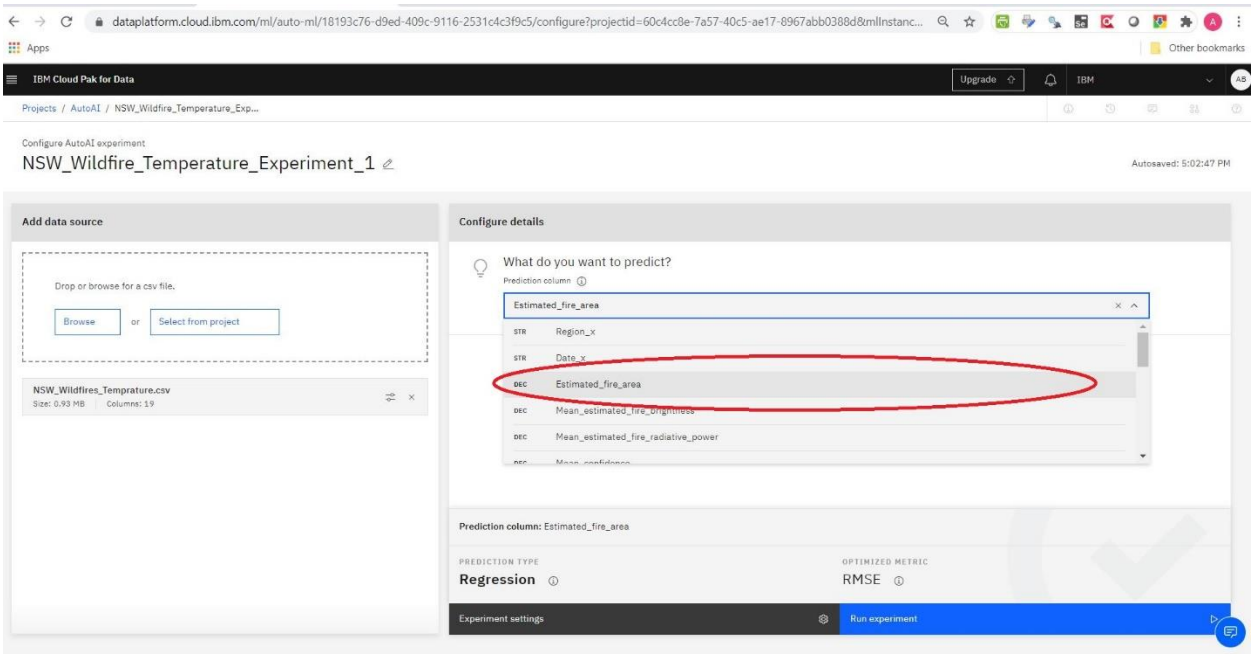


Browse to where the data is located, here the file is named NSW_Wildfires_Temprature.csv, once that is located and brought into the experiment it will show up as follows:



Notice the file name, the size and the number of columns.

Next, the drop down on the right allows you to select the column to predict. Notice that AutoAI engine has already determined the data types for each of the columns. Select “Estimated_fire_area” as shown.



Notice the AutoAI engine has already determined the prediction type as well as suggested the optimization metric to be RMSE.

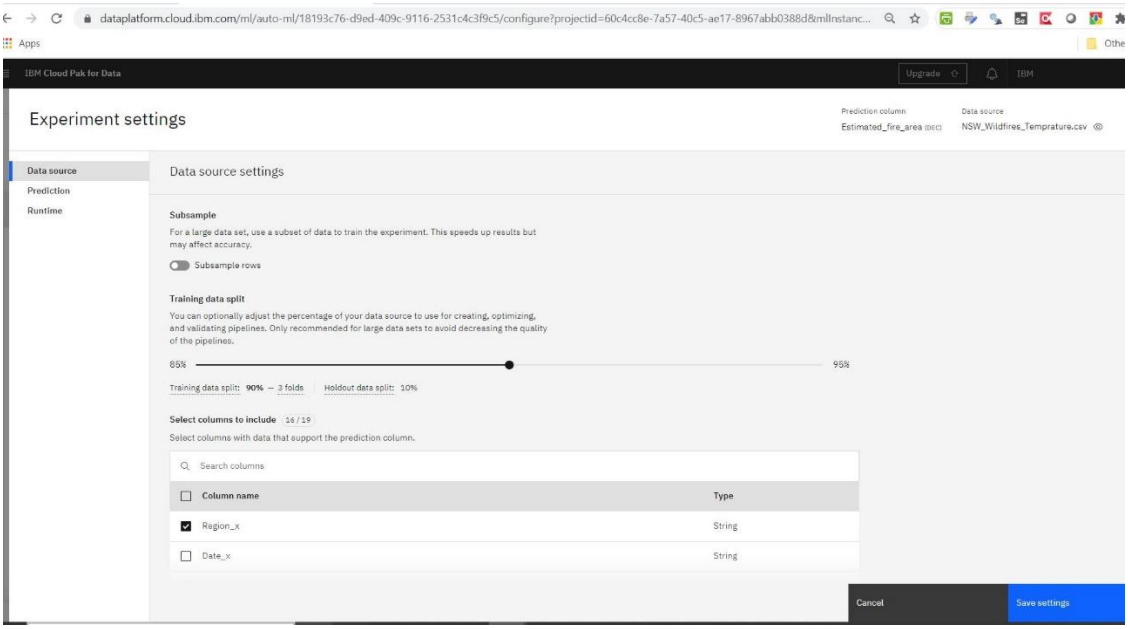
Click on the “Run experiment” button at the bottom right to invoke the running of the Experiment.

The “Experiment setting” button will let you select the setting associated with the experiment.

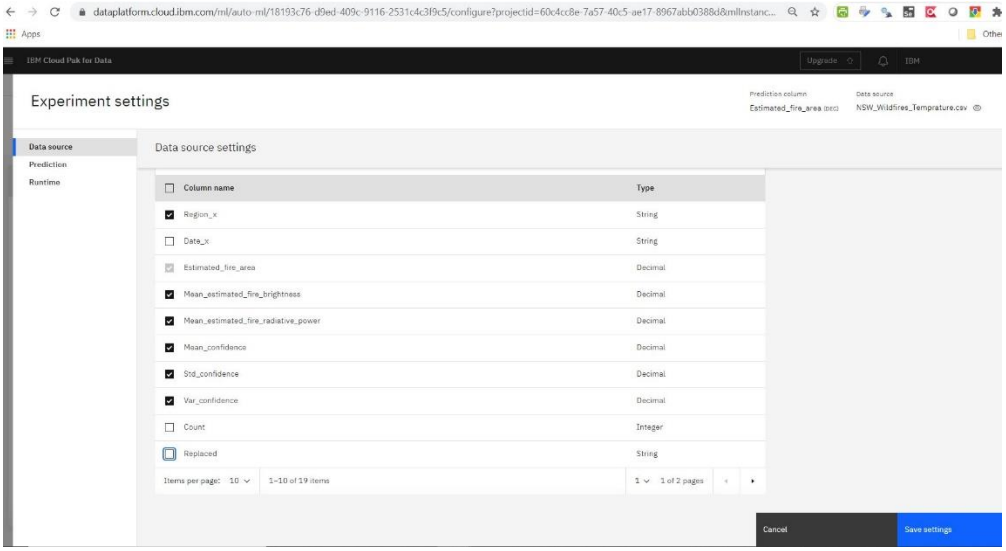
Experiment Settings

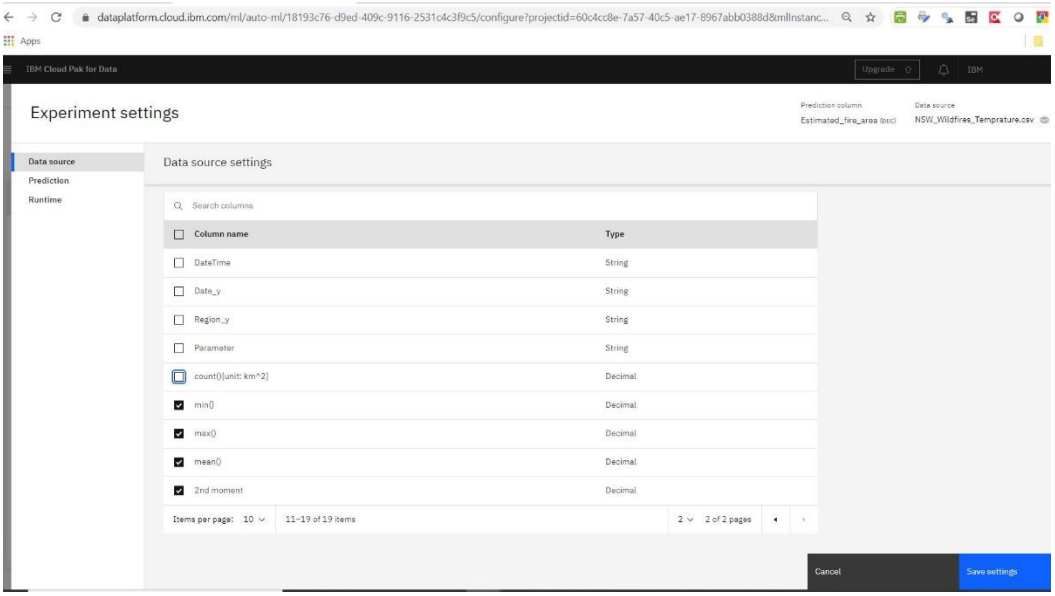
Data Source:

Set the data splitting, column selection to include in the experiment

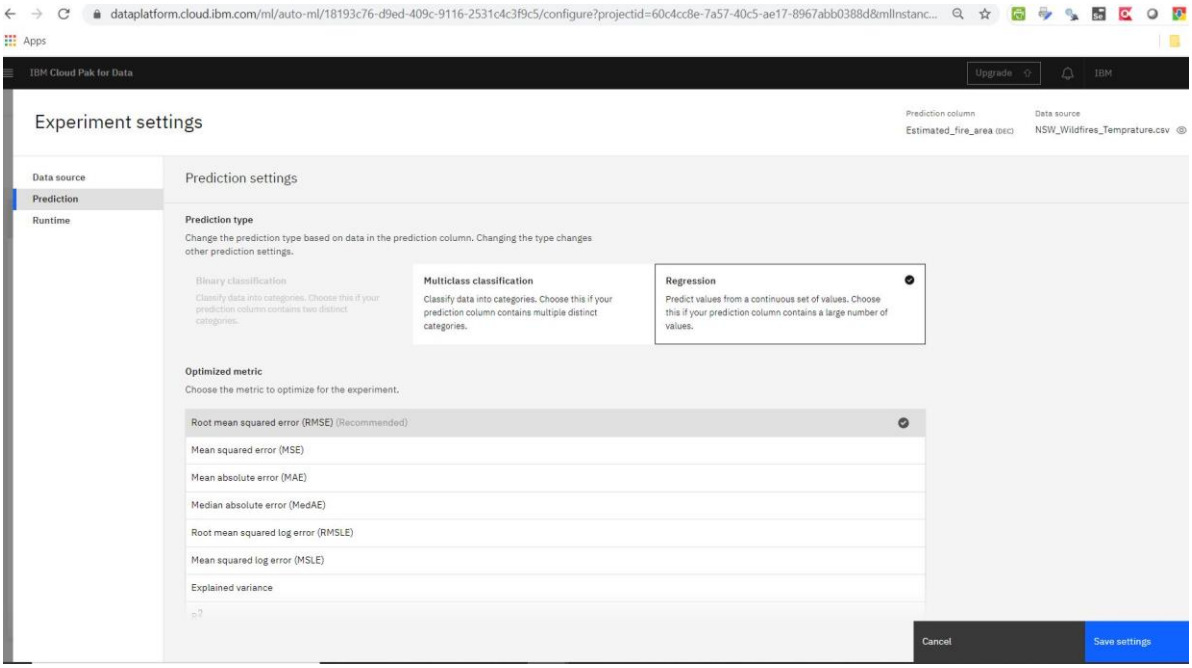


Uncheck box to exclude column data from the experiment

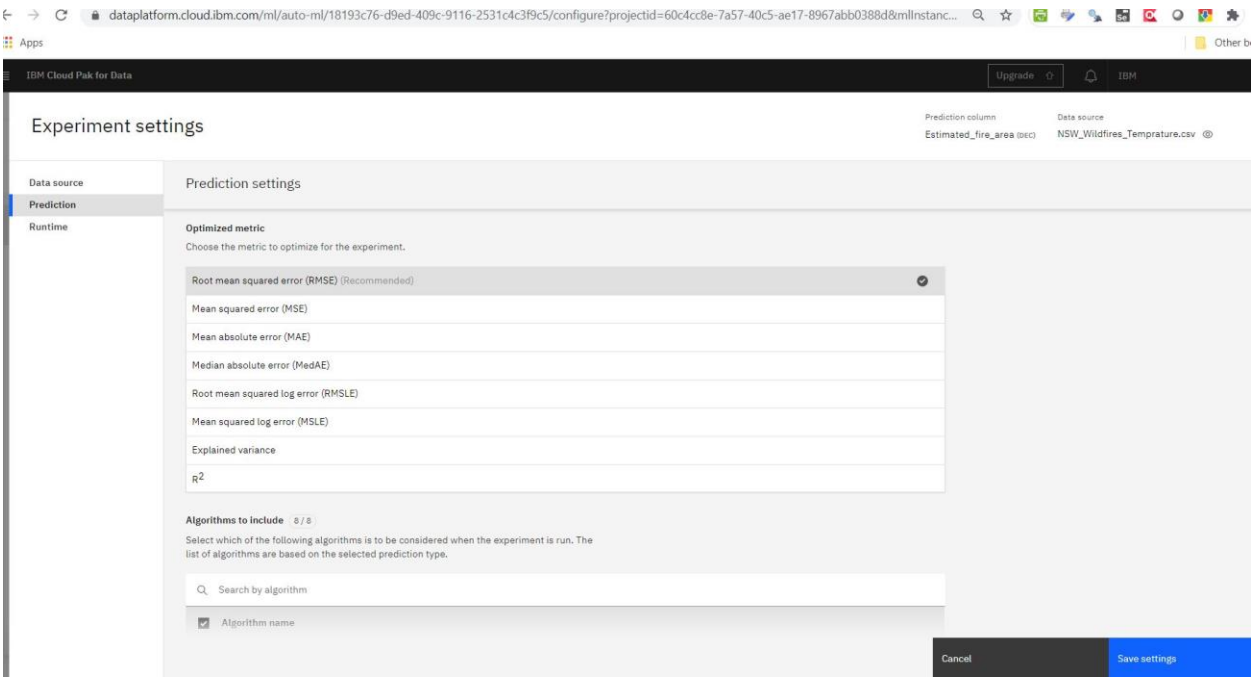



dataplatform.cloud.ibm.com/ml/auto-ml/18193c76-d9ed-409c-9116-2531c4c3f9c5/configure?projectid=60c4cc8e-7a57-40c5-ae17-8967abb0388d&mlInstanc...
Apps
IBM Cloud Pak for Data
Upgrade
IBM
Experiment settings
Prediction column
Estimated_fire_area (bec)
Data source
NSW_Wildfires_Temprature.csv
Data source
Prediction
Runtime
Data source settings
Search columns
Column name
Type
DateTime
String
Date_y
String
Region_y
String
Parameter
String
count()[unit: km^2]
Decimal
min()
Decimal
max()
Decimal
mean()
Decimal
2nd moment
Decimal
Items per page: 10
11–19 of 19 items
2
2 of 2 pages
Cancel
Save settings

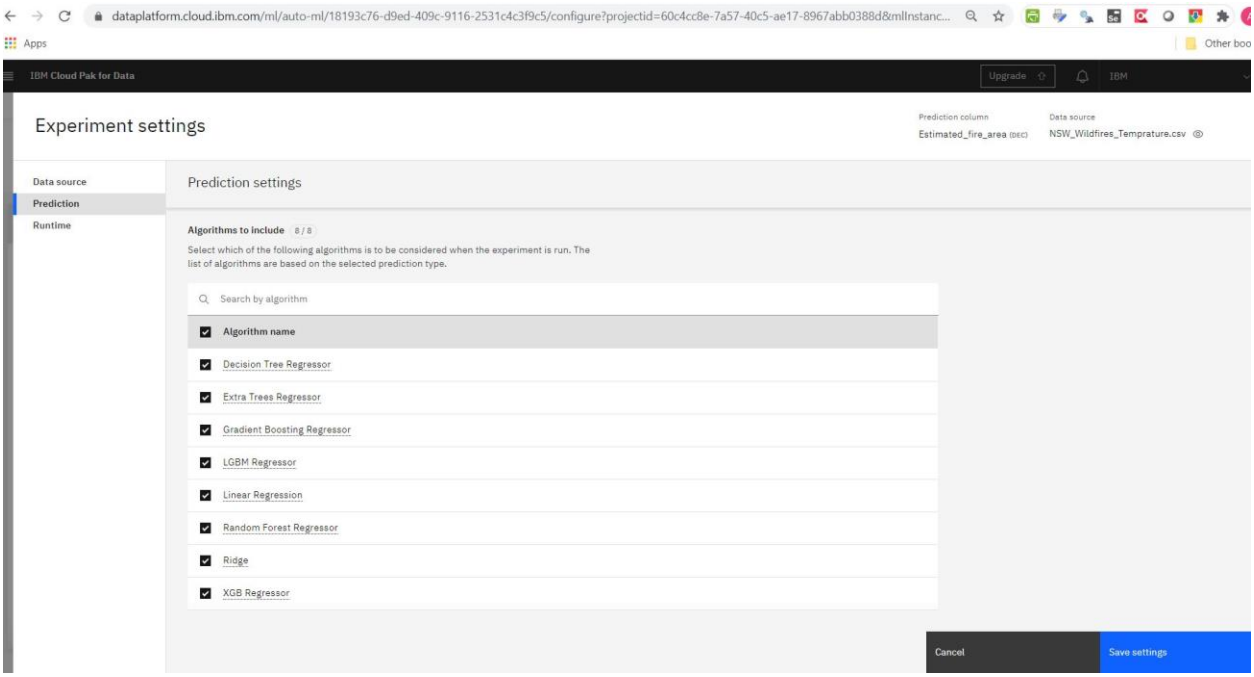
Prediction:



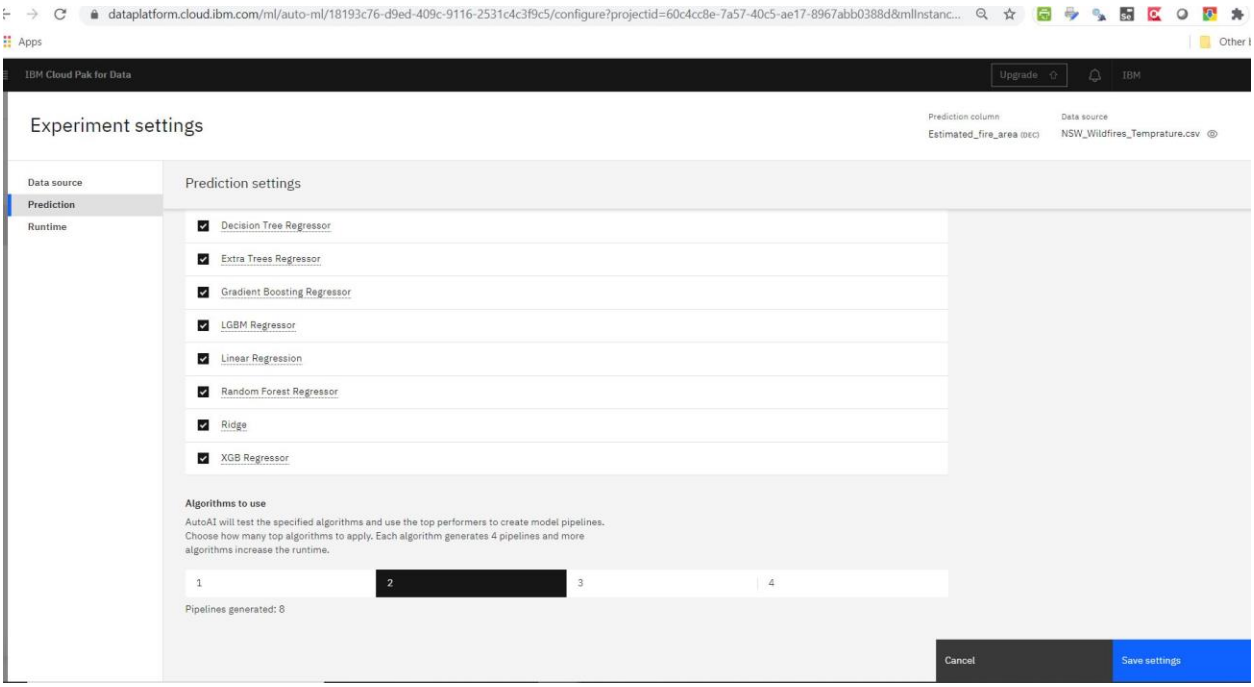
AutoAI engine uses Estimated_fire_area as the prediction column, as a Regression experiment with the choices of Optimized metrics, shown. RMSE is selected in this case.

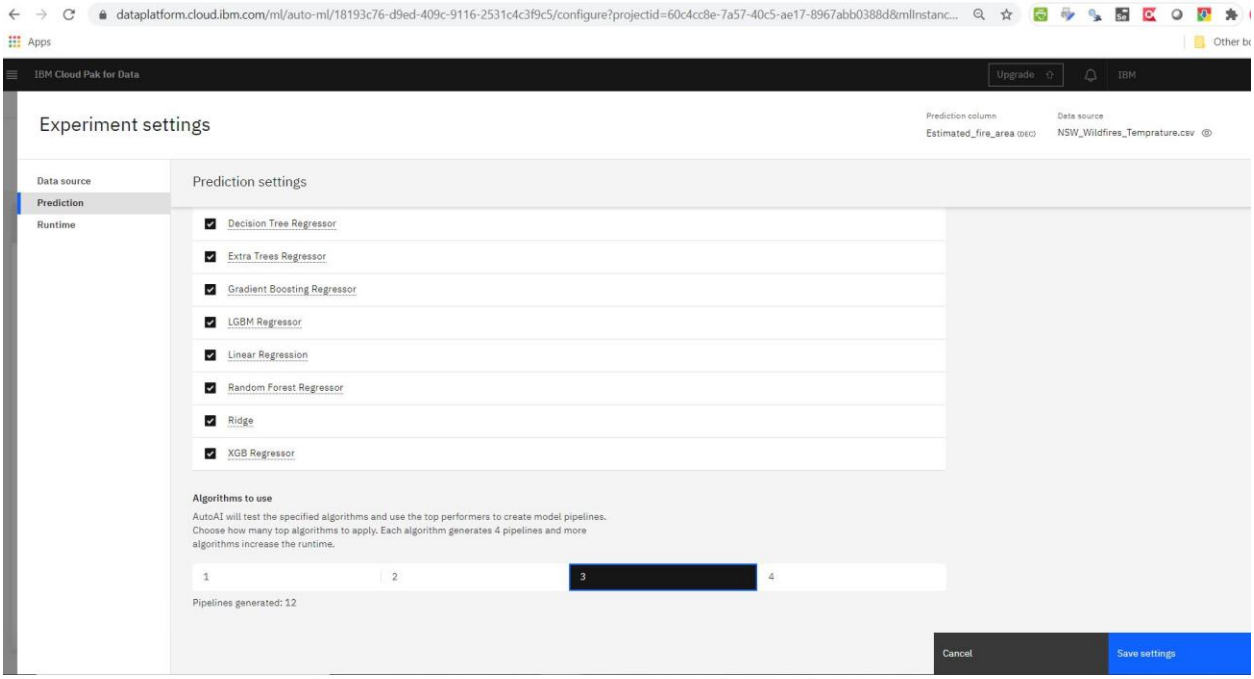


Algorithms that AutoAI will use is listed for selection next:

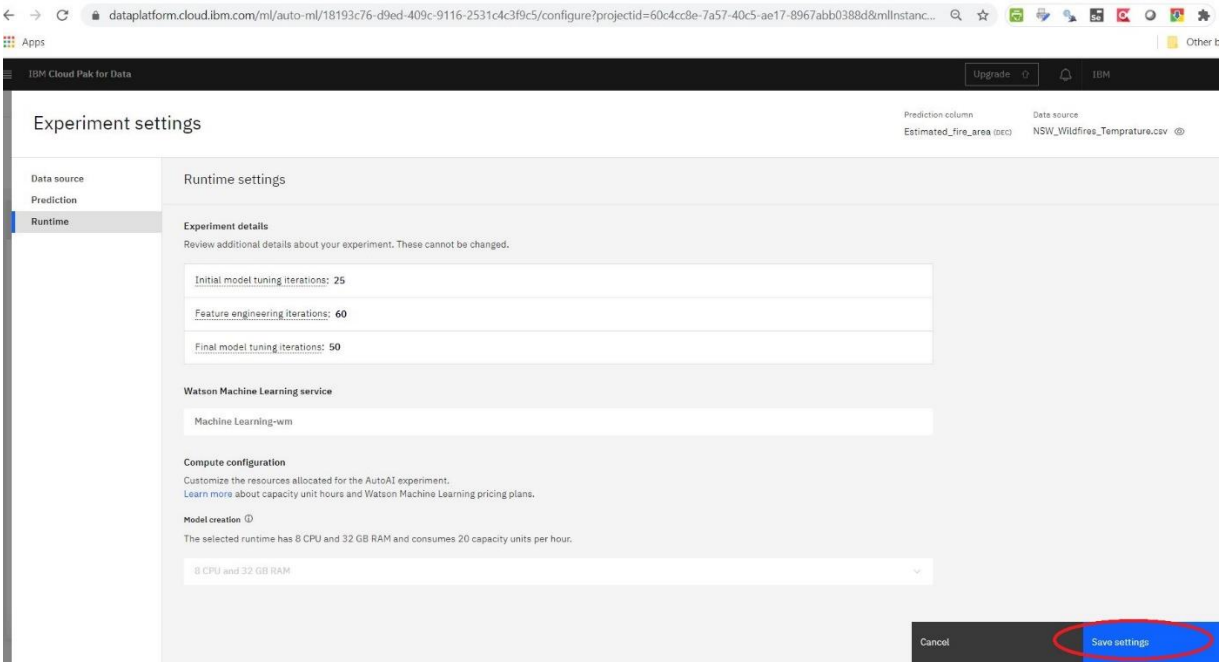


Select the number of top algorithms that AutoAI will test and use as shown: More algorithms increase the runtime as well as CUH.




dataplatform.cloud.ibm.com/ml/auto-ml/18193c76-d9ed-409c-9116-2531c4c3f9c5/configure?projectid=60c4cc8e-7a57-40c5-ae17-8967abb0388d&mlInstanc...
Apps
IBM Cloud Pak for Data
Upgrade
IBM
Experiment settings
Prediction column
Estimated_fire_area (DEC)
Data source
NSW_Wildfires_Temprature.csv
Data source
Prediction
Runtime
Prediction settings
Decision Tree Regressor
Extra Trees Regressor
Gradient Boosting Regressor
LGBM Regressor
Linear Regression
Random Forest Regressor
Ridge
XGB Regressor
Algorithms to use
AutoAI will test the specified algorithms and use the top performers to create model pipelines. Choose how many top algorithms to apply. Each algorithm generates 4 pipelines and more algorithms increase the runtime.
1
2
3
4
Pipelines generated: 12
Cancel
Save settings

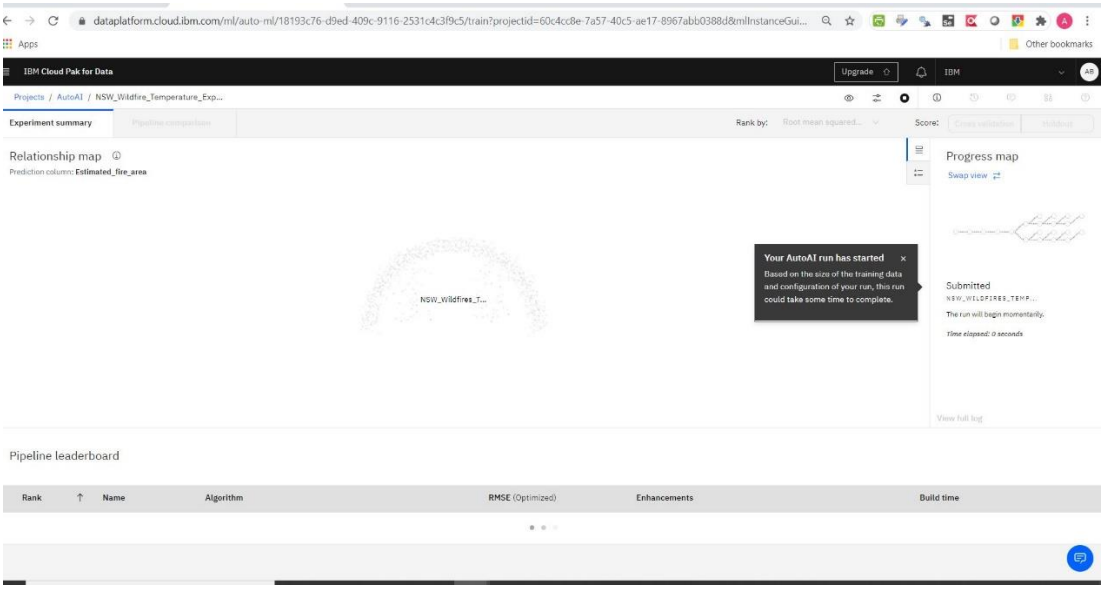
Runtime Settings:

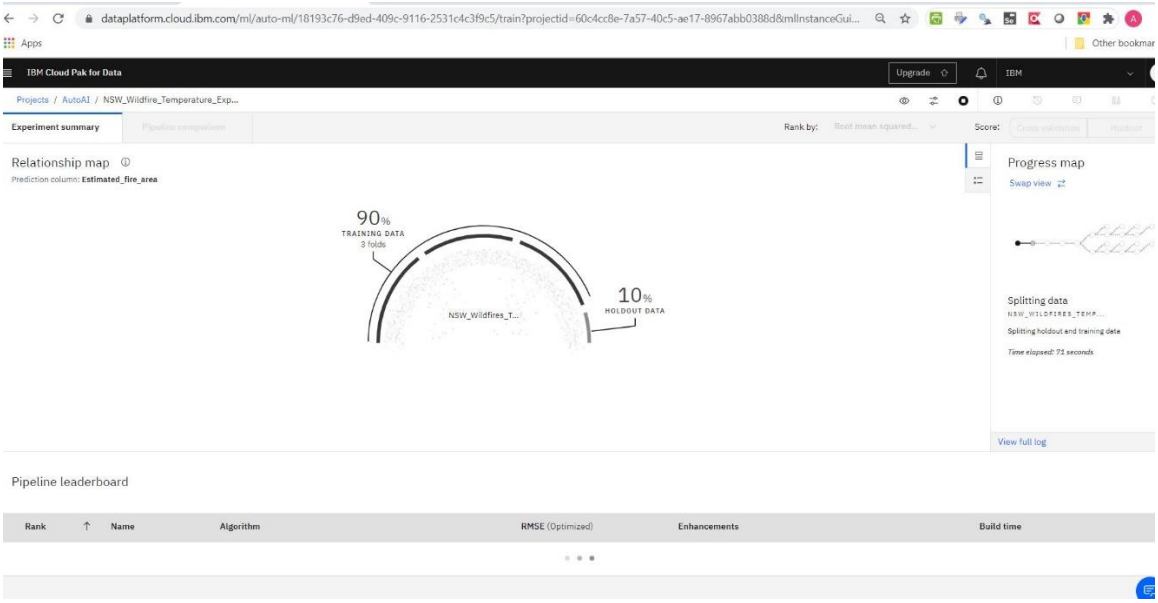


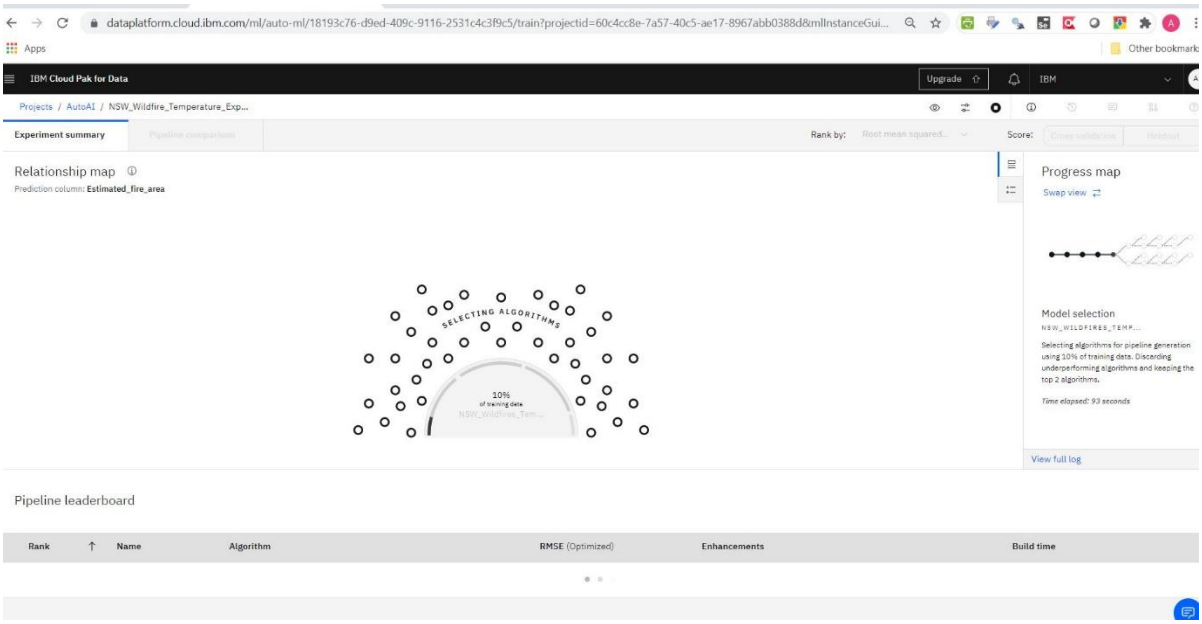
Save settings, before progressing to the next step.

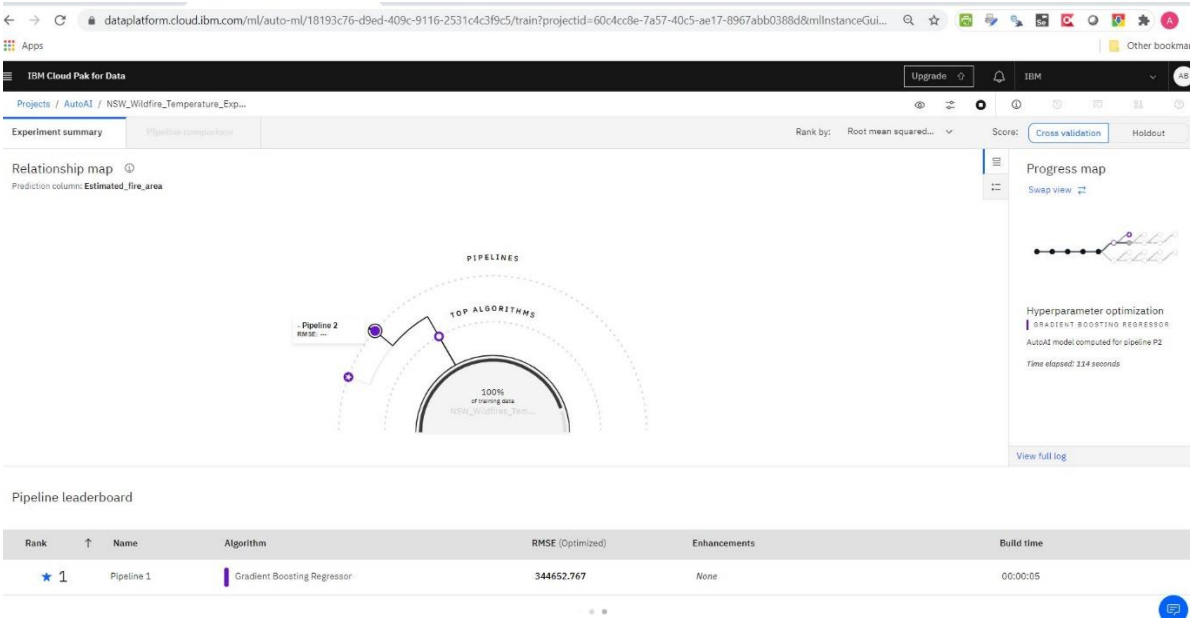
Run Experiment

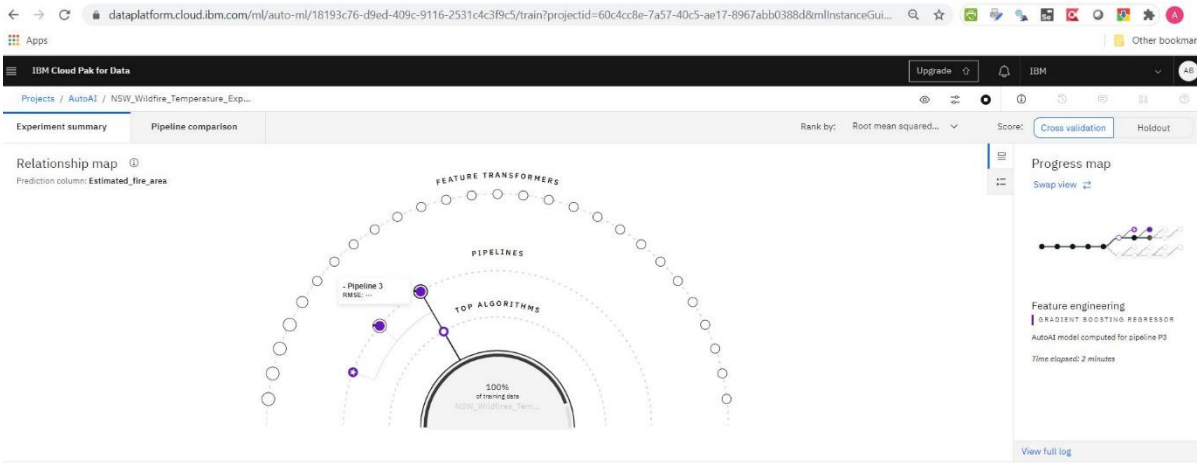
The next few screen captures detail the progress of the experiment.





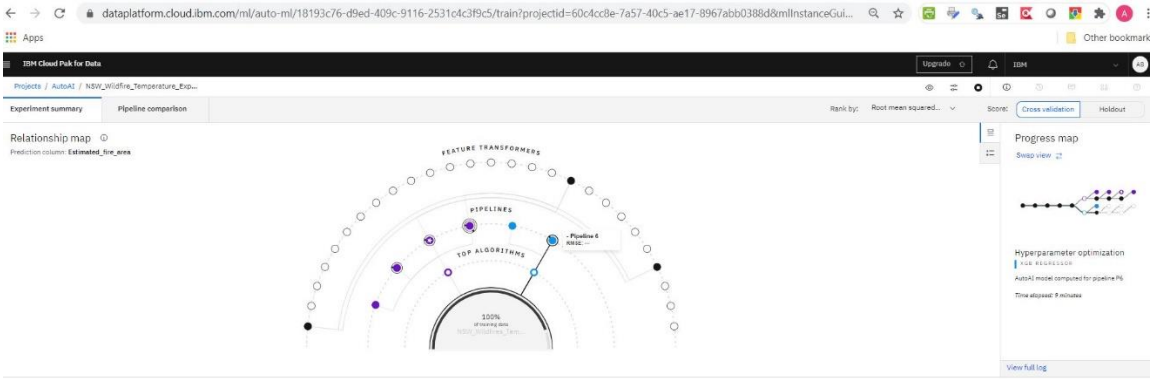

dataplatform.cloud.ibm.com/ml/auto-ml/18193c76-d9ed-409c-9116-2531c4c3f9c5/train?projectid=60c4cc8e-7a57-40c5-ae17-8967abb0388d&mlInstanceGui...
Apps
IBM Cloud Pak for Data
Upgrade
IBM
Projects / AutoAI / NSW_Wildfire_Temperature_Exp...
Experiment summary
Rank by:
Root mean squared...
Score:
Relationship map
Prediction column: Estimated_fire_area
SELECTING ALGORITHMS
10%
of training data
Progress map
Swap view
Model selection
NSW_WILDFIRES_TEMP...
Selecting algorithms for pipeline generation using 10% of training data. Discarding underperforming algorithms and keeping the top 2 algorithms.
Time elapsed: 93 seconds
View full log
Pipeline leaderboard
Rank
Name
Algorithm
RMSE (Optimized)
Enhancements
Build time


dataplatform.cloud.ibm.com/ml/auto-ml/18193c76-d9ed-409c-9116-2531c4c3f9c5/train?projectid=60c4cc8e-7a57-40c5-ae17-8967abb0388d&mlInstanceGui...
Apps
IBM Cloud Pak for Data
Upgrade
IBM
Projects / AutoAI / NSW_Wildfire_Temperature_Exp...
Experiment summary
Rank by:
Root mean squared...
Score:
Cross validation
Holdout
Relationship map
Prediction column: Estimated_fire_area
PIPELINES
TOP ALGORITHMS
Pipeline 2
100%
of training data
Progress map
Swap view
Hyperparameter optimization
GRADIENT BOOSTING REGRESSOR
AutoAI model computed for pipeline P2
Time elapsed: 114 seconds
View full log
Pipeline leaderboard
Rank
Name
Algorithm
RMSE (Optimized)
Enhancements
Build time
1
Pipeline 1
Gradient Boosting Regressor
344652.767
None
00:00:05



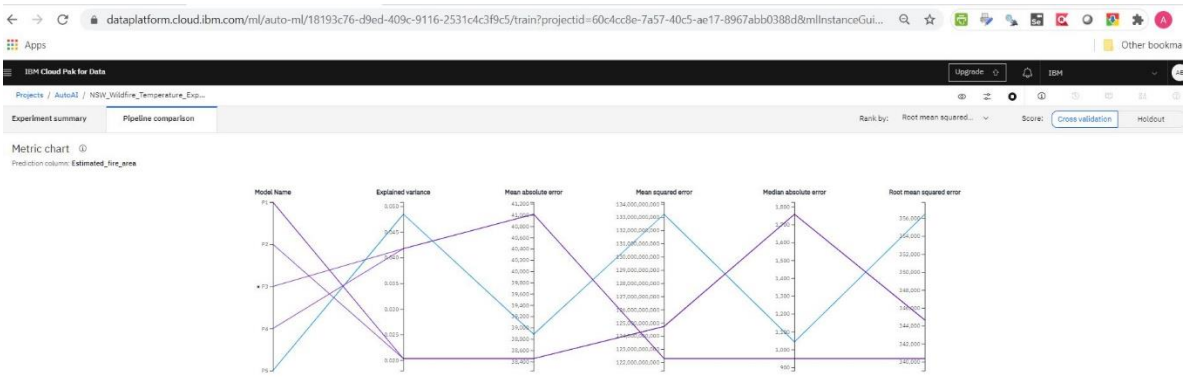
Pipeline leaderboard

Rank	Name	Algorithm	RMSE (Optimized)	Enhancements	Build time
★ 1	Pipeline 1	Gradient Boosting Regressor	344652.767	None	00:00:05
2	Pipeline 2	Gradient Boosting Regressor	344652.767	HPO-1	00:00:28



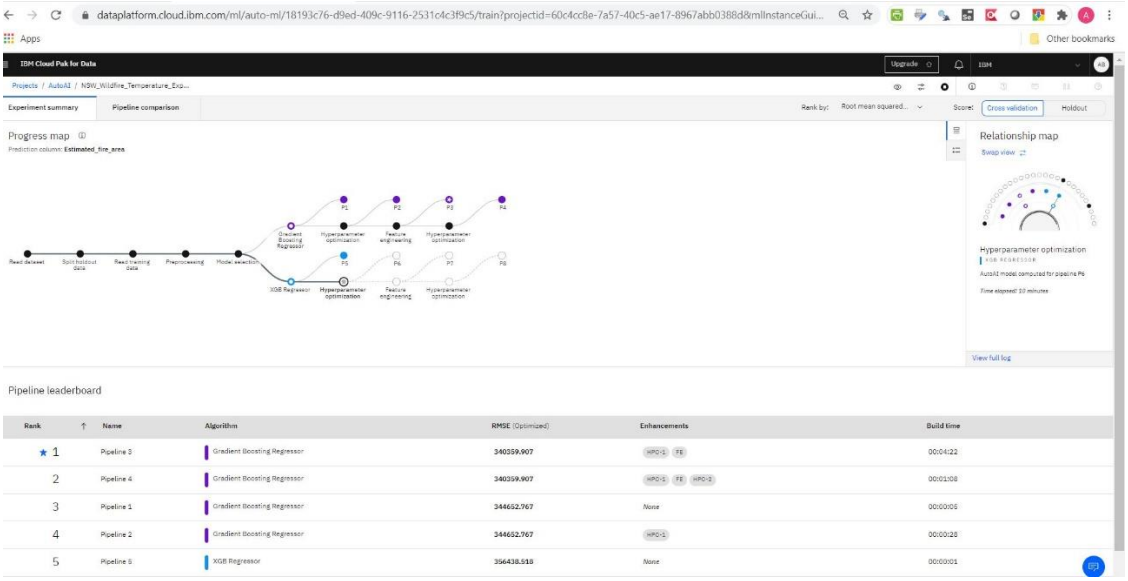
Pipeline leaderboard

Rank	Name	Algorithm	RMSE (Optimized)	Enhancements	Build time
★ 1	Pipeline 3	Gradient Boosting Regressor	340359.907	HPO-1 FE	00:04:22
2	Pipeline 4	Gradient Boosting Regressor	340359.907	HPO-1 FE HPO-2	00:01:08
3	Pipeline 1	Gradient Boosting Regressor	344652.767	None	00:00:05
4	Pipeline 2	Gradient Boosting Regressor	344652.767	HPO-1	00:00:28
5	Pipeline 6	XGB Regressor	356438.518	None	00:00:01



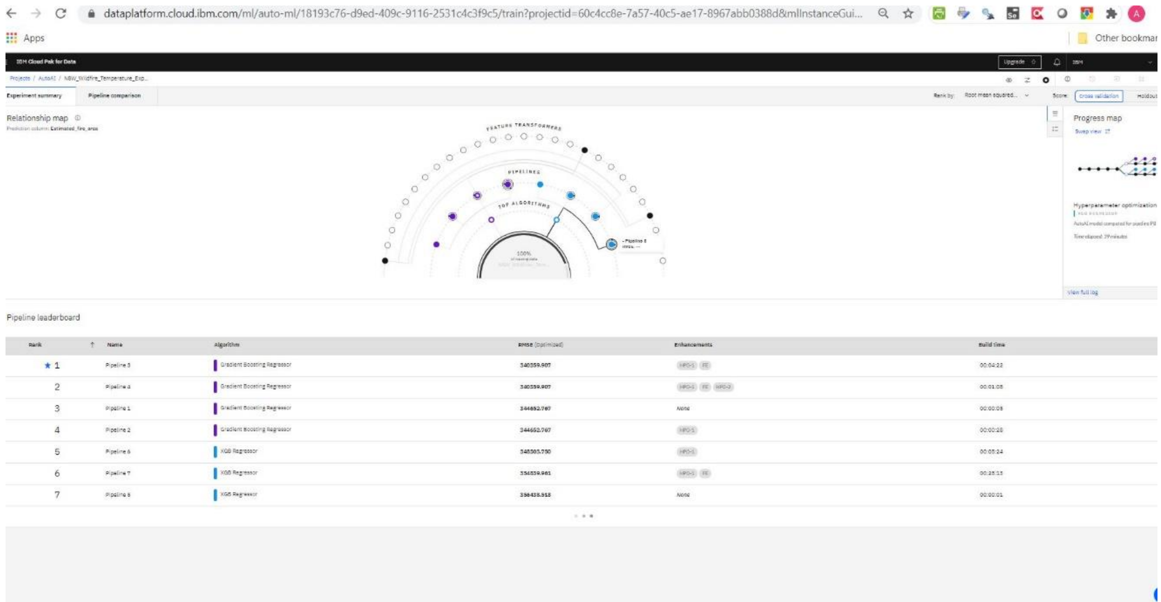
Pipeline leaderboard

Rank	Name	Algorithm	RMSE (Optimized)	Explained variance	Mean absolute error (MAE)	Mean squared error (MSE)	Median absolute error (Me...	R^2
★ 1	Pipeline 3	Gradient Boosting Regressor	340359.907	0.042	41012.752	122290835316.673	949.340	0.041
2	Pipeline 4	Gradient Boosting Regressor	340359.907	0.042	41012.752	122290835316.673	949.340	0.041
3	Pipeline 1	Gradient Boosting Regressor	344652.767	0.020	38453.033	124725419168.493	1758.102	0.020
4	Pipeline 2	Gradient Boosting Regressor	344652.767	0.020	38453.033	124725419168.493	1758.102	0.020
5	Pipeline 5	XGB Regressor	356438.518	-0.048	38885.389	133227394389.333	1042.307	-0.049

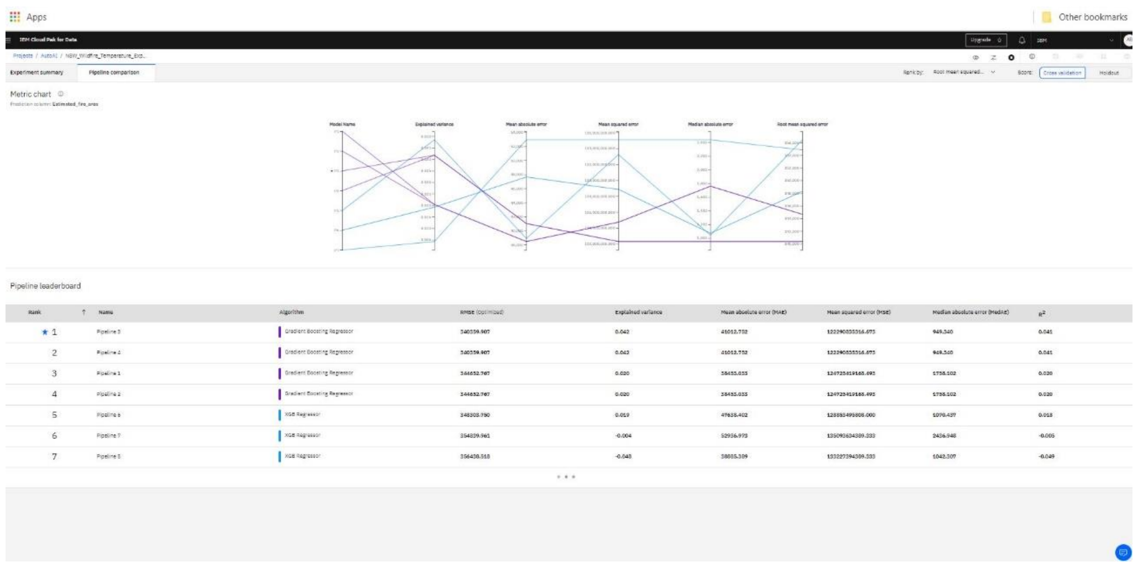


Pipeline leaderboard

Rank	Name	Algorithm	RMSE (Optimized)	Enhancements	Build time
★ 1	Pipeline 3	Gradient Boosting Regressor	340359.907	HPO-1 FE	00:04:22
2	Pipeline 4	Gradient Boosting Regressor	340359.907	HPO-1 FE HPO-2	00:01:08
3	Pipeline 1	Gradient Boosting Regressor	344652.767	None	00:00:05
4	Pipeline 2	Gradient Boosting Regressor	344652.767	HPO-1	00:00:28
5	Pipeline 5	XGB Regressor	356438.518	None	00:00:01

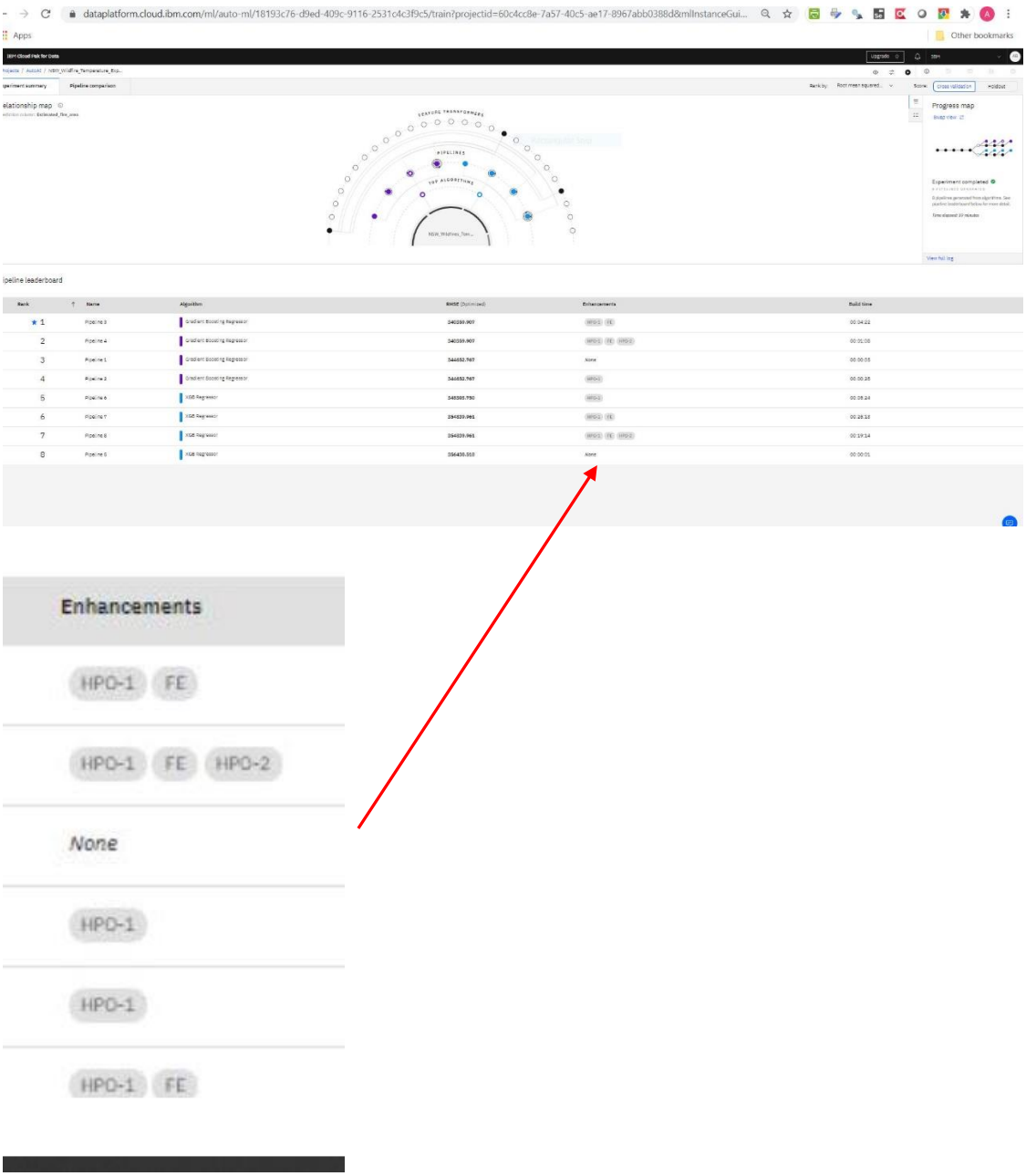


dataplatform.cloud.ibm.com/ml/auto-ml/18193c76-d9ed-409c-9116-2531c4c3f9c5/train?projectid=60c4cc8e-7a57-40c5-ae17-8967abb0388d&mlInstanceGui...
Apps
Other bookmar
IBM Cloud Pak for Data
Upgrade
Experiment summary
Pipeline comparison
Relationship map
Progress map
Hyperparameter optimization
Pipeline leaderboard
Rank
Name
Algorithm
RMSE (Optimized)
Enhancements
Build time
1 Pipeline 3 Gradient Boosting Regressor 340359.907 HPO-1 FE 00:04:22
2 Pipeline 4 Gradient Boosting Regressor 340359.907 HPO-1 FE HPO-2 00:01:08
3 Pipeline 1 Gradient Boosting Regressor 344652.767 None 00:00:05
4 Pipeline 2 Gradient Boosting Regressor 344652.767 HPO-1 00:00:28
5 Pipeline 6 XGB Regressor 348303.750 HPO-1 00:05:24
6 Pipeline 7 XGB Regressor 354839.961 HPO-1 FE 00:25:13
7 Pipeline 5 XGB Regressor 356438.518 None 00:00:01

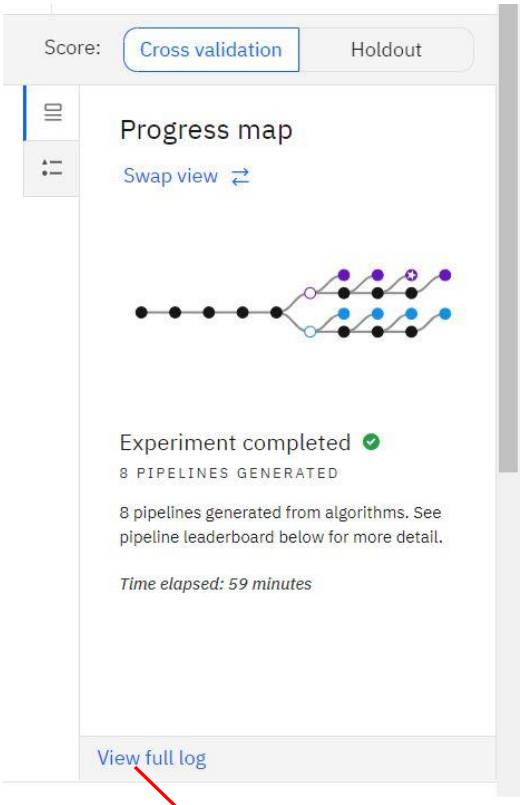


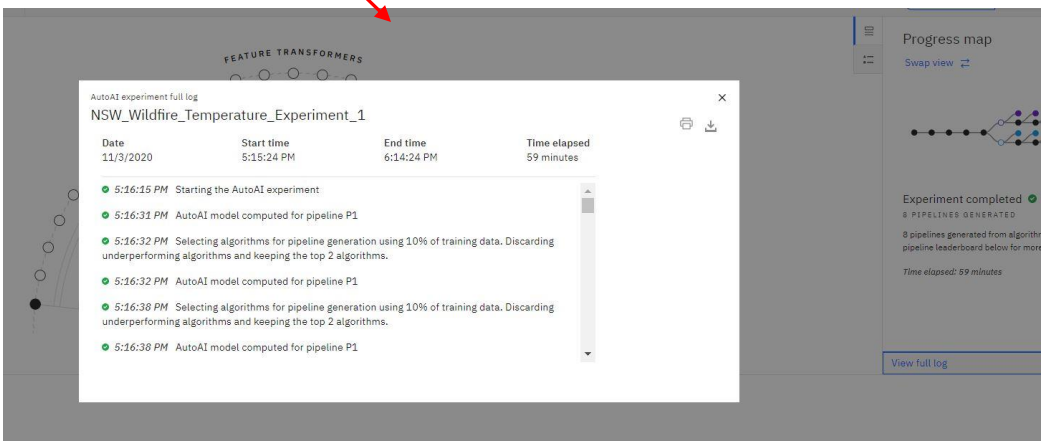
Apps
Other bookmarks
IBM Cloud Pak for Data
Experiment summary
Pipeline comparison
Metric chart
Model Name
Explained variance
Mean absolute error
Mean squared error
Median absolute error
Root mean squared error
Pipeline leaderboard
Rank
Name
Algorithm
RMSE (Optimized)
Explained variance
Mean absolute error (MAE)
Mean squared error (MSE)
Median absolute error (MedAE)
R2
1 Pipeline 3 Gradient Boosting Regressor 340359.907 0.042 41012.752 122290833316.673 949.340 0.041
2 Pipeline 4 Gradient Boosting Regressor 340359.907 0.042 41012.752 122290833316.673 949.340 0.041
3 Pipeline 1 Gradient Boosting Regressor 344652.767 0.020 38455.033 124725419168.493 1758.102 0.020
4 Pipeline 2 Gradient Boosting Regressor 344652.767 0.020 38455.033 124725419168.493 1758.102 0.020
5 Pipeline 6 XGB Regressor 348303.750 0.019 47638.402 128555495808.000 1070.437 0.018
6 Pipeline 7 XGB Regressor 354839.961 -0.004 52936.973 135093634389.333 2436.948 -0.005
7 Pipeline 5 XGB Regressor 356438.518 -0.048 38885.309 133227394389.333 1042.307 -0.049

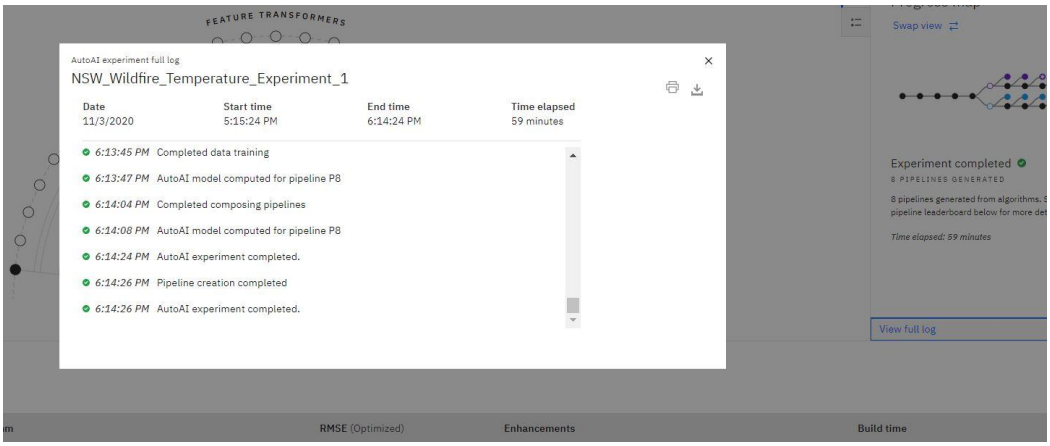
Experiment is Completed, as shown in the next screen.



Showing the top pipelines, the algorithms and the value of RMSE. The enhancements performed are also listed. HPO-Hyper Parameter Optimization and FE – Feature Engineering

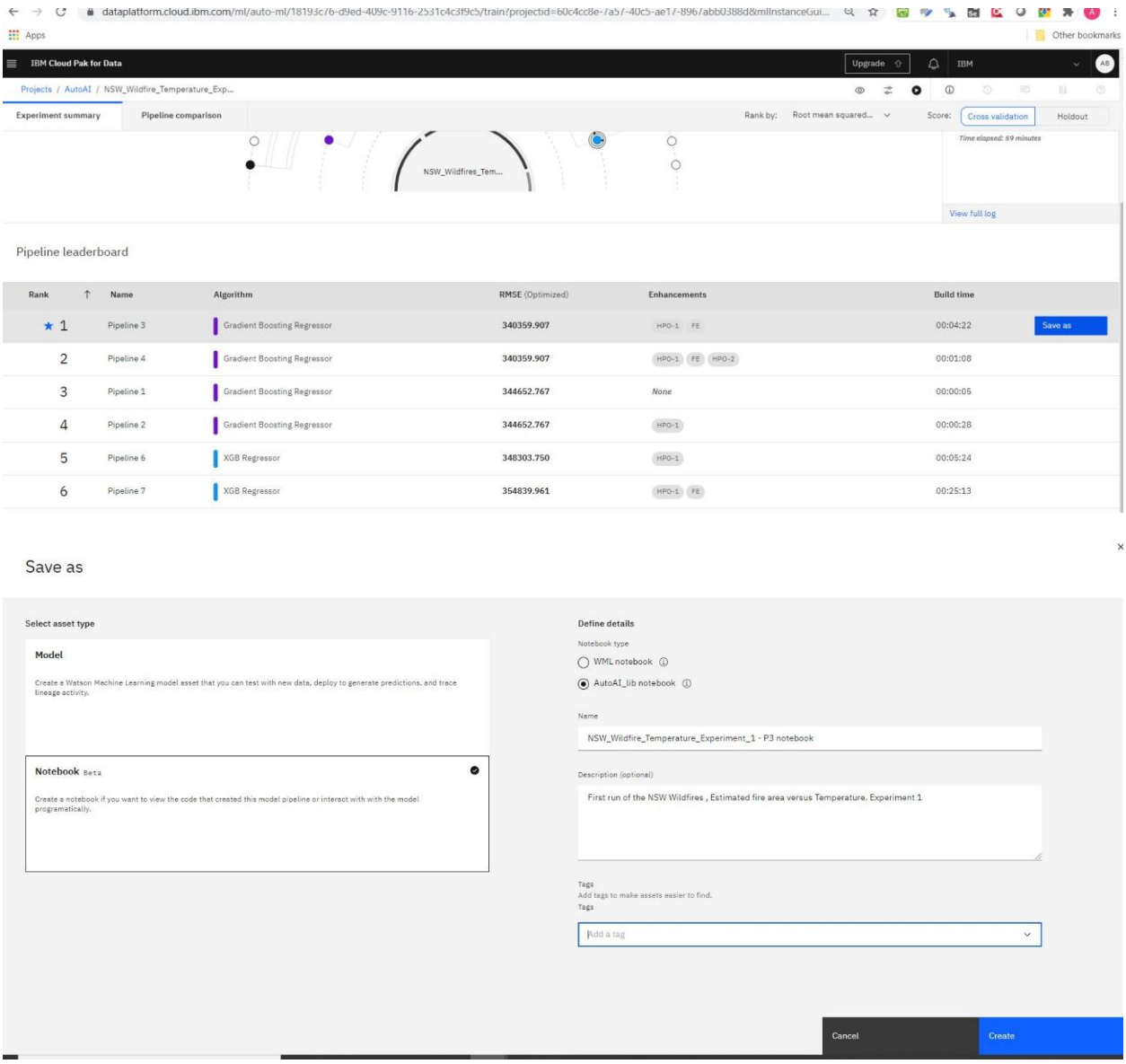

Score:
Cross validation
Holdout
Progress map
Swap view
Experiment completed
8 PIPELINES GENERATED
8 pipelines generated from algorithms. See pipeline leaderboard below for more detail.
Time elapsed: 59 minutes
View full log

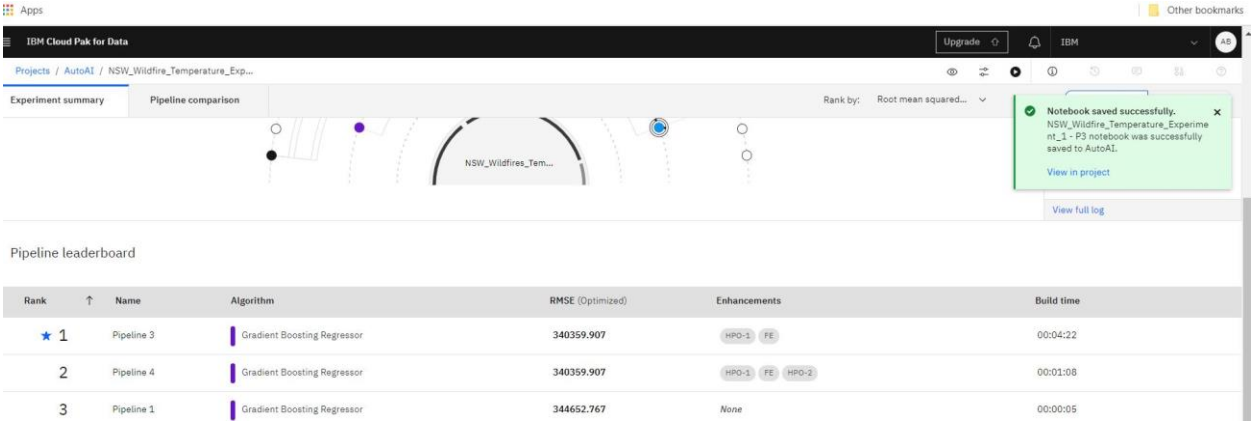

FEATURE TRANSFORMERS
AutoAI experiment full log
NSW_Wildfire_Temperature_Experiment_1
Date
11/3/2020
Start time
5:15:24 PM
End time
6:14:24 PM
Time elapsed
59 minutes
5:16:15 PM Starting the AutoAI experiment
5:16:31 PM AutoAI model computed for pipeline P1
5:16:32 PM Selecting algorithms for pipeline generation using 10% of training data. Discarding underperforming algorithms and keeping the top 2 algorithms.
5:16:32 PM AutoAI model computed for pipeline P1
5:16:38 PM Selecting algorithms for pipeline generation using 10% of training data. Discarding underperforming algorithms and keeping the top 2 algorithms.
5:16:38 PM AutoAI model computed for pipeline P1
Progress map
Swap view
Experiment completed
8 PIPELINES GENERATED
Time elapsed: 59 minutes
View full log

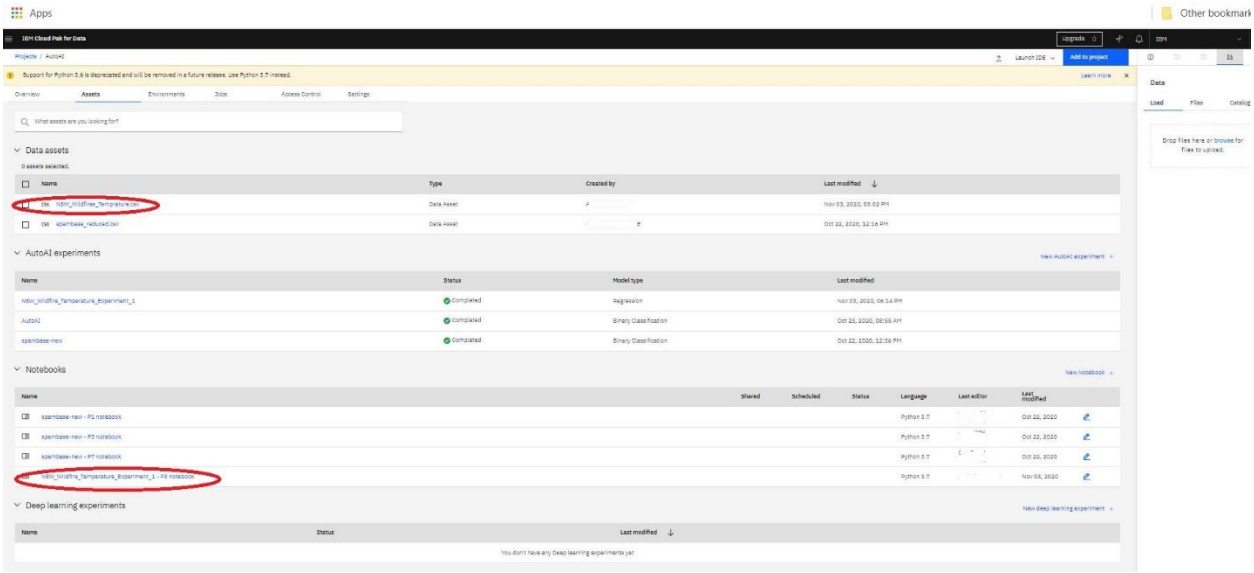

FEATURE TRANSFORMERS
AutoAI experiment full log
NSW_Wildfire_Temperature_Experiment_1
Date
11/3/2020
Start time
5:15:24 PM
End time
6:14:24 PM
Time elapsed
59 minutes
6:13:45 PM Completed data training
6:13:47 PM AutoAI model computed for pipeline P8
6:14:04 PM Completed composing pipelines
6:14:08 PM AutoAI model computed for pipeline P8
6:14:24 PM AutoAI experiment completed.
6:14:26 PM Pipeline creation completed
6:14:26 PM AutoAI experiment completed.
Swap view
Experiment completed
8 PIPELINES GENERATED
Time elapsed: 59 minutes
View full log
RMSE (Optimized)
Enhancements
Build time

Saving the work

Save the top pipeline as required.

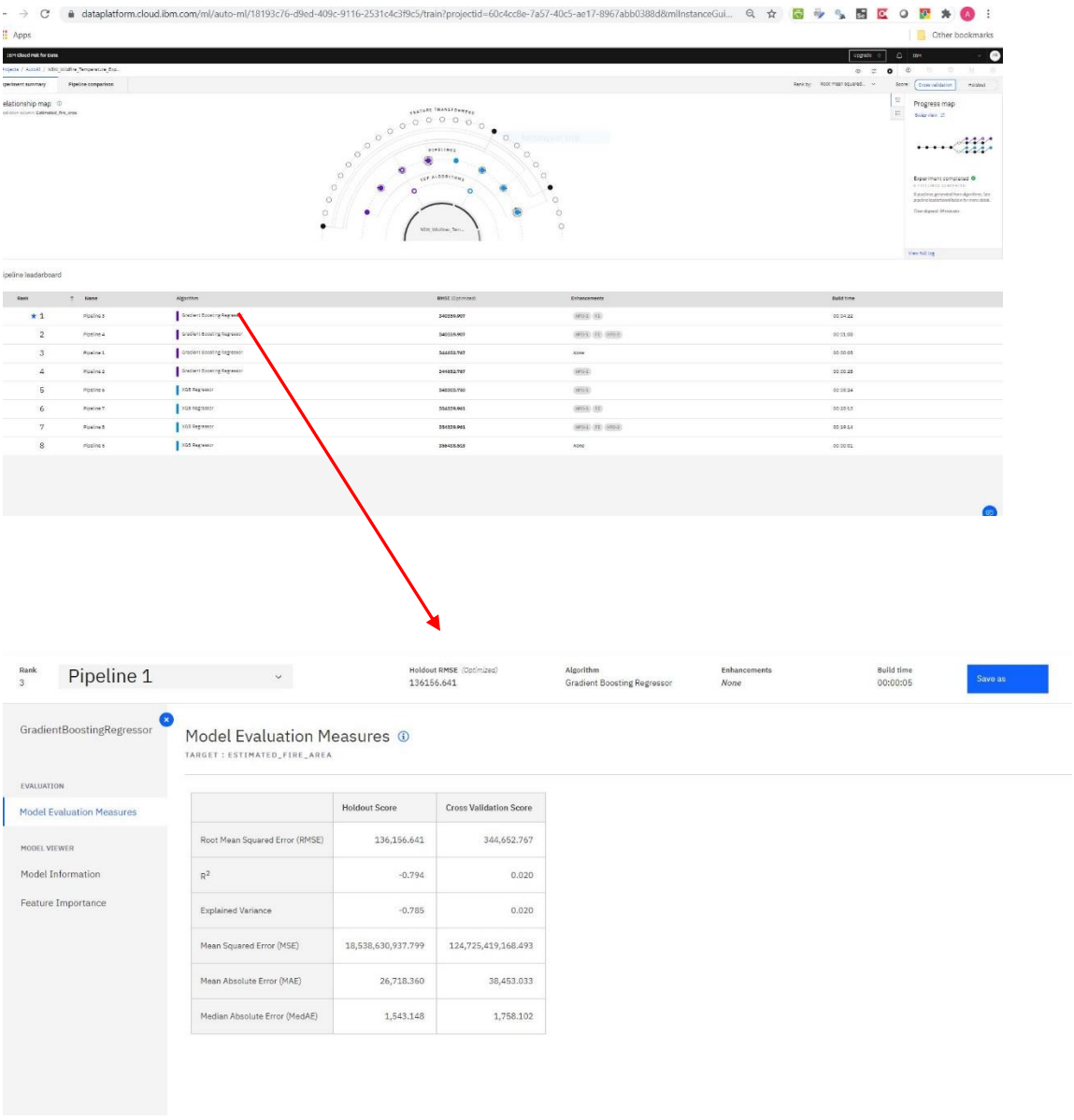


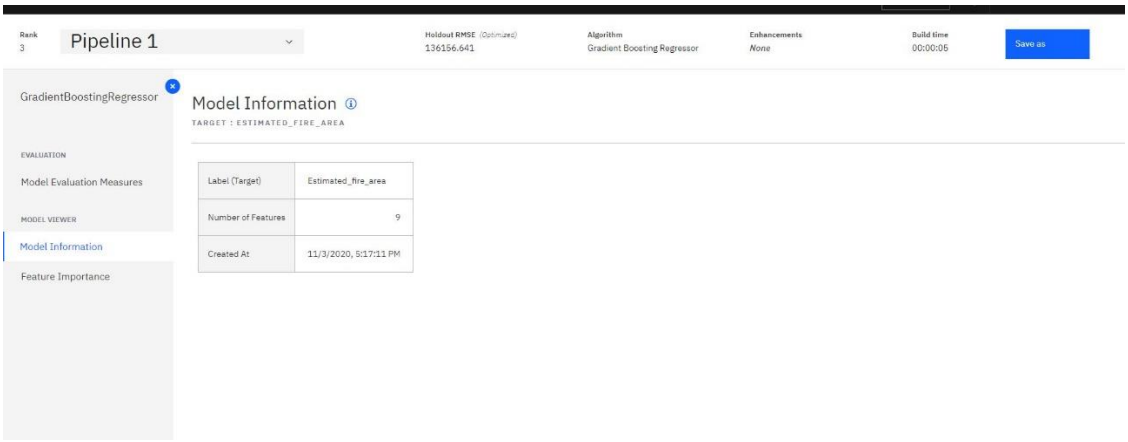

Apps
Other bookmarks
IBM Cloud Pak for Data
Upgrade
IBM
Projects / AutoAI / NSW_Wildfire_Temperature_Exp...
Experiment summary
Pipeline comparison
Rank by:
Root mean squared...
Notebook saved successfully.
NSW_Wildfire_Temperature_Experiment_1 - P3 notebook was successfully saved to AutoAI.
View in project
View full log
NSW_Wildfires_Tem...
Pipeline leaderboard
Rank
Name
Algorithm
RMSE (Optimized)
Enhancements
Build time
1
Pipeline 3
Gradient Boosting Regressor
340359.907
HPO-1 FE
00:04:22
2
Pipeline 4
Gradient Boosting Regressor
340359.907
HPO-1 FE HPO-2
00:01:08
3
Pipeline 1
Gradient Boosting Regressor
344652.767
None
00:00:05

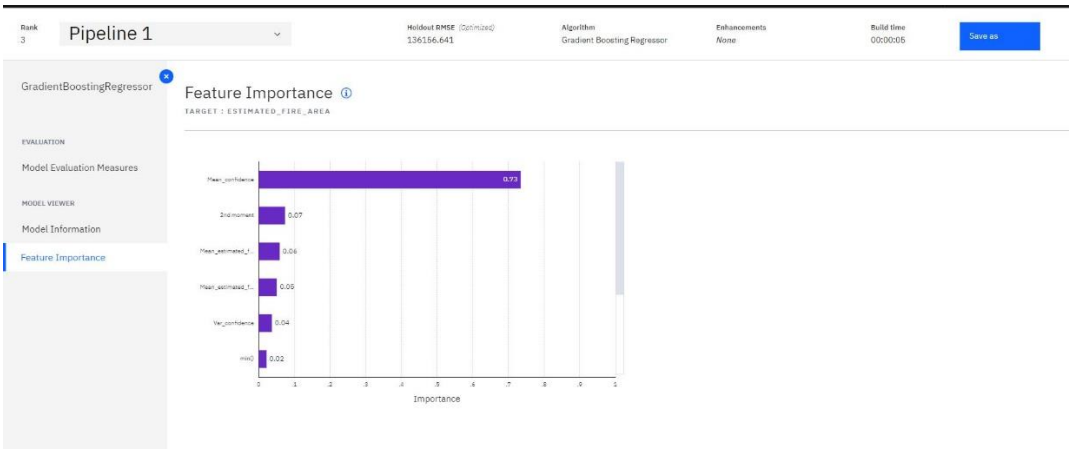

Apps
Other bookmark
Data assets
AutoAI experiments
Notebooks
Deep learning experiments

Details of the Algorithm

Click the link in the link in the Algorithm column of the top pipeline to drill down into the working of the pipeline as shown.




Rank
3
Pipeline 1
Holdout RMSE (Optimized)
136156.641
Algorithm
Gradient Boosting Regressor
Enhancements
None
Build time
00:00:05
Save as
GradientBoostingRegressor
EVALUATION
Model Evaluation Measures
MODEL VIEWER
Model Information
Feature Importance
Model Information
TARGET : ESTIMATED_FIRE_AREA
Label (Target)
Estimated_fire_area
Number of Features
9
Created At
11/3/2020, 5:17:11 PM


Rank
3
Pipeline 1
Holdout RMSE (Optimized)
136156.641
Algorithm
Gradient Boosting Regressor
Enhancements
None
Build time
00:00:05
Save as
GradientBoostingRegressor
EVALUATION
Model Evaluation Measures
MODEL VIEWER
Model Information
Feature Importance
Feature Importance
TARGET : ESTIMATED_FIRE_AREA
0.73
0.07
0.06
0.05
0.04
0.02
Importance


Rank
3
Pipeline 1
Holdout RMSE (Optimized)
136156.641
Algorithm
Gradient Boosting Regressor
Enhancements
None
Build time
00:00:05
Save as
GradientBoostingRegressor
EVALUATION
Model Evaluation Measures
MODEL VIEWER
Model Information
Feature Importance
Feature Importance
TARGET : ESTIMATED_FIRE_AREA
0.05
0.04
0.02
0.01
0.01
0.00
Importance

Rank 8 | Pipeline 5 | Holdout RMSE (Optimized) 151113.907 | Algorithm XGB Regressor | Enhancements *None* | Build time 00:00:01 | Save as

XGBRegressor

EVALUATION

Model Evaluation Measures

MODEL VIEWER

Model Information

Feature Importance

Model Evaluation Measures

TARGET : ESTIMATED_FIRE_AREA

	Holdout Score	Cross Validation Score
Root Mean Squared Error (RMSE)	151,113.907	356,438.518
R^2	-1.210	-0.049
Explained Variance	-1.198	-0.048
Mean Squared Error (MSE)	22,835,412,992.000	133,227,394,389.333
Mean Absolute Error (MAE)	26,708.814	38,885.389
Median Absolute Error (MedAE)	1,051.590	1,042.307

Rank 8 | Pipeline 5 | Holdout RMSE (Optimized) 151113.907 | Algorithm XGB Regressor | Enhancements *None* | Build time 00:00:01 | Save as

XGBRegressor

EVALUATION

Model Evaluation Measures

MODEL VIEWER

Model Information

Feature Importance

Model Information

TARGET : ESTIMATED_FIRE_AREA

Label (Target)	Estimated_fire_area
Number of Features	9
Created At	11/3/2020, 5:23:32 PM

Rank 8 | Pipeline 5 | Holdout RMSE (Optimized) 151113.907 | Algorithm XGB Regressor | Enhancements *None* | Build time 00:00:01 | Save as

XGBRegressor

EVALUATION

Model Evaluation Measures

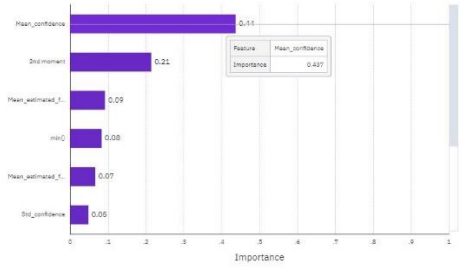
MODEL VIEWER

Model Information

Feature Importance

Feature Importance

TARGET : ESTIMATED_FIRE_AREA



Rank 8 | Pipeline 5 | Holdout RMSE (Optimized) 151113.907 | Algorithm XGB Regressor | Enhancements None | Build time 00:00:01 | Save as

XGBRegressor

EVALUATION

Model Evaluation Measures

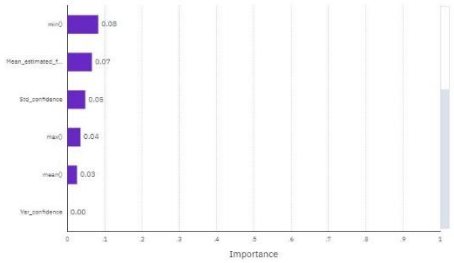
MODEL VIEWER

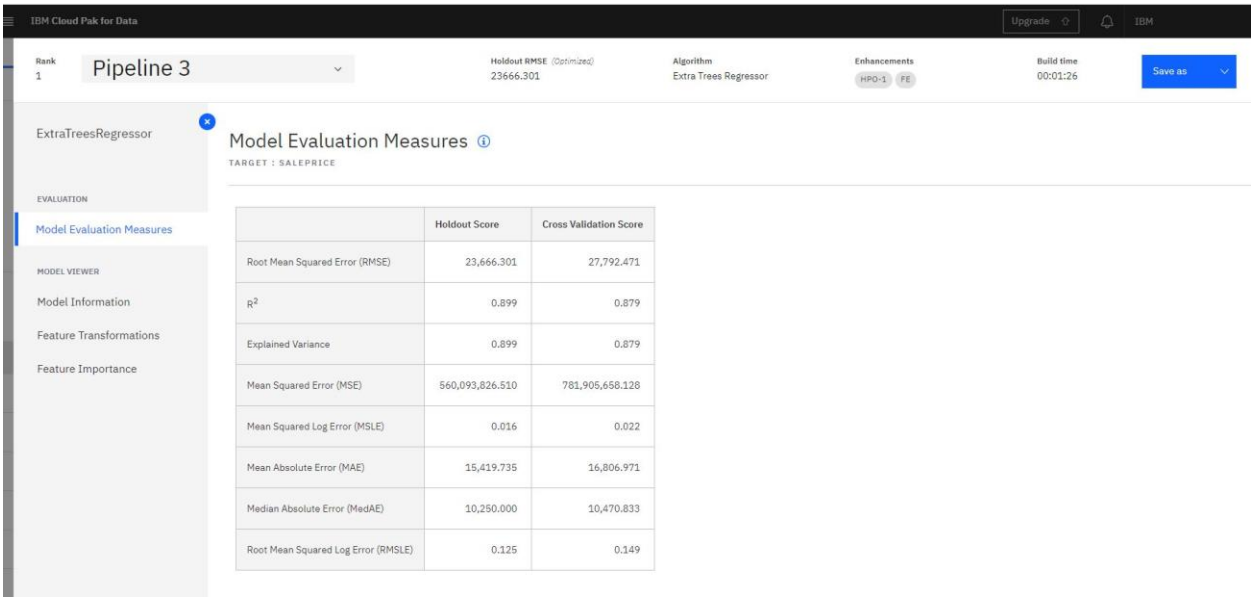
Model Information

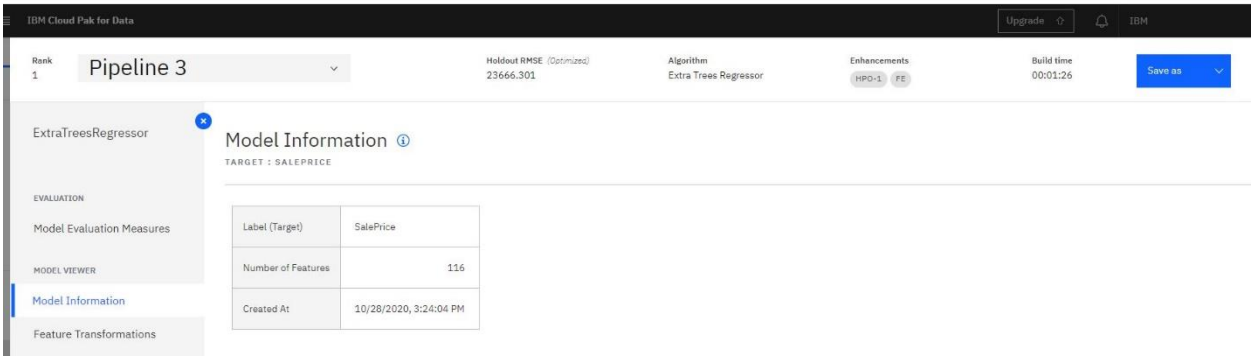
Feature Importance

Feature Importance

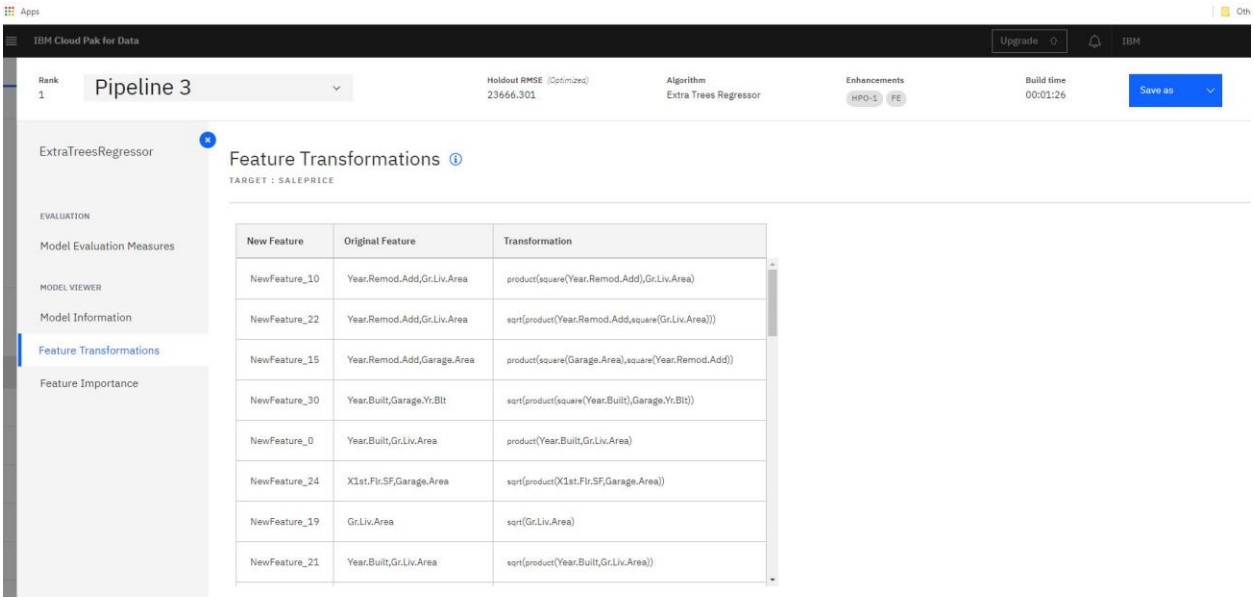
TARGET : ESTIMATED_FIRE_AREA



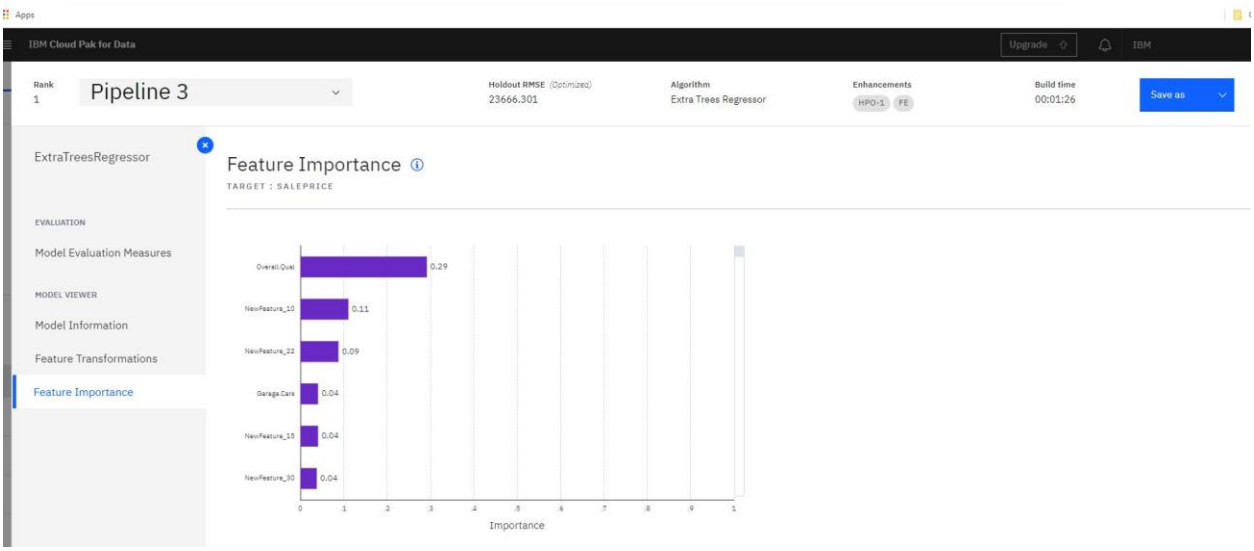

IBM Cloud Pak for Data
Upgrade
IBM
Rank
1
Pipeline 3
Holdout RMSE (Optimized)
23666.301
Algorithm
Extra Trees Regressor
Enhancements
HPO-1
FE
Build time
00:01:26
Save as
ExtraTreesRegressor
EVALUATION
Model Evaluation Measures
MODEL VIEWER
Model Information
Feature Transformations
Feature Importance
Model Evaluation Measures
TARGET : SALEPRICE
Holdout Score
Cross Validation Score
Root Mean Squared Error (RMSE) 23,666.301 27,792.471
R² 0.899 0.879
Explained Variance 0.899 0.879
Mean Squared Error (MSE) 560,093,826.510 781,905,658.128
Mean Squared Log Error (MSLE) 0.016 0.022
Mean Absolute Error (MAE) 15,419.735 16,806.971
Median Absolute Error (MedAE) 10,250.000 10,470.833
Root Mean Squared Log Error (RMSLE) 0.125 0.149

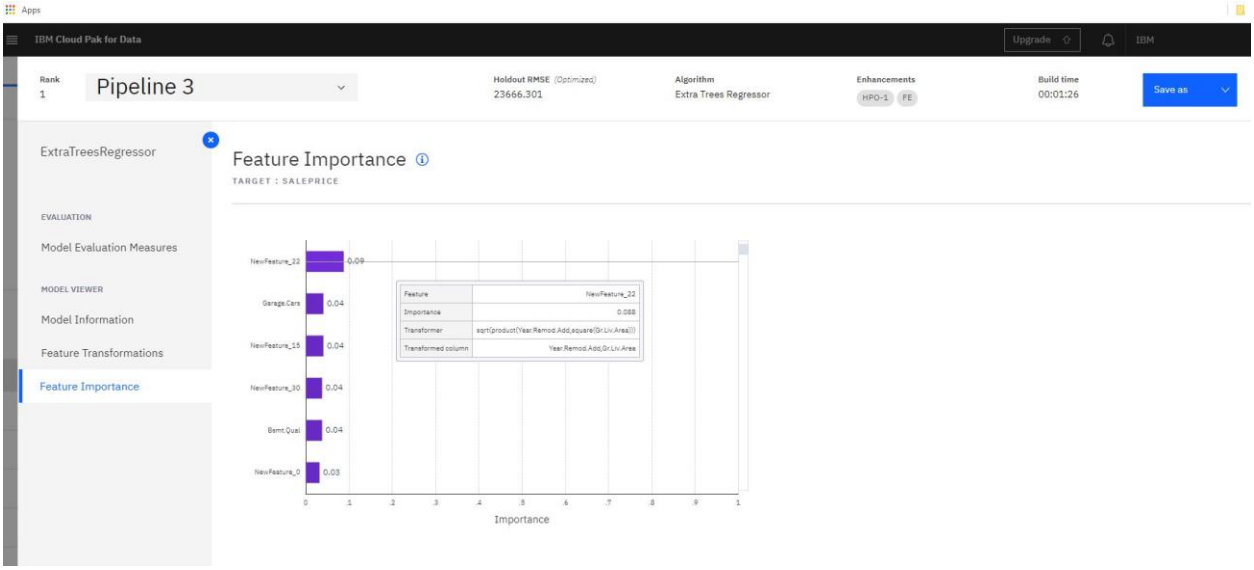

IBM Cloud Pak for Data
Upgrade
IBM
Rank
1
Pipeline 3
Holdout RMSE (Optimized)
23666.301
Algorithm
Extra Trees Regressor
Enhancements
HPO-1
FE
Build time
00:01:26
Save as
ExtraTreesRegressor
EVALUATION
Model Evaluation Measures
MODEL VIEWER
Model Information
Feature Transformations
Model Information
TARGET : SALEPRICE
Label (Target) SalePrice
Number of Features 116
Created At 10/28/2020, 3:24:04 PM

Feature Engineering



Feature Importance




Apps
IBM Cloud Pak for Data
Upgrade
IBM
Rank
1
Pipeline 3
Holdout RMSE (Optimized)
23666.301
Algorithm
Extra Trees Regressor
Enhancements
HPO-1
FE
Build time
00:01:26
Save as
ExtraTreesRegressor
EVALUATION
Model Evaluation Measures
MODEL VIEWER
Model Information
Feature Transformations
Feature Importance
Feature Importance
TARGET : SALEPRICE
NewFeature_22
Garage.Cars
NewFeature_15
NewFeature_10
Bsmt.Qual
NewFeature_0
0.09
0.04
0.04
0.04
0.04
0.03
Feature
NewFeature_22
Importance
0.088
Transformer
sqrt(product(Year.Remod.Add,square(Gr.Liv.Area)))
Transformed column
Year.Remod.Add,Gr.Liv.Area
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
Importance