

## # Optimizing an ML Pipeline in Azure

### ## Overview

This project is part of the Udacity Azure ML Nanodegree.

In this project, we build and optimize an Azure ML pipeline using the Python SDK and a provided Scikit-learn model.

This model is then compared to an Azure AutoML run.

### ## Summary

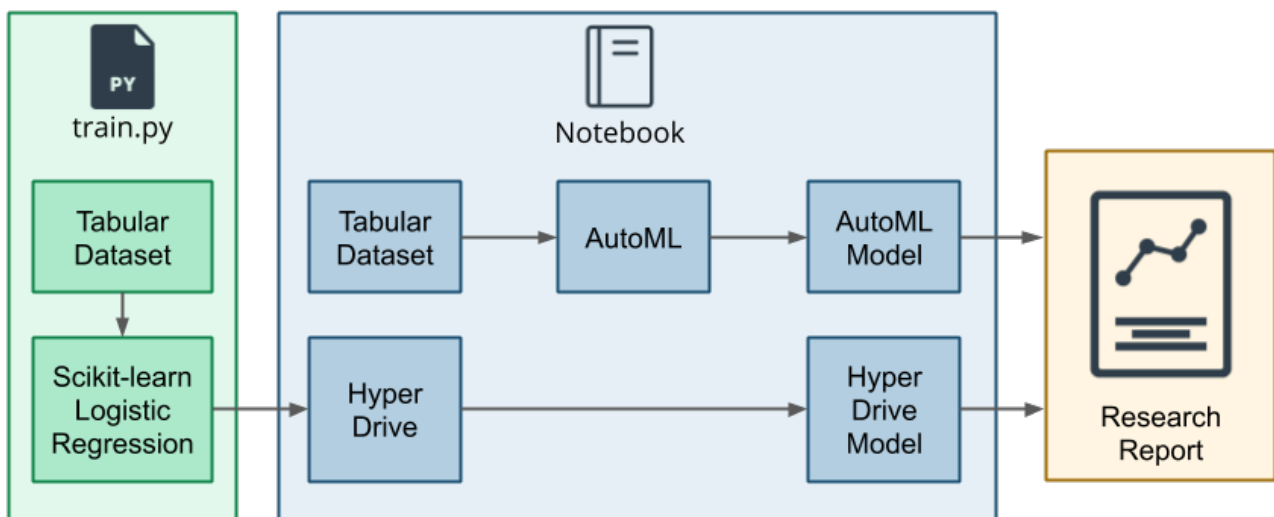
The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

The best performing model is an ensemble model VotingEnsemble produced by the AutoML run. It has an accuracy rate of 91.60% whereas it is 90.5% incase of HyperDrive assisted Scikit-learn LogisticRegression model.

### ## Scikit-learn Pipeline

The main steps and architecture is shown in below diagram.



The pipeline consists of a training script (train.py), a dataset downloaded from Portuguese banking institution, a Scikit-learn Logistic Regression, a HyperDrive for optimizing the hyperparameters. A compute instance is created and a Jupyter Notebook is used to run the training script.

### Benefits of the parameter sampler chosen

The random parameter sampler for HyperDrive supports discrete and continuous hyperparameters, as well as early termination of low-performance runs. It is simple to use, eliminates bias and increases the accuracy of the model.

### Benefits of the early stopping policy chosen

The early termination policy BanditPolicy for HyperDrive automatically terminates poorly performing runs and improves computational efficiency.

### ## AutoML

The AutoML run was executed with below AutoMLConfig settings:

```
automl_config = AutoMLConfig(
    experiment_timeout_minutes=30,
    task='classification',
    primary_metric='accuracy',
    training_data=x,
    label_column_name='y',
    n_cross_validations=2)
```

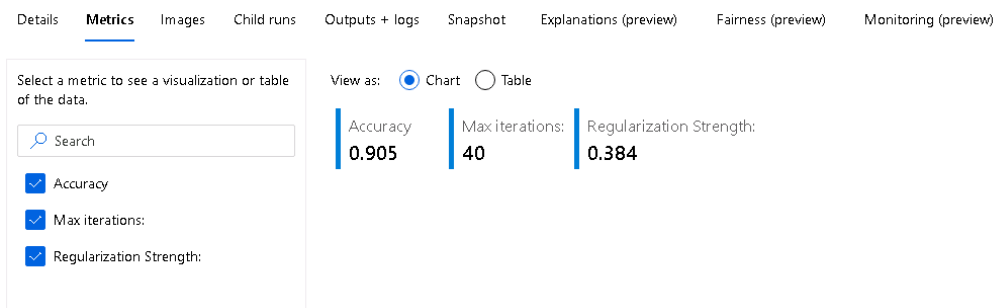
The best model generated from the run was a StackEnsemble model.

## ## Pipeline comparison

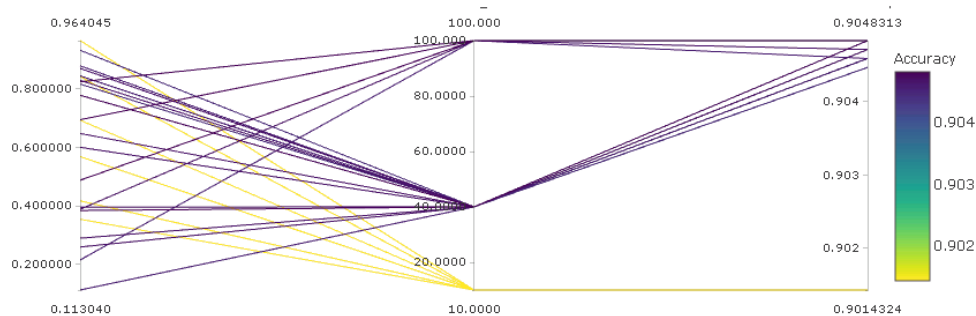
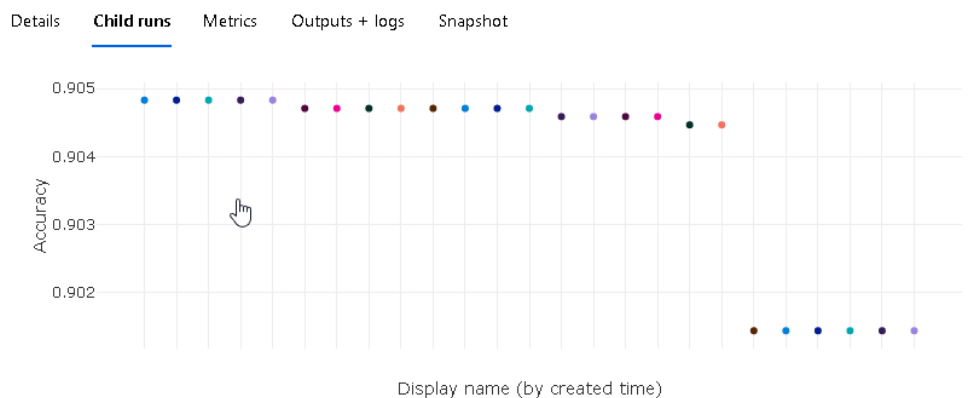
There is a little difference in accuracy.

and also a little difference between them from architecture point of view.

HyperDrive requires, a custom-coded machine learning model whereas AutoML requires selection of few parameters for AutoML config. AutoML model also have a feature for model interpretation.



The HyperDrive assisted Scikit-learn LogicRegression model gives the best accuracy of 90.50% as shown below:



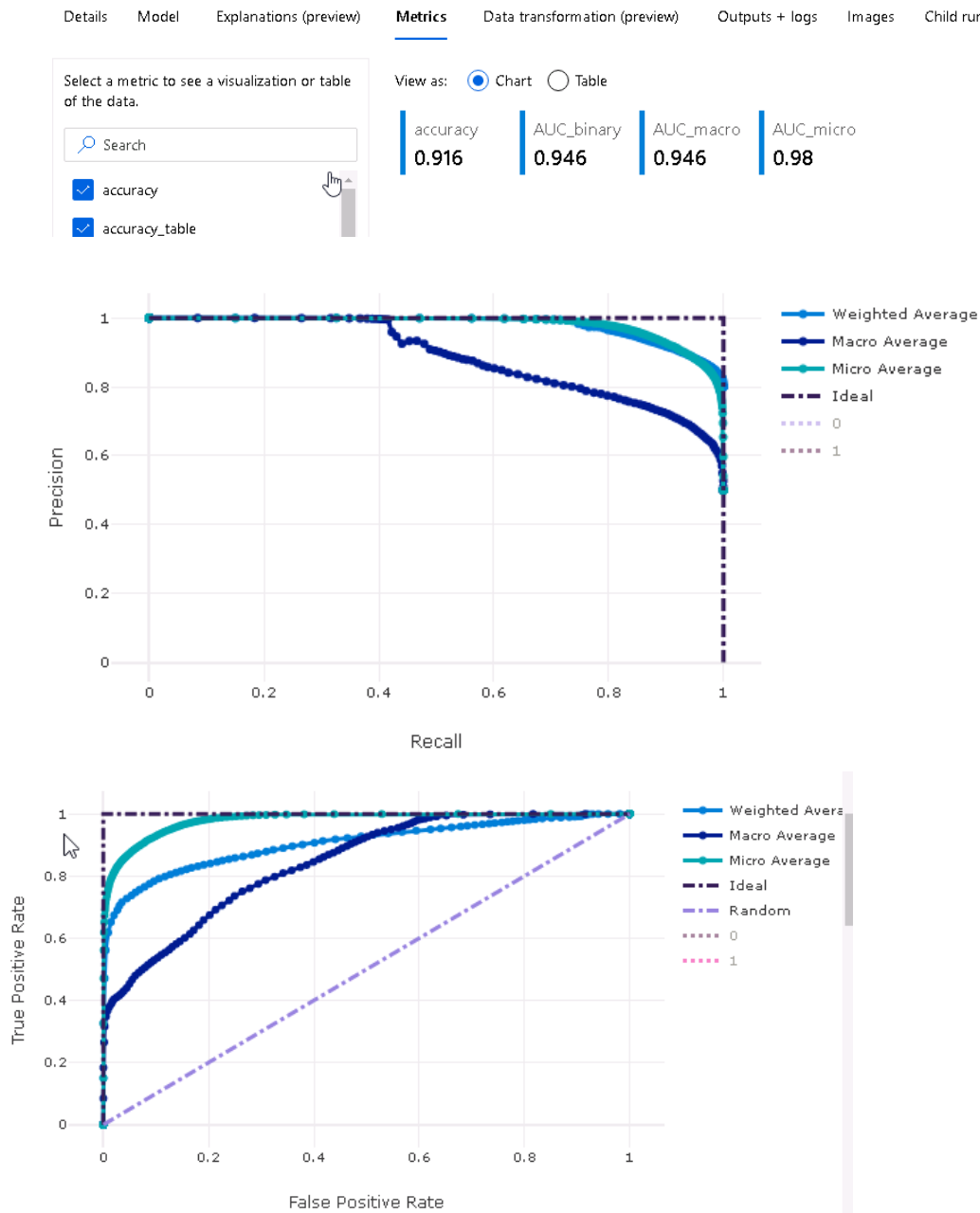
Details **Child runs** Metrics Outputs + logs Snapshot

Showing 1-25 of 30 Child runs

Display name	Status	Accuracy	--C	--max_iter	Parent run ID	Submitted time	Duration	Submitted by	Compute target	Tags
happy_sheep_0ljtk12	Completed	0.90483	0.694248043...	100	HD_d8c9a31e-dd68-4...	Nov 3, 2021 5:20 AM	32s	ODL_User 162649	computeCluster	hyperparameters : ['--C':
hungry_cumin_tbbmh6q	Completed	0.90483	0.384095577...	40	HD_d8c9a31e-dd68-4...	Nov 3, 2021 5:31 AM	1m 23s	ODL_User 162649	computeCluster	hyperparameters : ['--C':
gifted_garage_wrf686d	Completed	0.90483	0.823546982...	100	HD_d8c9a31e-dd68-4...	Nov 3, 2021 5:20 AM	1m 58s	ODL_User 162649	computeCluster	hyperparameters : ['--C':
keen_eagle_gxolv9s9	Completed	0.90483	0.777385636...	40	HD_d8c9a31e-dd68-4...	Nov 3, 2021 5:23 AM	1m 27s	ODL_User 162649	computeCluster	hyperparameters : ['--C':
nice_chicken_5srdarg0	Completed	0.90483	0.487638786...	100	HD_d8c9a31e-dd68-4...	Nov 3, 2021 5:20 AM	1m 59s	ODL_User 162649	computeCluster	hyperparameters : ['--C':
bold_corn_fxtn0hg	Completed	0.90471	0.646980437...	40	HD_d8c9a31e-dd68-4...	Nov 3, 2021 5:27 AM	1m 19s	ODL_User 162649	computeCluster	hyperparameters : ['--C':
sweet_arch_rfhv46fm	Completed	0.90471	0.259708342...	40	HD_d8c9a31e-dd68-4...	Nov 3, 2021 5:29 AM	1m 19s	ODL_User 162649	computeCluster	hyperparameters : ['--C':
upbeat_hand_6wv437d0	Completed	0.90471	0.388753533...	100	HD_d8c9a31e-dd68-4...	Nov 3, 2021 5:20 AM	1m 59s	ODL_User 162649	computeCluster	hyperparameters : ['--C':

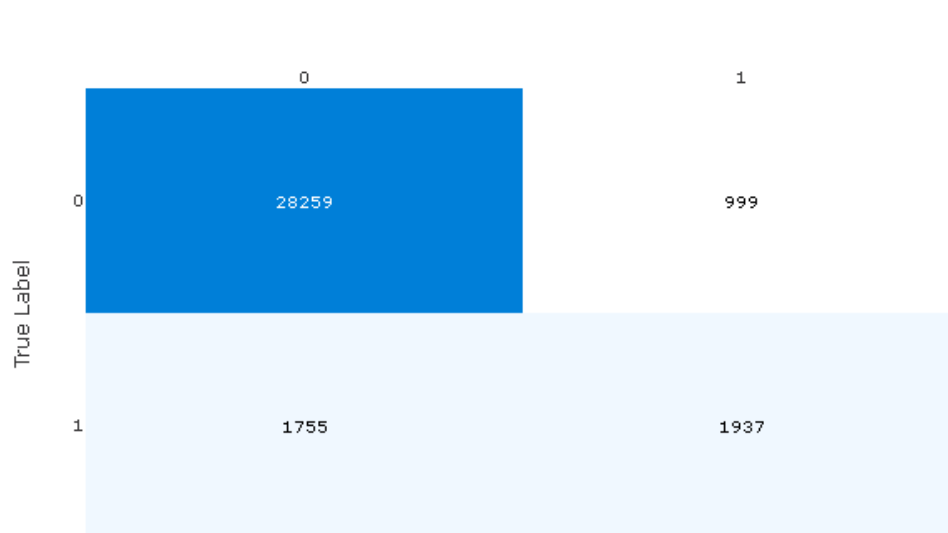
Page size: 25

And the AutoML (VotingEnsemble) model gives the best accuracy of 91.60%:



below are the AutoML generated visual feature based explanation and confusion matrix **(not normalized)**:





### ## Future work

Apply model interpretability of AutoML on more complex and larger datasets, to gain speed and valuable insights in feature engineering, which can in turn be used to refine complex model accuracy

Experiment with different hyperparameter sampling methods like Grid sampling or Bayesian sampling on the Scikit-learn LogisticRegression model or other custom-coded machine learning models.