Week 8: Python for Machine Learning

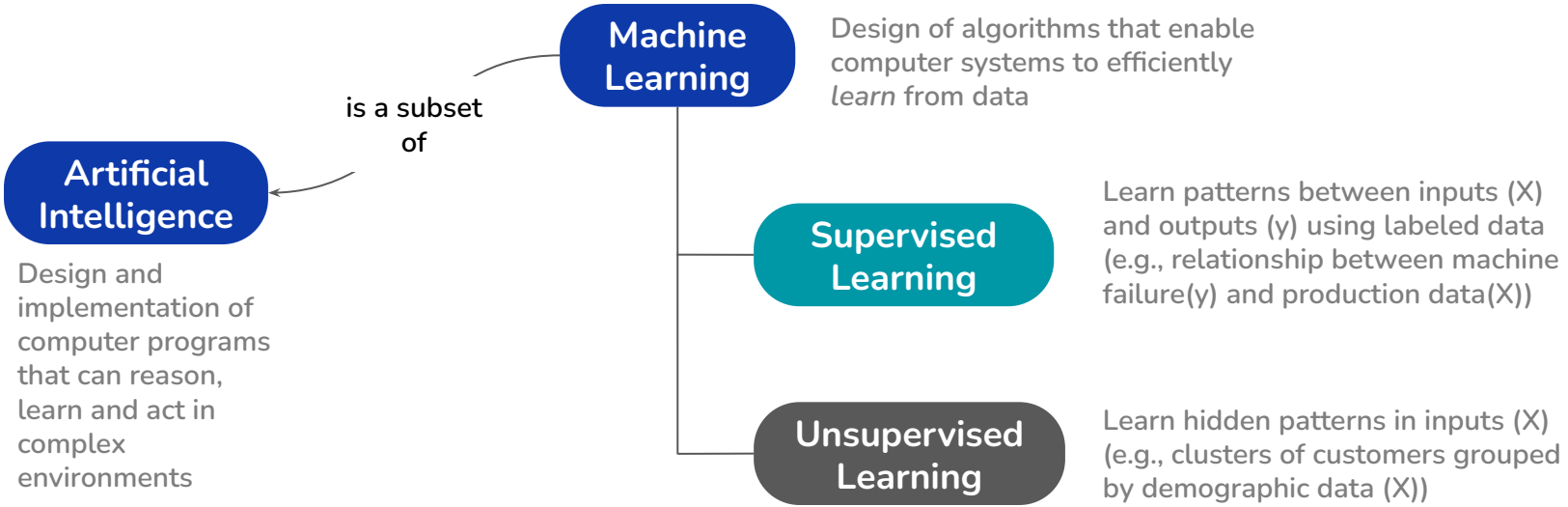
Agenda

In this session, we will discuss:

- Key Ideas of Machine Learning (ML)
- ML Workflow - An Overview
- ML for Regression and Classification Tasks
- Hyperparameter Tuning and Model Evaluation
- Serializing ML Models

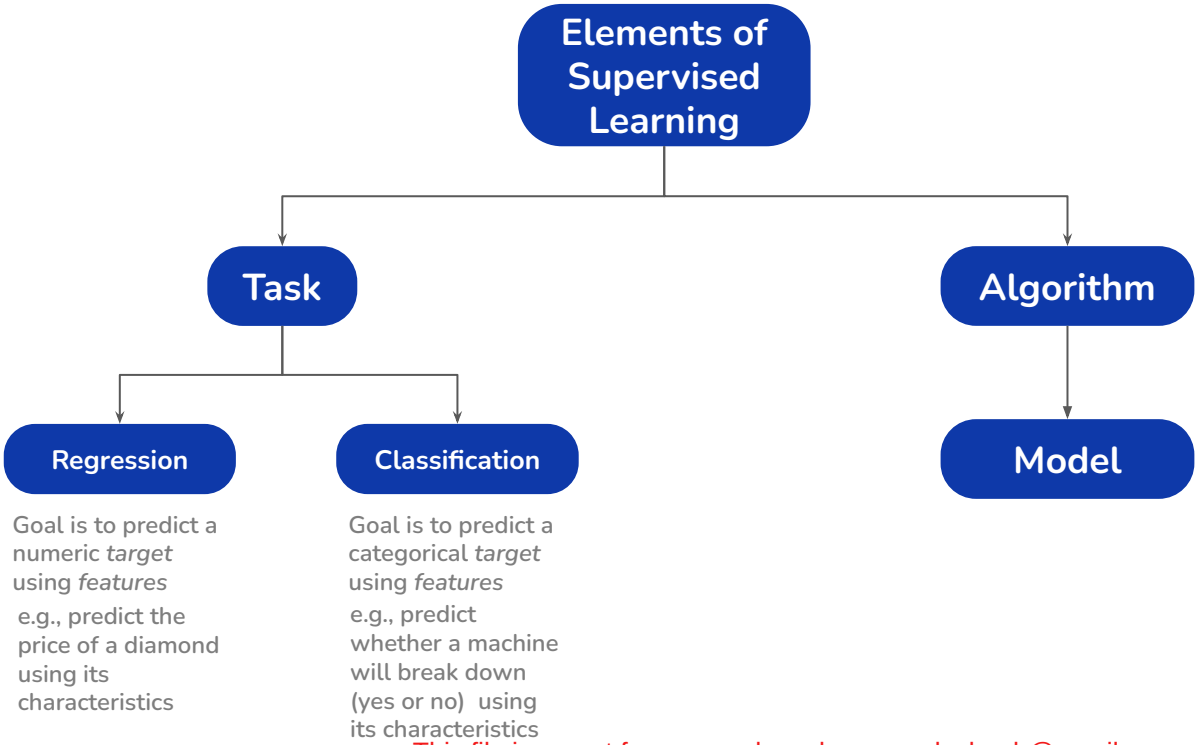
Machine Learning - An Introduction

Terminology

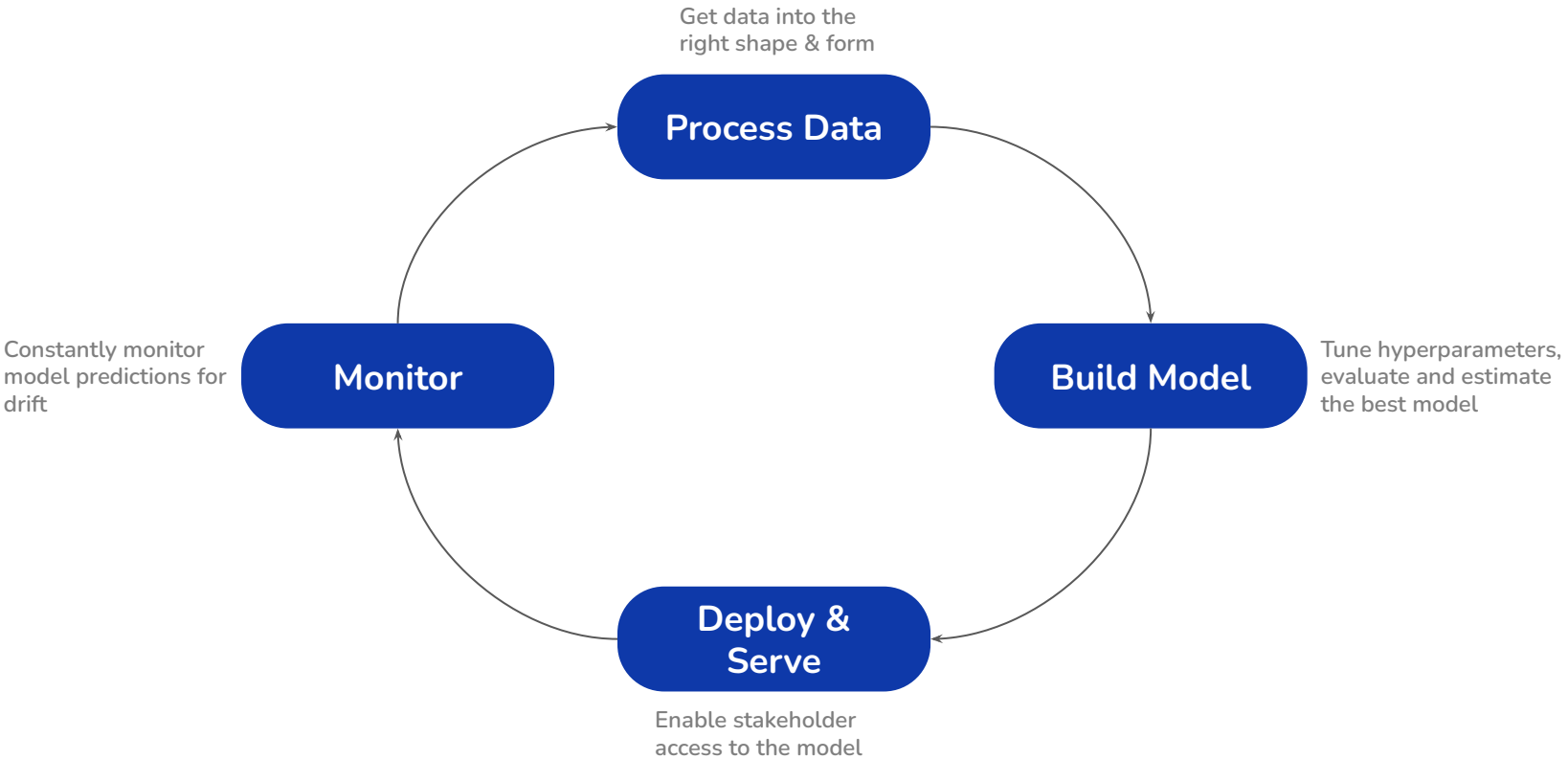


Machine Learning - An Introduction

Terminology

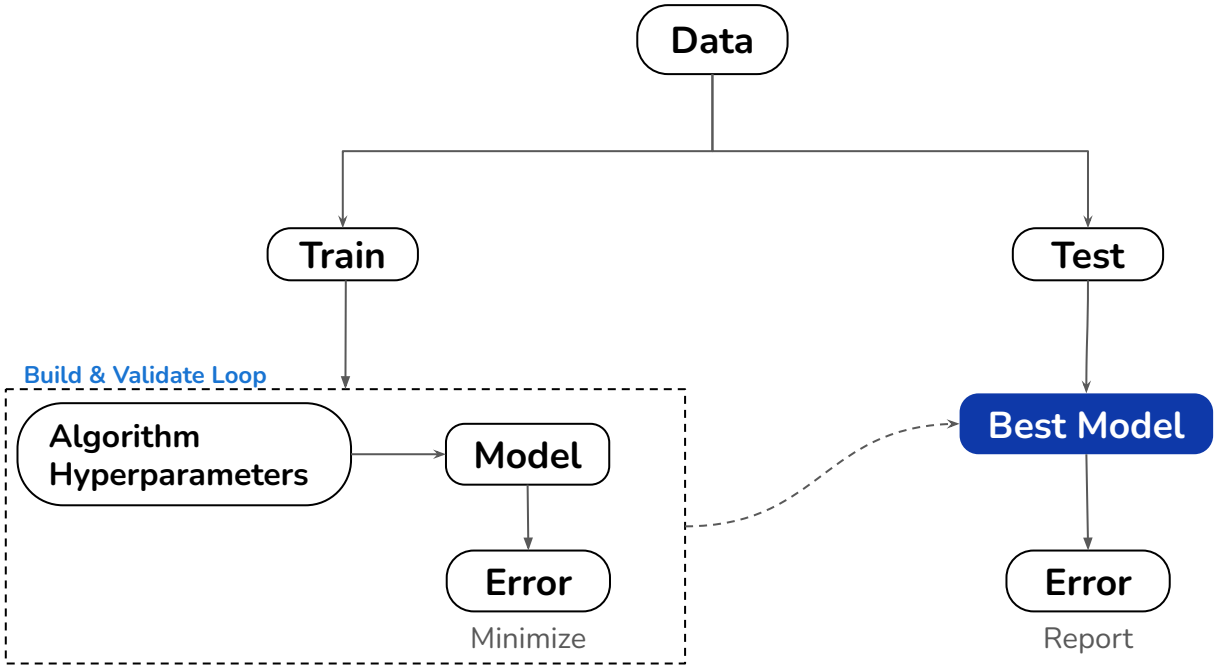


ML Workflow - An Overview



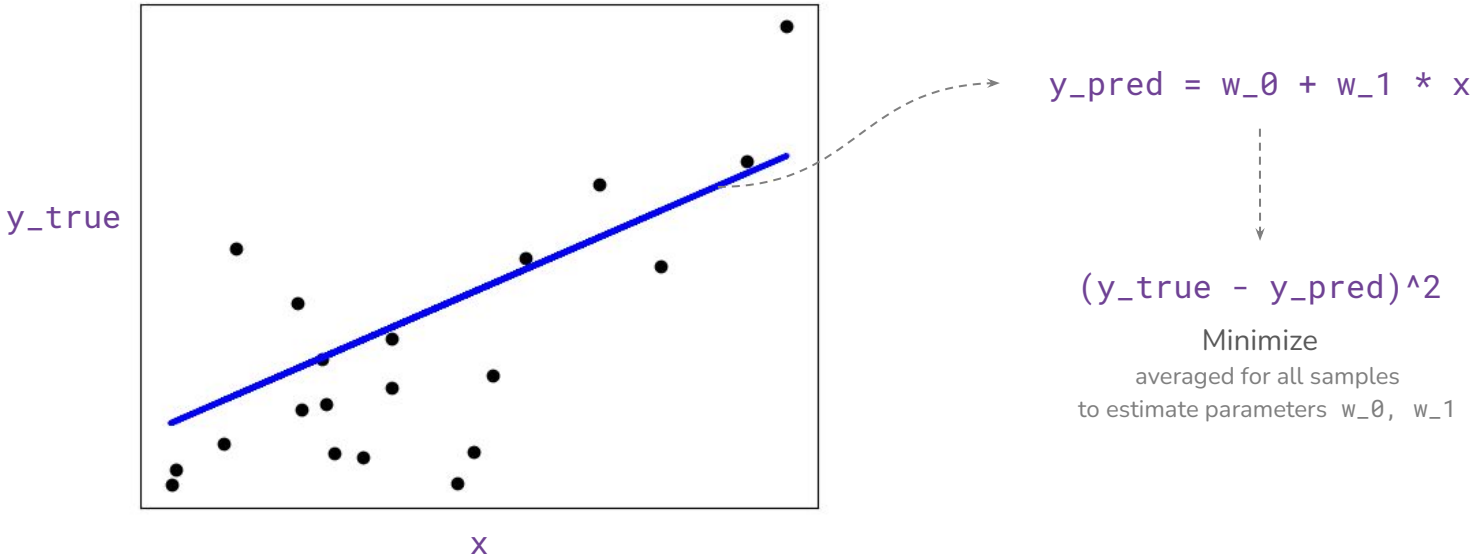
Machine Learning - An Introduction

Supervised Learning Components for Efficient Learning



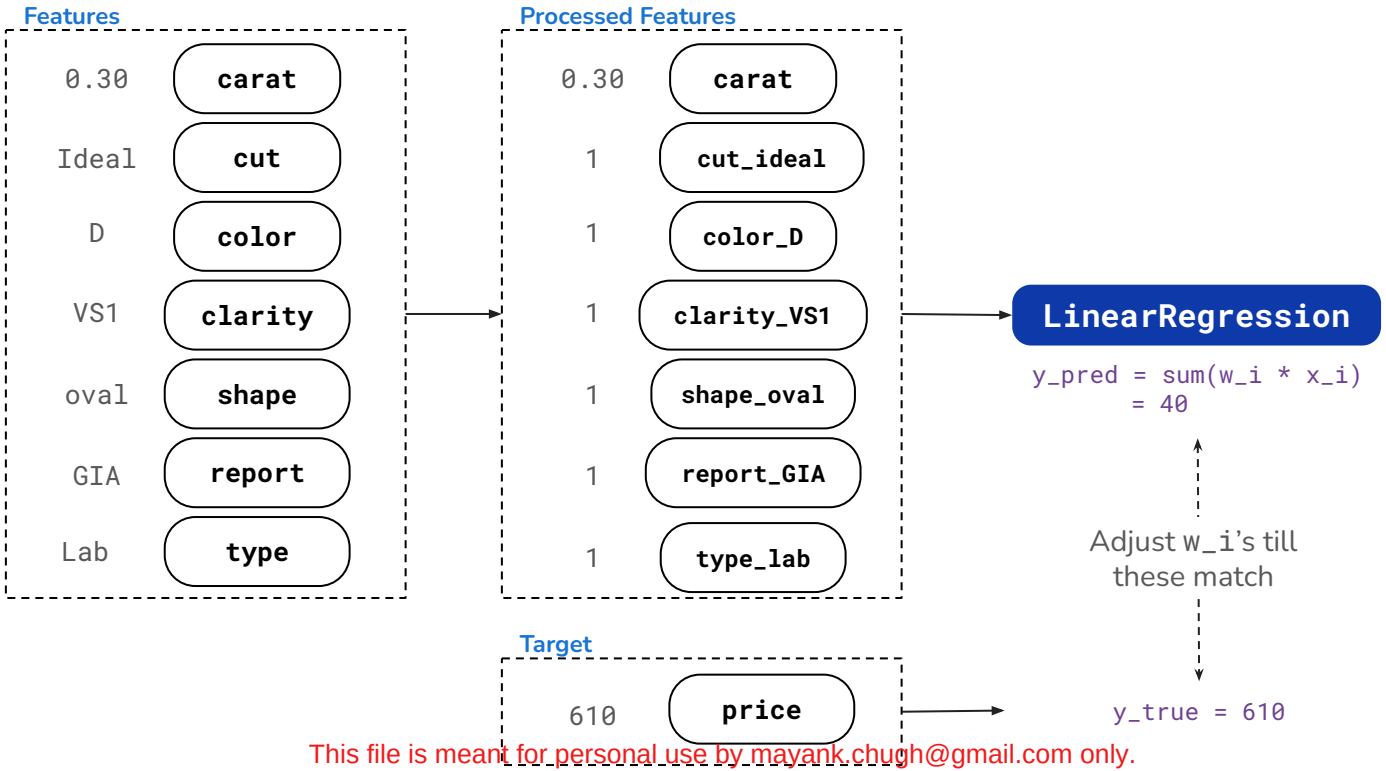
Supervised Learning for Regression Tasks

Algorithm: *Linear Regression estimates a linear model that is the best predictor of the target given the features in the training data. The best predictor is derived by minimizing the mean squared differences between the true targets and the targets predicted by the linear approximation.*



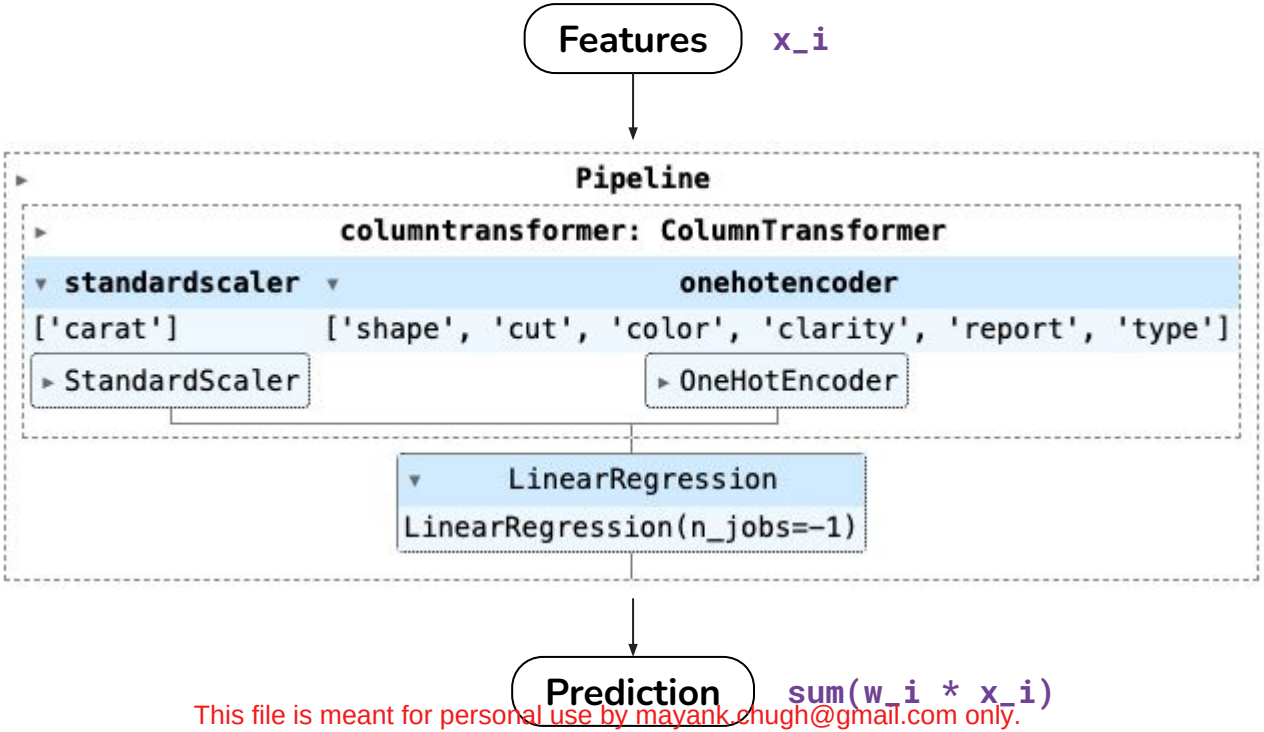
Supervised Learning for Regression Tasks

Example: Diamond Price Prediction ***Model***



Supervised Learning for Regression Tasks

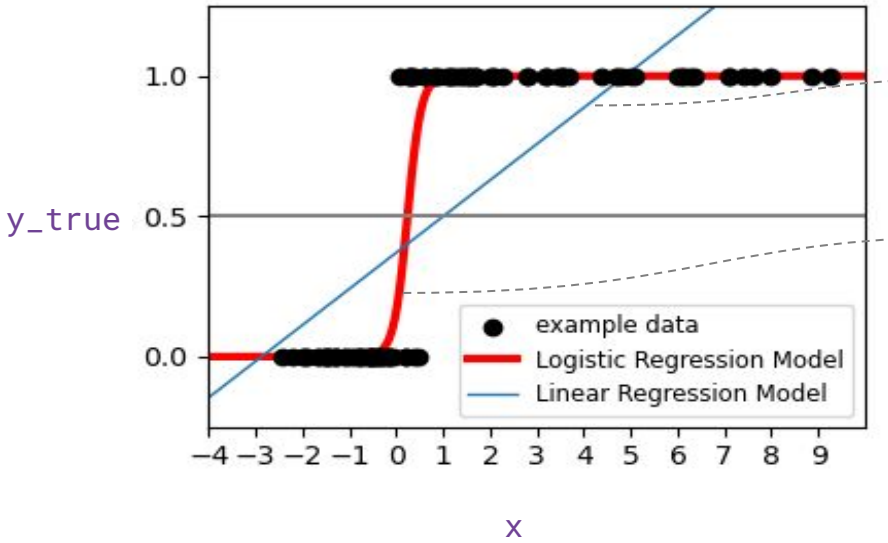
Example: Diamond Price Prediction with scikit-learn



[Notebook]

Supervised Learning for Classification Tasks

Algorithm: *Logistic Regression squishes a linear model through a logistic function to estimate probabilities of the target classes. The best model minimizes the negative log probability predicted by the linear approximation + logistic squisher.*



$s = \text{sum}(w_i * x_i)$

$h(s) = 1/(1+e^{-s})$

$-\log(h(s)) + C * \text{sum}(|w|)$

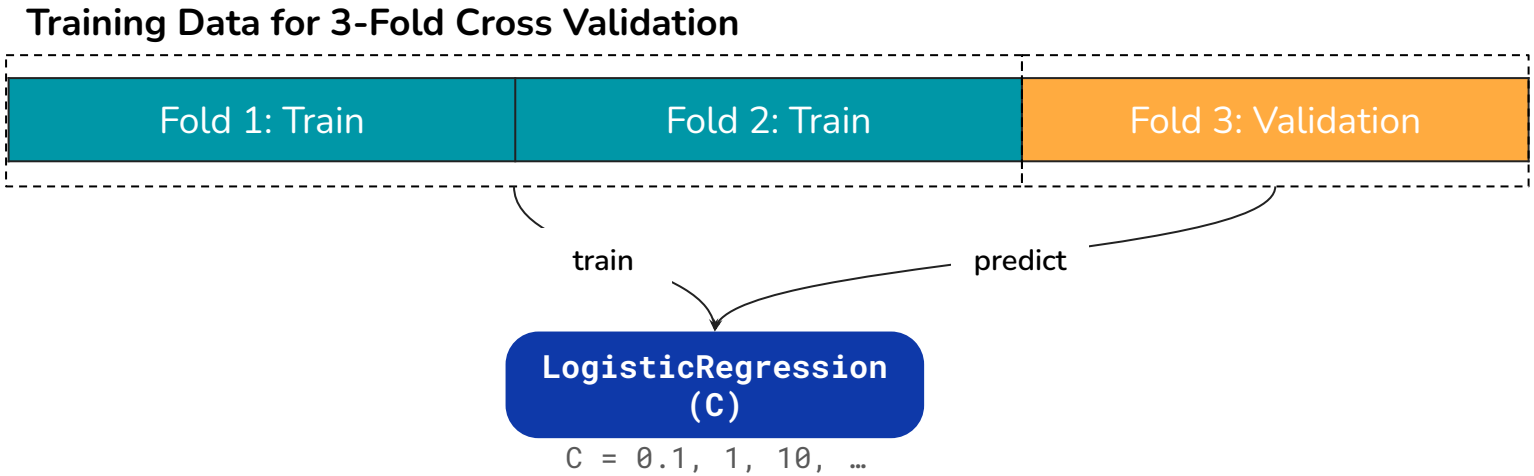
Minimize

averaged for all samples & a value of hyperparameter C

to estimate parameters w_0, w_1

Supervised Learning for Classification Tasks

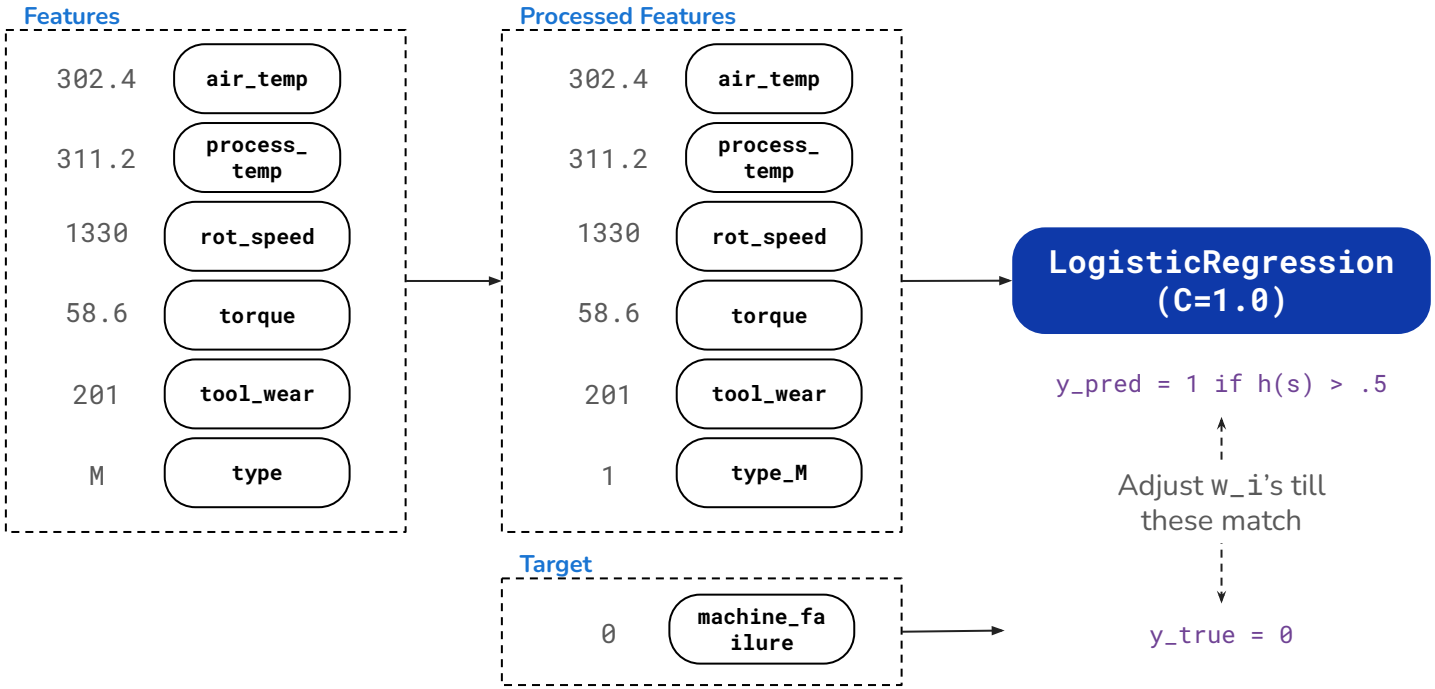
Algorithm: *To estimate the best value of the hyperparameter C, we divide training data into folds and estimate models with different values of C on the training folds. The performance of the model is then measured on the validation fold. The best value of C corresponds to the model with highest validation accuracy. This is called* ***hyperparameter tuning***.



Shuffle folds and estimate best C

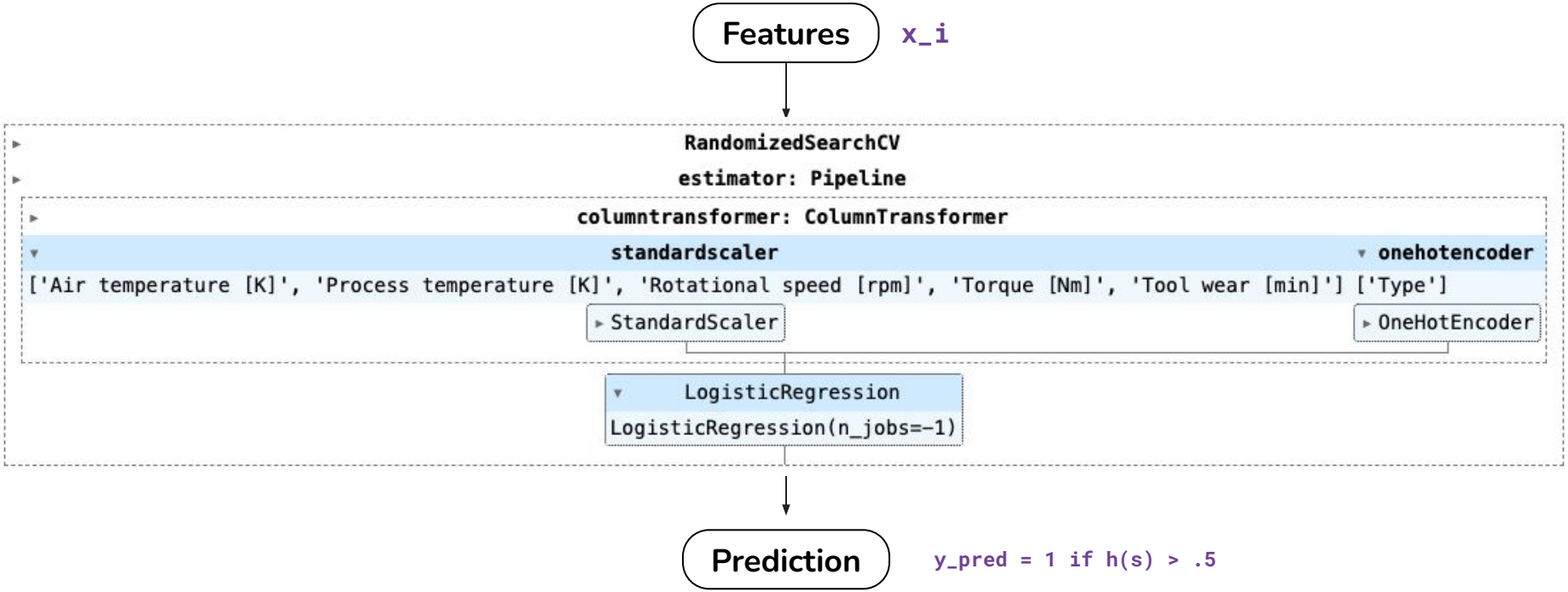
Supervised Learning for Classification Tasks

Example: Machine Failure Prediction ***Model***



Supervised Learning for Classification Tasks

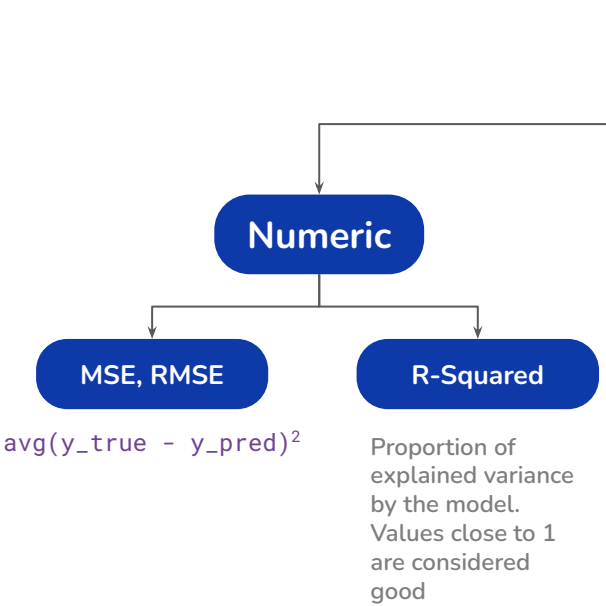
Example: Machine Failure Prediction with scikit-learn



[Notebook]

Model Evaluation

Model evaluation is conducted by comparing predictions and targets on train, validation and test data



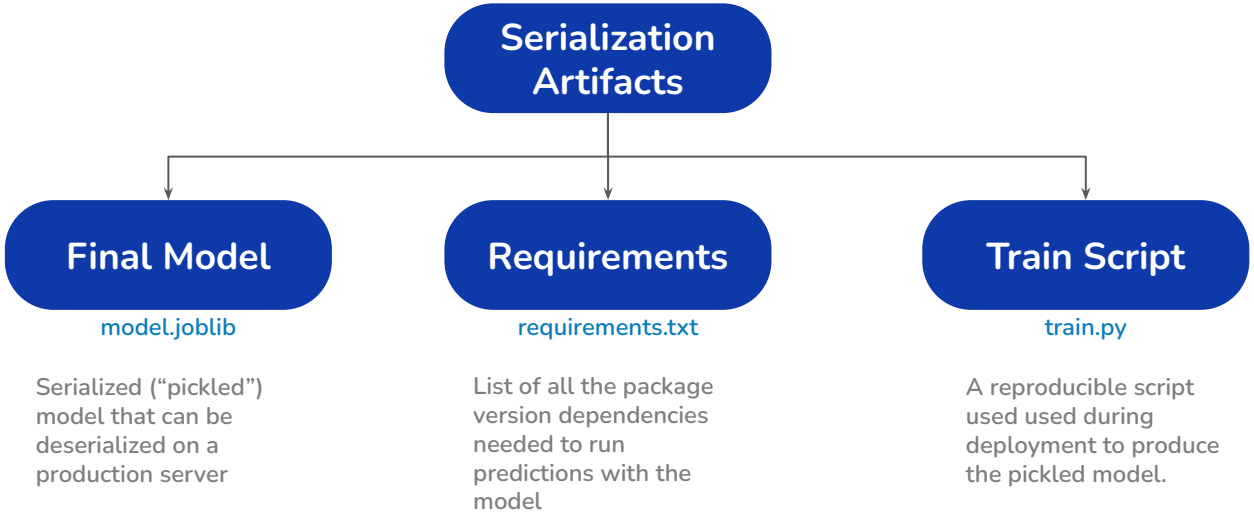
		Actual	
		Positive Label (1)	Negative Label (0)
Predicted	Positive Label (1)	50 TP	10 FP
	Negative Label (0)	5 FN	100

Micro-precision = TP / (TP + FP) = 50 / (50 + 10) = 0.833
Micro-recall = TP / (TP + FN) = 50 / (50 + 5) = 0.909
Micro-F1 score = 2 * (Micro-precision * Micro-recall) / (Micro-precision + Micro-recall) = 2 * (0.833 * 0.909) / (0.833 + 0.909) = 0.870

[Notebook]

Model Serialization

Model serialization is the process of creating persistent, deployable artifacts after model training is complete. Three artifacts are created during serialization - a byte-stream conversion (e.g., binary file) of the model Python object, a requirements file and a training script.



[Notebook]

Summary