PyCaret

An open-source, low-code machine learning library in Python

Pycaret?

- PyCaret is an open-source, low-code machine learning library in Python that automates machine learning workflows.
- PyCaret is essentially a Python wrapper around several machine learning libraries and frameworks, such as scikit-learn, XGBoost, LightGBM, CatBoost, spaCy, Optuna, Hyperopt, Ray, and a few more.
- inspired by the emerging role of citizen data scientists, a term first used by Gartner.

Pycaret?

Data Preparation

Model Training

Hyperparameter Tuning

Analysis & Interpretability

Model Selection

Experiment Logging

PyCaret is ideal for:

- Experienced Data Scientists who want to increase productivity.
- Citizen Data Scientists who prefer a low code machine learning solution.
- Data Science Professionals who want to build rapid prototypes.
- Data Science and Machine Learning students and enthusiasts.

data analysis with Pycaret

EDA

data preprocessing (set up)

Model Training

Optimize

Analysis

Deploy

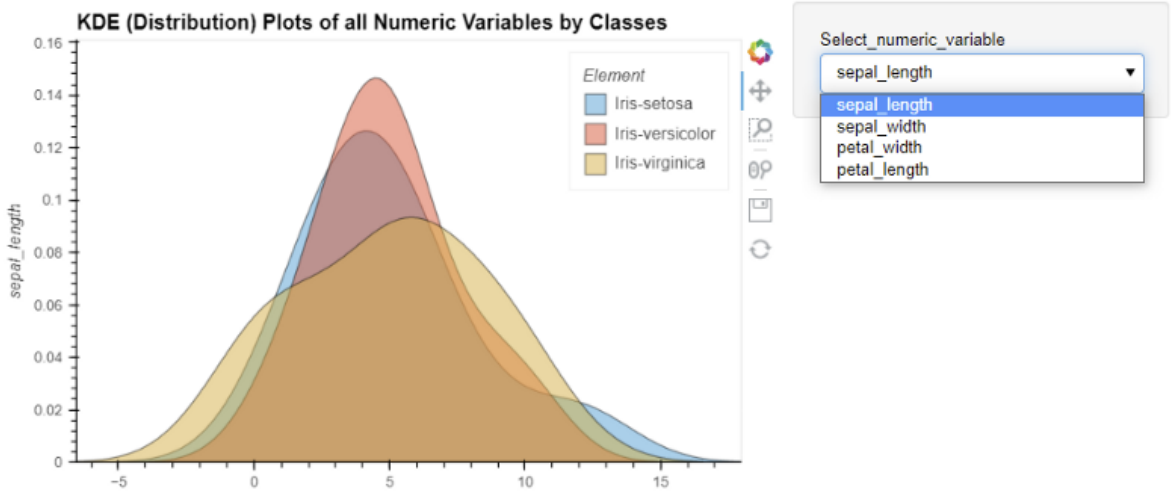
📊 Exploratory Data Analysis (EDA)

This function will generate automated EDA using the AutoViz integration.

```
# load dataset
from pycaret.datasets import get_data
data = get_data('iris')

# init setup
from pycaret.classification import *
s = setup(data, target = 'species', session_id = 123)

# generate EDA
eda()
```



- AutoViz를 통해 손쉬운 EDA가능
- 시각화를 짧은 코드로 가능

data analysis with Pycaret

EDA
data preprocessing (set up)
Model Training
Optimize
Analysis
Deploy

```
from pycaret.classification import *
s = setup(data, target = 'Class variable')
```

- data : 학습에 사용할 데이터
- Target : 데이터 중 라벨 컬럼 이름

Processing:

Initiated	10:59:20
Status	Preprocessing Data

Following data types have been inferred automatically, if they are correct press enter to continue or type 'quit' otherwise.

	Data Type
Number of times pregnant	Categorical
Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Numeric
Diastolic blood pressure (mm Hg)	Numeric
Triceps skin fold thickness (mm)	Numeric
2-Hour serum insulin (mu U/ml)	Numeric
Body mass index (weight in kg/(height in m)^2)	Numeric
Diabetes pedigree function	Numeric
Age (years)	Numeric
Class variable	Label

- 파라미터들로 파이캐럿의 실험을 초기화하는 과정
- 데이터와 타겟 입력하면 자동으로 모든 변수의 데이터 타입 추론

data analysis with Pycaret

EDA
data preprocessing (set up)
Model Training
Optimize
Analysis
Deploy

```
best = compare_models()
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
catboost	CatBoost Classifier	0.7767	0.8309	0.6056	0.7114	0.6413	0.4823	0.4950	1.3870
lr	Logistic Regression	0.7564	0.8043	0.5056	0.6941	0.5786	0.4145	0.4285	1.1230
gbc	Gradient Boosting Classifier	0.7562	0.8239	0.5667	0.6731	0.6031	0.4314	0.4431	0.0900
ada	Ada Boost Classifier	0.7526	0.8016	0.5889	0.6524	0.6091	0.4310	0.4394	0.0800
lightgbm	Light Gradient Boosting Machine	0.7524	0.8028	0.5778	0.6614	0.6086	0.4299	0.4381	0.1430
rf	Random Forest Classifier	0.7488	0.8035	0.5111	0.6849	0.5740	0.4023	0.4182	0.2350
ridge	Ridge Classifier	0.7452	0.0000	0.4722	0.6844	0.5492	0.3816	0.3997	0.0150
lda	Linear Discriminant Analysis	0.7452	0.7912	0.4833	0.6783	0.5563	0.3859	0.4017	0.0130
xgboost	Extreme Gradient Boosting	0.7449	0.7896	0.5722	0.6442	0.5984	0.4140	0.4207	0.2640
knn	K Neighbors Classifier	0.7153	0.7261	0.5111	0.5962	0.5405	0.3379	0.3467	0.0220
et	Extra Trees Classifier	0.7134	0.7573	0.4333	0.6079	0.4968	0.3072	0.3204	0.1810
dt	Decision Tree Classifier	0.7075	0.6741	0.5722	0.5635	0.5630	0.3445	0.3481	0.0130
nb	Naive Bayes	0.6817	0.7064	0.2389	0.5527	0.3288	0.1657	0.1905	0.0110
svm	SVM - Linear Kernel	0.6015	0.0000	0.3611	0.3419	0.3251	0.0851	0.0924	0.0170
qda	Quadratic Discriminant Analysis	0.5759	0.5889	0.4833	0.4062	0.3705	0.1011	0.1281	0.0180

- 다양한 모델을 다양한 평가지표를 이용해 비교 가능
- 성능 비교에 사용되는 지표 및 걸리는 시간까지 제공

data analysis with Pycaret

EDA
data preprocessing (set up)
Model Training
Optimize
Analysis
Deploy

	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.4596	0.3575	0.5979	0.6187	-0.0000	0.1100
1	0.5786	0.6307	0.7941	0.3693	-0.0000	0.2108
2	0.6451	0.7284	0.8535	0.5144	-0.0000	0.1818
3	0.9047	1.4243	1.1934	0.1788	-0.0000	0.4036
4	0.7391	0.8451	0.9193	0.4749	-0.0000	0.2118
5	0.5919	0.8052	0.8973	0.4632	-0.0000	0.2440
6	0.3923	0.2809	0.5300	0.7146	-0.0000	0.1642
7	0.5588	0.9039	0.9507	0.2388	-0.0000	0.2776
8	0.3440	0.2470	0.4970	0.7118	-0.0000	0.1335
9	0.3991	0.3054	0.5526	0.7229	-0.0000	0.2179
Mean	0.5613	0.6528	0.7786	0.5008	0.0000	0.2155
SD	0.1644	0.3515	0.2159	0.1859	0.0000	0.0786

- 모델 생성 : 모델의 이름을 생성 가능. 디폴트로 10개의 fold를 생성 후 평가
- 각종 평가지표의 평균과 표준편차를 모델 생성의 결과로 보여줌

data analysis with Pycaret

EDA
data preprocessing (set up)
Model Training
Optimize
Analysis
Deploy

```
GradientBoostingRegressor(alpha=0.9, ccp_alpha=0.0, criterion='friedman_mse',
                          init=None, learning_rate=0.1, loss='ls', max_depth=3,
                          max_features=None, max_leaf_nodes=None,
                          min_impurity_decrease=0.0, min_impurity_split=None,
                          min_samples_leaf=1, min_samples_split=2,
                          min_weight_fraction_leaf=0.0, n_estimators=100,
                          n_iter_no_change=None, presort='deprecated',
                          random_state=123, subsample=1.0, tol=0.0001,
                          validation_fraction=0.1, verbose=0, warm_start=False)
```

- 튜닝에 사용된 모델 하이퍼 파라미터를 print()로 확인 가능
- Iteration 횟수 조정, 커스텀 메트릭 사용가능, 그리드 서치 가능

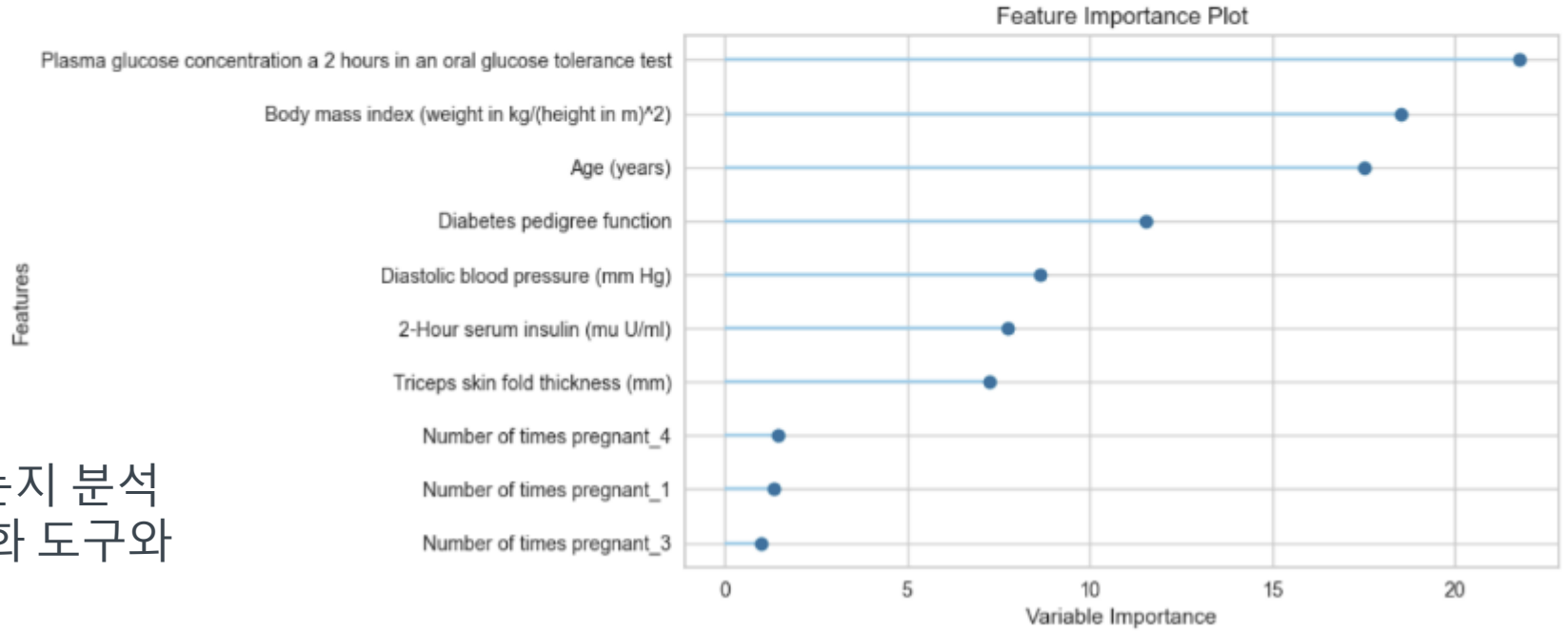
• 그리드서치 : 다양한 모델 하이퍼파라미터의 조합을 순서대로 실험해보고 가장 높은 성능을 보이는 하이퍼 파라미터 조합을 갖는 탐색 방법

data analysis with Pycaret

EDA
data preprocessing (set up)
Model Training
Optimize
Analysis
Deploy

```
evaluate_model(best)
```

Plot Type:					
	Hyperparameters	AUC	Confusion Matrix	Threshold	Precision Recall
	Prediction Error	Class Report	Feature Selection	Learning Curve	Manifold Learning
	Calibration Curve	Validation Curve	Dimensions	Feature Importance	Feature Importance...
	Decision Boundary	Lift Chart	Gain Chart	Decision Tree	KS Statistic Plot



- 모델 튜닝 후 모델이 잘 만들어졌는지 분석
- 분류, 회귀 등을 위한 다양한 시각화 도구와 커스터마이징 속성 제공

data analysis with Pycaret

EDA
data preprocessing (set up)
Model Training
Optimize
Analysis
Deploy

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	CatBoost Classifier	0.7835	0.8553	0.7045	0.7209	0.7126	0.5391	0.5392

Age (years)	Number of times pregnant_0	Number of times pregnant_1	Number of times pregnant_10	...	Number of times pregnant_3	Number of times pregnant_4	Number of times pregnant_5	Number of times pregnant_6	Number of times pregnant_7	Number of times pregnant_8	Number of times pregnant_9	Class variable	Label	Score
51.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1	1	0.9382
25.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0	0.7577
35.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	1	0.6828
25.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0	0.9621
23.0	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0	0.9245
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
23.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0	0.8802
26.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0	1	0.9338
21.0	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0	0.9844
28.0	0.0	1.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0	0.8812
33.0	0.0	0.0	0.0	...	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0	0	0.7411

- 분류 모델의 경우 예측시 결과 뿐 아니라 예측 확률도 확인 가능
- 데이터 드리프트를 위한 모니터링 리포트 출력가능

* 데이터 드리프트 : 입력되는 데이터의 특성이 변경되어 모델의 성능 저하를 초래하는 현상

EOD.