Data Drift & Model Drift Detection

Data Drift

If there is changes in the data, we normally call it as Data Drift or Data Shift. A Data Drift can also refer to

- changes in the input data
- changes in the values of the features used to define or predict a target label.
- changes in the properties of the independent variable

Model Drift

This refers to changes in the performance of the model over time. It is the deterioration of models over time in the case of accuracy and prediction. ML Models do not live in a static environment hence they will deteriorate or decay over time.

Deepchecks

- Useful for detecting data drift,data integrity,model performance,etc
- pip install deepchecks

```
# Load Packages
import pandas as pd
import numpy as np
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
```

```
#### Build A Model
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
```

```
# load data
df = pd.read_csv("data/bank-additional-full_encoded.csv")
```

```
df.head()
```

	age	job	marital	education	default	housing	loan	contact	month	day_
0	56	0	0	0	0	0	0	0	0	
1	57	1	0	1	1	0	0	0	0	
2	37	1	0	1	0	1	0	0	0	
3	40	2	0	2	0	0	0	0	0	
4	56	1	0	1	0	0	1	0	0	

5 rows × 21 columns

```
# Features & Labels
Xfeatures = df.drop('y',axis=1)
# Select last column of dataframe as a dataframe object
ylabels = df.iloc[: , -1:]
```

```
Xfeatures.columns
```

```
Index(['age', 'job', 'marital', 'education', 'default', 'housing', 'loan',
       'contact', 'month', 'day_of_week', 'duration', 'campaign', 'pdays',
       'previous', 'poutcome', 'emp.var.rate', 'cons.price.idx',
       'cons.conf.idx', 'euribor3m', 'nr.employed'],
      dtype='object')
```

```
# Split Dataset
x_train,x_test,y_train,y_test = train_test_split(Xfeatures,ylabels,test_size=0.3,random_state=7)
```

Requirements

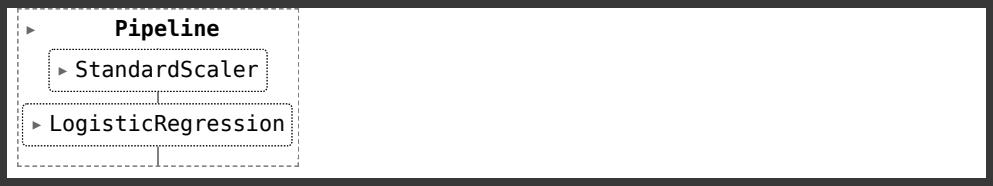
- Datasets
  - train,test data
- Model

Components

- Suites
- Checks
- Dataset

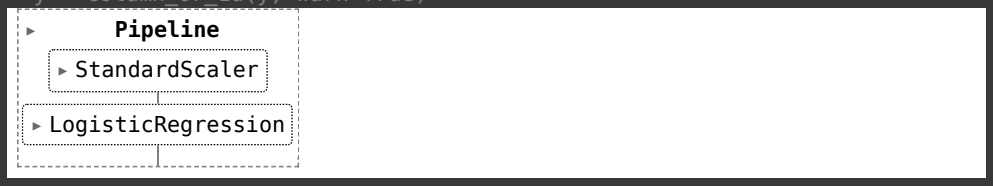
```
# Build the Model
pipe_lr = Pipeline(steps=[('sc',StandardScaler()),('lr',LogisticRegression())])
```

```
pipe_lr
```




```
# Train to Fit
pipe_lr.fit(x_train,y_train)
```

```
/usr/local/lib/python3.9/dist-packages/sklearn/utils/validation.py:1143: Da
  y = column_or_1d(y, warn=True)
```




```
# Accuarcy
pipe_lr.score(x_test,y_test)
```

```
0.9105770008901837
```

```
!pip install deepchecks
```

```
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.9/dist-packages (from matplot
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.9/dist-packages (from matplo
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.9/dist-packages (from mat
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.9/dist-packages (from matplo
Requirement already satisfied: importlib-resources>=3.2.0 in /usr/local/lib/python3.9/dist-packages (fr
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.9/dist-packages (from pandas>=1.1
Requirement already satisfied: tenacity>=6.2.0 in /usr/local/lib/python3.9/dist-packages (from plotly>=
Requirement already satisfied: python-utils in /usr/local/lib/python3.9/dist-packages (from PyNomaly>=0
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.9/dist-packages (from sci
Requirement already satisfied: joblib>=1.1.1 in /usr/local/lib/python3.9/dist-packages (from scikit-lea
Requirement already satisfied: zipp>=3.1.0 in /usr/local/lib/python3.9/dist-packages (from importlib-re
Requirement already satisfied: parso<0.9.0,>=0.8.0 in /usr/local/lib/python3.9/dist-packages (from jedi
Requirement already satisfied: six in /usr/local/lib/python3.9/dist-packages (from patsy>=0.5.1->categ
Requirement already satisfied: ptyprocess>=0.5 in /usr/local/lib/python3.9/dist-packages (from pexpect>
Requirement already satisfied: wcwidth in /usr/local/lib/python3.9/dist-packages (from prompt-toolkit!=
Requirement already satisfied: notebook>=4.4.1 in /usr/local/lib/python3.9/dist-packages (from widgetsn
Requirement already satisfied: jupyter-core>=4.6.0 in /usr/local/lib/python3.9/dist-packages (from jupy
Requirement already satisfied: platformdirs>=2.5 in /usr/local/lib/python3.9/dist-packages (from jupyte
Requirement already satisfied: nbconvert in /usr/local/lib/python3.9/dist-packages (from notebook>=4.4
Requirement already satisfied: prometheus-client in /usr/local/lib/python3.9/dist-packages (from noteb
Requirement already satisfied: argon2-cffi in /usr/local/lib/python3.9/dist-packages (from notebook>=4
Requirement already satisfied: jinja2 in /usr/local/lib/python3.9/dist-packages (from notebook>=4.4.1->
Requirement already satisfied: nbformat in /usr/local/lib/python3.9/dist-packages (from notebook>=4.4.1
Requirement already satisfied: terminado>=0.8.3 in /usr/local/lib/python3.9/dist-packages (from noteboo
Requirement already satisfied: Send2Trash>=1.5.0 in /usr/local/lib/python3.9/dist-packages (from noteb
Requirement already satisfied: argon2-cffi-bindings in /usr/local/lib/python3.9/dist-packages (from arg
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.9/dist-packages (from jinja2->
Requirement already satisfied: bleach in /usr/local/lib/python3.9/dist-packages (from nbconvert->noteb
Requirement already satisfied: mistune<2,>=0.8.1 in /usr/local/lib/python3.9/dist-packages (from nbconv
Requirement already satisfied: jupyterlab-pygments in /usr/local/lib/python3.9/dist-packages (from nbco
Requirement already satisfied: tinycss2 in /usr/local/lib/python3.9/dist-packages (from nbconvert->note
Requirement already satisfied: entrypoints>=0.2.2 in /usr/local/lib/python3.9/dist-packages (from nbcon
Requirement already satisfied: defusedxml in /usr/local/lib/python3.9/dist-packages (from nbconvert->nb
Requirement already satisfied: pandocfilters>=1.4.1 in /usr/local/lib/python3.9/dist-packages (from nbc
Requirement already satisfied: lxml in /usr/local/lib/python3.9/dist-packages (from nbconvert->notebook
Requirement already satisfied: nbclient>=0.5.0 in /usr/local/lib/python3.9/dist-packages (from nbconver
Requirement already satisfied: jsonschema>=2.6 in /usr/local/lib/python3.9/dist-packages (from nbformat
Requirement already satisfied: fastjsonschema in /usr/local/lib/python3.9/dist-packages (from nbformat-
Requirement already satisfied: attrs>=17.4.0 in /usr/local/lib/python3.9/dist-packages (from jsonschema
Requirement already satisfied: pyrsistent!=0.17.0,!=0.17.1,!=0.17.2,>=0.14.0 in /usr/local/lib/python3
Requirement already satisfied: cffi>=1.0.1 in /usr/local/lib/python3.9/dist-packages (from argon2-cffi-
Requirement already satisfied: webencodings in /usr/local/lib/python3.9/dist-packages (from bleach->nbc
Requirement already satisfied: pycparser in /usr/local/lib/python3.9/dist-packages (from cffi>=1.0.1->a
Building wheels for collected packages: PyNomaly
  Building wheel for PyNomaly (setup.py) ... done
  Created wheel for PyNomaly: filename=PyNomaly-0.3.3-py3-none-any.whl size=8497 sha256=a0f11d9d5d53a0c
  Stored in directory: /root/.cache/pip/wheels/72/c2/79/a077f8b4ef4549086016b0202e61193518262377b6d2331
Successfully built PyNomaly
Installing collected packages: PyNomaly, jedi, category-encoders, deepchecks
```

Using Deepchecks for Offline ML Data Drift Detection

```
import deepchecks
```

```
# Method
dir(deepchecks)
```

```
['BaseCheck',
 'BaseSuite',
 'CheckFailure',
 'CheckResult',
 'Condition',
 'ConditionCategory',
 'ConditionResult',
 'Context',
 'Dataset',
 'ModelComparisonCheck',
 'ModelComparisonSuite',
```

```
 'ModelOnlyBaseCheck',
 'ModelOnlyCheck',
 'SingleDatasetBaseCheck',
 'SingleDatasetCheck',
 'Suite',
 'SuiteResult',
 'TrainTestBaseCheck',
 'TrainTestCheck',
 '_SubstituteModule',
 '__all__',
 '__builtins__',
 '__cached__',
 '__doc__',
 '__file__',
 '__loader__',
 '__name__',
 '__original_module__',
 '__package__',
 '__path__',
 '__spec__',
 '__version__',
 '__warningregistry__',
 '_init_module_attrs',
 'analytics',
 'core',
 'get_verbosity',
 'is_notebook',
 'matplotlib',
 'os',
 'pio',
 'pio_backends',
 'set_verbosity',
 'sys',
 'tabular',
 'types',
 'utils',
 'validate_latest_version',
 'version',
 'warnings']
```

Full Suite

- Data Drift Detection
- Model Performance /Confidence
- Data Integrity Check
- Label Ambuiguity
- Other checks

```
from deepchecks.suites import full_suite
```

```
/usr/local/lib/python3.9/dist-packages/deepchecks/suites.py:21: DeprecationWarning:

Ability to import tabular suites from the `deepchecks.suites` is deprecated, please import from `deepche
```

```
# Create the Dataset Objects
ds_train = deepchecks.Dataset(df=x_train,label=y_train,cat_features=[])
ds_test = deepchecks.Dataset(df=x_test,label=y_test,cat_features=[])
```

```
/usr/local/lib/python3.9/dist-packages/deepchecks/__init__.py:136: DeprecationWarning:

Ability to import base tabular functionality from the `deepchecks` package directly is deprecated, pleas

/usr/local/lib/python3.9/dist-packages/deepchecks/__init__.py:136: DeprecationWarning:
```

```
Ability to import base tabular functionality from the `deepchecks` package directly is deprecated, pleas
```

```
# Create the suite
fsuite = full_suite()
```

```
results = fsuite.run(train_dataset=ds_train,test_dataset=ds_test,model=pipe_lr)
```

```
deepchecks - WARNING - Cannot use model's built-in feature importance on a
WARNING:deepchecks:Cannot use model's built-in feature importance on a Scik
deepchecks - INFO - Calculating permutation feature importance. Expected to
INFO:deepchecks:Calculating permutation feature importance. Expected to fir
```

```
results
```

▾ **Full Suite**

Full Suite

The suite is composed of various checks such as: New Category Train Test, Label Drift, Simple Model Comparison, etc...

Each check may contain conditions (which will result in pass ✓ / fail ✖ / warning ! / error ?!) as well as other outputs such as plots or tables.

Suites, checks and conditions can all be modified. Read more about custom suites.

▸ Didn't Pass

▸ Passed

▸ Other

▸ Didn't Run

▾ Feature/Data Drift

```
from deepchecks.checks import TrainTestFeatureDrift
```

```
/usr/local/lib/python3.9/dist-packages/deepchecks/checks.py:21: DeprecationWarning:

Ability to import tabular checks from the `deepchecks.checks` is deprecated, please import from `deepche
```

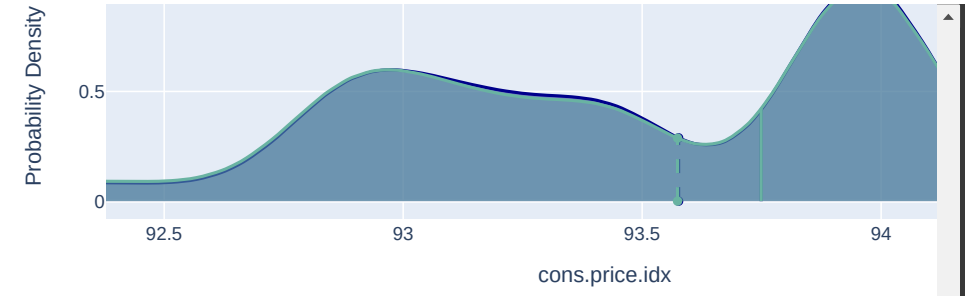
```
check = TrainTestFeatureDrift()
```

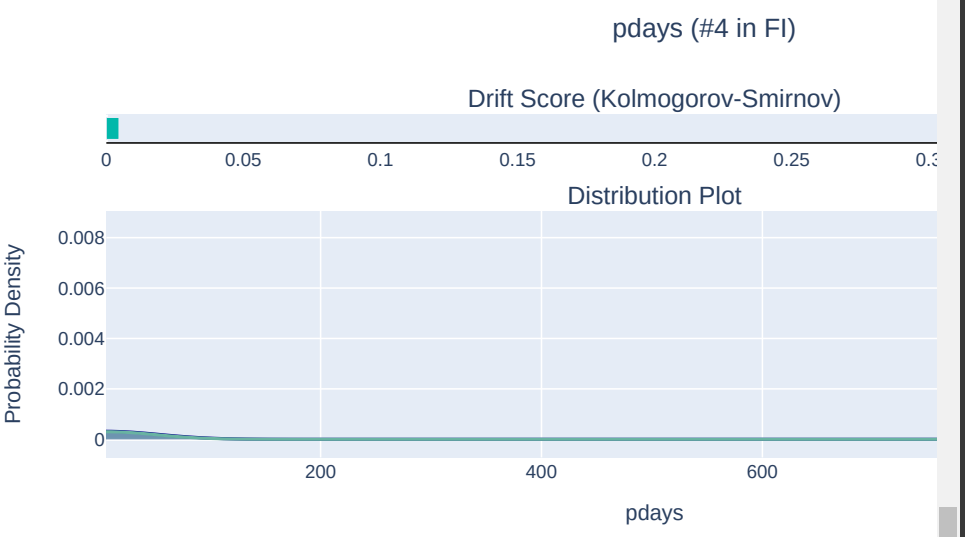
```
<ipython-input-23-58b7fee6c441>:1: DeprecationWarning:

The TrainTestFeatureDrift check is deprecated and will be removed in the 0.14 version. Please use the Fe
```

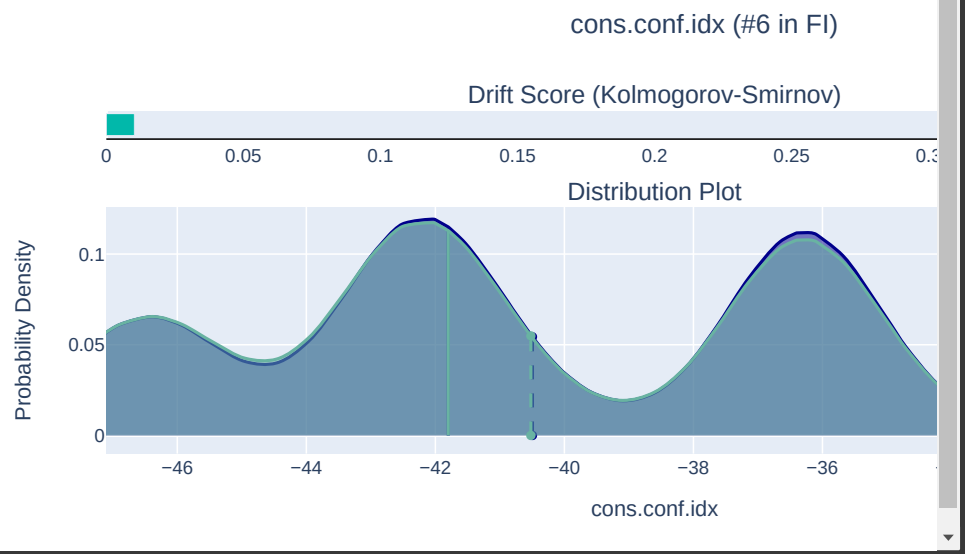
```
result = check.run(train_dataset=ds_train, test_dataset=ds_test, model=pipe_lr)
```

```
deepchecks - WARNING - Cannot use model's built-in feature importance on a Scikit-learn Pipeline, using p
WARNING:deepchecks:Cannot use model's built-in feature importance on a Scikit-learn Pipeline, using perm
deepchecks - INFO - Calculating permutation feature importance. Expected to finish in 32 seconds
INFO:deepchecks:Calculating permutation feature importance. Expected to finish in 32 seconds
```

```
result
```


Probability Density
0.5
0
92.5
93
93.5
94
cons.price.idx


pdays (#4 in FI)
Drift Score (Kolmogorov-Smirnov)
0
0.05
0.1
0.15
0.2
0.25
Distribution Plot
Probability Density
0.008
0.006
0.004
0.002
0
200
400
600
pdays


cons.conf.idx (#6 in FI)
Drift Score (Kolmogorov-Smirnov)
0
0.05
0.1
0.15
0.2
0.25
Distribution Plot
Probability Density
0.1
0.05
0
−46
−44
−42
−40
−38
−36
cons.conf.idx

```
### Label Drift
```

```
from deepchecks.checks import TrainTestLabelDrift
lcheck = TrainTestLabelDrift()
lresult = lcheck.run(train_dataset=ds_train, test_dataset=ds_test)
```

```
<ipython-input-26-a0c3317c2313>:3: DeprecationWarning:

The TrainTestLabelDrift check is deprecated and will be removed in the 0.14 version.Please use the LabelI
```

```
lresult
```

Label Drift

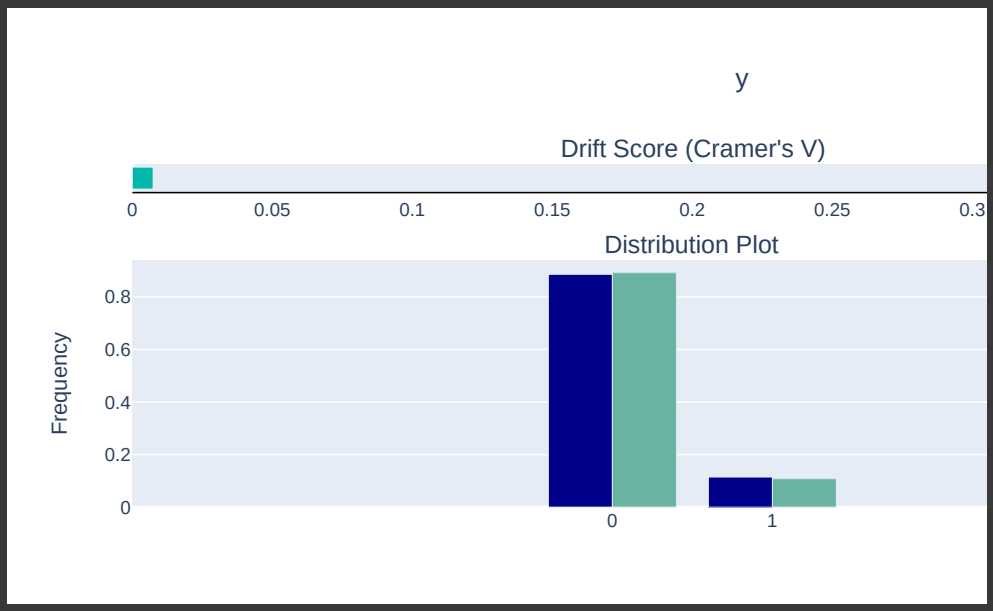
The TrainTestLabelDrift check is deprecated and will be removed in the 0.14 version. Read More...

Additional Outputs

The Drift score is a measure for the difference between two distributions, in this check - the test and train distributions.

The check shows the drift score and distributions for the label.

For discrete distribution plots, showing the top 10 categories with largest difference between train and test.



Dataset Integrity Checks using Deepchecks

- pip install deepchecks

Components

- checks
- suites
- Dataset

```
import pandas as pd
import deepchecks
```

```
# Load Dataset
df = pd.read_csv("data/bank-additional-full_encoded.csv")
```

```
dir(deepchecks)
```

```
['BaseCheck',
 'BaseSuite',
 'CheckFailure',
 'CheckResult',
 'Condition',
 'ConditionCategory',
 'ConditionResult',
 'Context',
 'Dataset',
 'ModelComparisonCheck',
 'ModelComparisonSuite',
 'ModelOnlyBaseCheck',
 'ModelOnlyCheck',
 'SingleDatasetBaseCheck',
 'SingleDatasetCheck',
 'Suite',
 'SuiteResult',
 'TrainTestBaseCheck',
 'TrainTestCheck',
 '_SubstituteModule',
 '__all__',
 '__builtins__',
 '__cached__',
 '__doc__',
 '__file__',
 '__loader__',
 '__name__',
 '__original_module__',
 '__package__',
 '__path__',
 '__spec__',
 '__version__',
 '__warningregistry__',
 '_init_module_attrs',
 'analytics',
 'checks',
 'core',
 'get_verbosity',
 'is_notebook',
 'matplotlib',
 'os',
 'pio',
 'pio_backends',
 'ppscore',
 'set_verbosity',
 'suites',
 'sys',
 'tabular',
 'types',
 'utils',
 'validate_latest_version',
 'version',
 'warnings']
```

```
!pip install single_dataset_integrity
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
ERROR: Could not find a version that satisfies the requirement single_dataset_integrity (from versions: n
ERROR: No matching distribution found for single_dataset_integrity
```

```
from deepchecks.suites import single_dataset_integrity
```