# Python Notebook Viewer

## Mastering Gradient Boosting with CatBoost

In this tutorial we will use dataset Amazon Employee Access Challenge from Kaggle competition for our experiments. Here is the link to the challenge, that we will be exploring.

### Libraries installation

In [1]:
```
#!pip install --user --upgrade catboost
# !pip install --user --upgrade ipywidgets
# !pip install shap
# #!pip install sklearn
# !jupyter nbextension enable --py widgetsnbextension
```

Out [1]:
```
Collecting ipywidgets
  Downloading https://files.pythonhosted.org/packages/56/a0
/dbcf5881bb2f51e8db678211907f16ea0a182b232c591a6d6f276985ca95
/ipywidgets-7.5.1-py2.py3-none-any.whl (121kB)
K    100% |████████████████████████████████| 122kB 4.7MB/s
ent already up-to-date: nbformat>=4.2.0 in /usr/local
/lib/python2.7/dist-packages (from ipywidgets)
Requirement already up-to-date: ipython<6.0.0,>=4.0.0;
python_version < "3.3" in /usr/local/lib/python2.7/dist-
packages (from ipywidgets)
Collecting widgetsnbextension~=3.5.0 (from ipywidgets)
  Downloading https://files.pythonhosted.org/packages/6c/7b
/7ac231c20d2d33c445eaacf8a433f4e22c60677eb9776c7c5262d7ddee2d
/widgetsnbextension-3.5.1-py2.py3-none-any.whl (2.2MB)
K    100% |████████████████████████████████| 2.2MB 415kB/s
 ipywidgets)
  Downloading https://files.pythonhosted.org/packages/59/9e
/e16335ee2d645ee48f082e4207d4fad9bdce09cc1537e76320c341d8d75c
/ipykernel-4.10.1-py2-none-any.whl (109kB)
K    100% |████████████████████████████████| 112kB 9.6MB/s
ent already up-to-date: traitlets>=4.3.1 in /usr/local
/lib/python2.7/dist-packages (from ipywidgets)
Requirement already up-to-date: jupyter-core in /usr/local
/lib/python2.7/dist-packages (from
nbformat>=4.2.0->ipywidgets)
Requirement already up-to-date: ipython-genutils in /usr/local
/lib/python2.7/dist-packages (from
```

```
nbformat>=4.2.0->ipywidgets)
Collecting jsonschema!=2.5.0,>=2.4 (from
nbformat>=4.2.0->ipywidgets)
  Downloading https://files.pythonhosted.org/packages/54/48
/f5f11003ceddcd4ad292d4d9b5677588e9169eef41f88e38b2888e7ec6c4
/jsonschema-3.0.2-py2.py3-none-any.whl (54kB)
K    100% |████████████████████████████████| 61kB 10.4MB/s
ent already up-to-date: pickleshare in /usr/local
/lib/python2.7/dist-packages (from ipython<6.0.0,>=4.0.0;
python_version < "3.3"->ipywidgets)
Requirement already up-to-date: simplegeneric>0.8 in
/usr/local/lib/python2.7/dist-packages (from
ipython<6.0.0,>=4.0.0; python_version < "3.3"->ipywidgets)
Requirement already up-to-date: backports.shutil-get-terminal-
size; python_version == "2.7" in /usr/local/lib/python2.7
/dist-packages (from ipython<6.0.0,>=4.0.0; python_version <
"3.3"->ipywidgets)
Requirement already up-to-date: decorator in /usr/local
/lib/python2.7/dist-packages (from ipython<6.0.0,>=4.0.0;
python_version < "3.3"->ipywidgets)
Requirement already up-to-date: pygments in /usr/local
/lib/python2.7/dist-packages (from ipython<6.0.0,>=4.0.0;
python_version < "3.3"->ipywidgets)
Requirement already up-to-date: pexpect; sys_platform !=
"win32" in /usr/local/lib/python2.7/dist-packages (from
ipython<6.0.0,>=4.0.0; python_version < "3.3"->ipywidgets)
Requirement already up-to-date: pathlib2; python_version ==
"2.7" or python_version == "3.3" in /usr/local/lib/python2.7
/dist-packages (from ipython<6.0.0,>=4.0.0; python_version <
"3.3"->ipywidgets)
Collecting setuptools>=18.5 (from ipython<6.0.0,>=4.0.0;
python_version < "3.3"->ipywidgets)
  Using cached https://files.pythonhosted.org/packages/b2/86
/095d2f7829badc207c893dd4ac767e871f6cd547145df797ea26baea4e2e
/setuptools-41.2.0-py2.py3-none-any.whl
Collecting prompt-toolkit<2.0.0,>=1.0.4 (from
ipython<6.0.0,>=4.0.0; python_version < "3.3"->ipywidgets)
  Downloading https://files.pythonhosted.org/packages/4e/4e
/e5545e8bebf89790455b31112a0f886580c7b3624a05abec416efa5c3a08
/prompt_toolkit-1.0.16-py2-none-any.whl (244kB)
K    100% |████████████████████████████████| 245kB 5.2MB/s
ent already up-to-date: notebook>=4.4.1 in /usr/local
/lib/python2.7/dist-packages (from
widgetsnbextension~=3.5.0->ipywidgets)
Requirement already up-to-date: tornado>=4.0 in /usr/local
/lib/python2.7/dist-packages (from
ipykernel>=4.5.1->ipywidgets)
Collecting jupyter-client (from ipykernel>=4.5.1->ipywidgets)
  Downloading https://files.pythonhosted.org/packages/af/4c
/bf613864ae0644e2ac7d4a40bd209c40c8c71e3dc88d5f1d0aa92a68e716
/jupyter_client-5.3.1-py2.py3-none-any.whl (91kB)
K    100% |████████████████████████████████| 92kB 10.2MB/s
ent already up-to-date: enum34; python_version == "2.7" in
/usr/lib/python2.7/dist-packages (from
traitlets>=4.3.1->ipywidgets)
Requirement already up-to-date: six in /usr/local
/lib/python2.7/dist-packages (from
traitlets>=4.3.1->ipywidgets)
```

```
Requirement already up-to-date: functools32; python_version <
"3" in /usr/local/lib/python2.7/dist-packages (from
jsonschema!=2.5.0,>=2.4->nbformat>=4.2.0->ipywidgets)
Collecting pyrsistent>=0.14.0 (from
jsonschema!=2.5.0,>=2.4->nbformat>=4.2.0->ipywidgets)
  Downloading https://files.pythonhosted.org/packages/b9/66
/b2638d96a2d128b168d0dba60fdc77b7800a9b4a5340cefcc5fc4eae6295
/pyrsistent-0.15.4.tar.gz (107kB)
K    100% |████████████████████████████████| 112kB 9.5MB/s
ent already up-to-date: attrs>=17.4.0 in /usr/local
/lib/python2.7/dist-packages (from
jsonschema!=2.5.0,>=2.4->nbformat>=4.2.0->ipywidgets)
Requirement already up-to-date: ptyprocess>=0.5 in /usr/local
/lib/python2.7/dist-packages (from pexpect; sys_platform !=
"win32"->ipython<6.0.0,>=4.0.0; python_version <
"3.3"->ipywidgets)
Requirement already up-to-date: scandir; python_version <
"3.5" in /usr/local/lib/python2.7/dist-packages (from
pathlib2; python_version == "2.7" or python_version ==
"3.3"->ipython<6.0.0,>=4.0.0; python_version <
"3.3"->ipywidgets)
Requirement already up-to-date: wcwidth in /usr/local
/lib/python2.7/dist-packages (from prompt-
toolkit<2.0.0,>=1.0.4->ipython<6.0.0,>=4.0.0; python_version <
"3.3"->ipywidgets)
Collecting nbconvert (from
notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets)
  Downloading https://files.pythonhosted.org/packages/f9/df
/4505c0a7fea624cac461d0f41051f33456ae656753f65cee8c2f43121cb2
/nbconvert-5.6.0-py2.py3-none-any.whl (453kB)
K    100% |████████████████████████████████| 460kB 2.7MB/s
ent already up-to-date: Send2Trash in /usr/local/lib/python2.7
/dist-packages (from
notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets)
Requirement already up-to-date: prometheus-client in
/usr/local/lib/python2.7/dist-packages (from
notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets)
Requirement already up-to-date: terminado>=0.8.1 in /usr/local
/lib/python2.7/dist-packages (from
notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets)
Requirement already up-to-date: ipaddress; python_version ==
"2.7" in /usr/local/lib/python2.7/dist-packages (from
notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets)
Collecting pyzmq>=17 (from
notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets)
  Downloading https://files.pythonhosted.org/packages/27/d7
/d8aa6dde15f077f20efdf657367e26cbd2fcb96d182d0e690be9b3f2b76c
/pyzmq-18.1.0-cp27-cp27mu-manylinux1_x86_64.whl (1.0MB)
K    100% |████████████████████████████████| 1.0MB 860kB/s
ent already up-to-date: jinja2 in /usr/local/lib/python2.7
/dist-packages (from
notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets)
Requirement already up-to-date: singledispatch in /usr/local
/lib/python2.7/dist-packages (from
tornado>=4.0->ipykernel>=4.5.1->ipywidgets)
Collecting futures (from
tornado>=4.0->ipykernel>=4.5.1->ipywidgets)
  Downloading https://files.pythonhosted.org/packages/d8/a6
```

```
/f46ae3f1da0cd4361c344888f59ec2f5785e69c872e175a748ef6071cdb5
/futures-3.3.0-py2-none-any.whl
Requirement already up-to-date: backports-abc>=0.4 in
/usr/local/lib/python2.7/dist-packages (from
tornado>=4.0->ipykernel>=4.5.1->ipywidgets)
Requirement already up-to-date: python-dateutil>=2.1 in
/usr/local/lib/python2.7/dist-packages (from jupyter-
client->ipykernel>=4.5.1->ipywidgets)
Requirement already up-to-date: pandocfilters>=1.4.1 in
/usr/local/lib/python2.7/dist-packages (from
nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidg
ets)
Requirement already up-to-date: mistune<2,>=0.8.1 in
/usr/local/lib/python2.7/dist-packages (from
nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidg
ets)
Requirement already up-to-date: defusedxml in /usr/local
/lib/python2.7/dist-packages (from
nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidg
ets)
Requirement already up-to-date: testpath in /usr/local
/lib/python2.7/dist-packages (from
nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidg
ets)
Requirement already up-to-date: entrypoints>=0.2.2 in
/usr/local/lib/python2.7/dist-packages (from
nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidg
ets)
Requirement already up-to-date: bleach in /usr/local
/lib/python2.7/dist-packages (from
nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidg
ets)
Requirement already up-to-date: MarkupSafe>=0.23 in /usr/local
/lib/python2.7/dist-packages (from
jinja2->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets
)
Collecting configparser>=3.5; python_version == "2.7" (from
entrypoints>=0.2.2->nbconvert->notebook>=4.4.1->widgetsnbexten
sion~=3.5.0->ipywidgets)
  Downloading https://files.pythonhosted.org/packages/ab/1a
/ec151e5e703ac80041eaccef923611bbcec2b667c20383655a06962732e9
/configparser-3.8.1-py2.py3-none-any.whl
Requirement already up-to-date: webencodings in /usr/local
/lib/python2.7/dist-packages (from
bleach->nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0-
>ipywidgets)
Building wheels for collected packages: pyrsistent
  Running setup.py bdist_wheel for pyrsistent ... ?25ldone
e/jupyter/.cache/pip/wheels/bb/46
/00/6d471ef0b813e3621f0abe6cb723c20d529d39a061de3f7c51
Successfully built pyrsistent
Installing collected packages: widgetsnbextension, pyzmq,
jupyter-client, ipykernel, ipywidgets, pyrsistent, setuptools,
jsonschema, prompt-toolkit, nbconvert, futures, configparser
Successfully installed configparser-3.8.1 futures-3.3.0
ipykernel-4.10.1 ipywidgets-7.5.1 jsonschema-3.0.2 jupyter-
client-5.3.1 nbconvert-5.6.0 prompt-toolkit-1.0.16 pyrsistent-
0.15.4 pyzmq-18.1.0 setuptools-41.2.0 widgetsnbextension-3.5.1
```

```
Collecting shap
  Downloading https://files.pythonhosted.org/packages/3b/98
/16829410426bdd08b836c30e164c56646d6a102afb9eadd81a6bd3a8bb65
/shap-0.30.0.tar.gz (240kB)
K    100% |████████████████████████████████| 245kB 4.2MB/s
ent already satisfied: numpy in /home/jupyter/.local
/lib/python2.7/site-packages (from shap)
Requirement already satisfied: scipy in /home/jupyter/.local
/lib/python2.7/site-packages (from shap)
Requirement already satisfied: scikit-learn in /home/jupyter
/.local/lib/python2.7/site-packages (from shap)
Requirement already satisfied: matplotlib in /usr/local
/lib/python2.7/dist-packages (from shap)
Requirement already satisfied: pandas in /usr/local
/lib/python2.7/dist-packages (from shap)
Requirement already satisfied: tqdm>4.25.0 in /usr/local
/lib/python2.7/dist-packages (from shap)
Requirement already satisfied: ipython in /usr/local
/lib/python2.7/dist-packages (from shap)
Requirement already satisfied: scikit-image in /usr/local
/lib/python2.7/dist-packages (from shap)
Requirement already satisfied: cycler>=0.10 in /usr/local
/lib/python2.7/dist-packages (from matplotlib->shap)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local
/lib/python2.7/dist-packages (from matplotlib->shap)
Requirement already satisfied: python-dateutil>=2.1 in
/usr/local/lib/python2.7/dist-packages (from matplotlib->shap)
Requirement already satisfied: six>=1.10 in /usr/local
/lib/python2.7/dist-packages (from matplotlib->shap)
Requirement already satisfied: pytz in /home/jupyter/.local
/lib/python2.7/site-packages (from matplotlib->shap)
Requirement already satisfied: subprocess32 in /usr/local
/lib/python2.7/dist-packages (from matplotlib->shap)
Requirement already satisfied: backports.functools-lru-cache
in /usr/local/lib/python2.7/dist-packages (from
matplotlib->shap)
Requirement already satisfied:
pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /usr/local
/lib/python2.7/dist-packages (from matplotlib->shap)
Requirement already satisfied: pickleshare in /usr/local
/lib/python2.7/dist-packages (from ipython->shap)
Requirement already satisfied: simplegeneric>0.8 in /usr/local
/lib/python2.7/dist-packages (from ipython->shap)
Requirement already satisfied: traitlets>=4.2 in /usr/local
/lib/python2.7/dist-packages (from ipython->shap)
Requirement already satisfied: backports.shutil-get-terminal-
size; python_version == "2.7" in /usr/local/lib/python2.7
/dist-packages (from ipython->shap)
Requirement already satisfied: decorator in /usr/local
/lib/python2.7/dist-packages (from ipython->shap)
Requirement already satisfied: pygments in /usr/local
/lib/python2.7/dist-packages (from ipython->shap)
Requirement already satisfied: pexpect; sys_platform !=
"win32" in /usr/local/lib/python2.7/dist-packages (from
ipython->shap)
Requirement already satisfied: pathlib2; python_version ==
"2.7" or python_version == "3.3" in /usr/local/lib/python2.7
/dist-packages (from ipython->shap)
```

```
Requirement already satisfied: setuptools>=18.5 in
/home/jupyter/.local/lib/python2.7/site-packages (from
ipython->shap)
Requirement already satisfied: prompt-toolkit<2.0.0,>=1.0.4 in
/home/jupyter/.local/lib/python2.7/site-packages (from
ipython->shap)
Requirement already satisfied: PyWavelets>=0.4.0 in /usr/local
/lib/python2.7/dist-packages (from scikit-image->shap)
Requirement already satisfied: networkx>=1.8 in /usr/local
/lib/python2.7/dist-packages (from scikit-image->shap)
Requirement already satisfied: cloudpickle>=0.2.1 in
/usr/local/lib/python2.7/dist-packages (from scikit-
image->shap)
Collecting pillow>=4.3.0 (from scikit-image->shap)
  Downloading https://files.pythonhosted.org/packages/cc/a4
/79b5f36d1e1a2b426073bd62217d1530fcd939950c2936651e6b39127a9b
/Pillow-6.1.0-cp27-cp27mu-manylinux1_x86_64.whl (2.1MB)
K    100% |████████████████████████████████| 2.1MB 597kB/s
ent already satisfied: enum34; python_version == "2.7" in
/usr/lib/python2.7/dist-packages (from
traitlets>=4.2->ipython->shap)
Requirement already satisfied: ipython-genutils in /usr/local
/lib/python2.7/dist-packages (from
traitlets>=4.2->ipython->shap)
Requirement already satisfied: ptyprocess>=0.5 in /usr/local
/lib/python2.7/dist-packages (from pexpect; sys_platform !=
"win32"->ipython->shap)
Requirement already satisfied: scandir; python_version < "3.5"
in /usr/local/lib/python2.7/dist-packages (from pathlib2;
python_version == "2.7" or python_version ==
"3.3"->ipython->shap)
Requirement already satisfied: wcwidth in /usr/local
/lib/python2.7/dist-packages (from prompt-
toolkit<2.0.0,>=1.0.4->ipython->shap)
Building wheels for collected packages: shap
  Running setup.py bdist_wheel for shap ... ?25ldone
e/jupyter/.cache/pip/wheels/67/7a
/5b/34feab81170fb8bf642a7536b5127e54e00bce373564435808
Successfully built shap
Installing collected packages: shap, pillow
Successfully installed pillow-6.1.0 shap-0.30.0
Enabling notebook extension jupyter-js-widgets/extension...
      - Validating: OK
```

In [2]:

```
import os
import pandas as pd
import numpy as np
np.set_printoptions(precision=4)

import catboost
print(catboost.__version__)
```

Out [2]:

```
0.16.5
```

# Reading the data

In [3]:
```
from catboost.datasets import amazon

# If you have "URLError: SSL: CERTIFICATE_VERIFY_FAILED
# import ssl
# ssl._create_default_https_context = ssl._create_unver

# If you have any other error:
# Download datasets from http://bit.ly/2ZUXTSv and unco
# train_df = pd.read_csv('train.csv', sep=',', header='

(train_df, test_df) = amazon()
```

In [4]:
```
train_df.head()
```

Out [4]:

| | ACTION | RESOURCE | MGR_ID | ROLE_ROLLUP_1 | ROLE_ROLI |
|---|---|---|---|---|---|
| 0 | 1 | 39353 | 85475 | 117961 | 118300 |
| 1 | 1 | 17183 | 1540 | 117961 | 118343 |
| 2 | 1 | 36724 | 14457 | 118219 | 118220 |
| 3 | 1 | 36135 | 5396 | 117961 | 118343 |
| 4 | 1 | 42680 | 5905 | 117929 | 117930 |

# Exploring the data

Label values extraction

In [5]:
```
y = train_df.ACTION
X = train_df.drop('ACTION', axis=1)
```

Categorical features declaration

In [6]:
```
cat_features = list(range(0, X.shape[1]))
print(cat_features)
```

Out [6]:
```
[0, 1, 2, 3, 4, 5, 6, 7, 8]
```

Looking on label balance in dataset

In [7]:
```
print('Labels: {}'.format(set(y)))
print('Zero count = {}, One count = {}'.format(len(y) -
```

```
Labels: {0, 1}
Zero count = 1897, One count = 30872
```

# Training the first model

In [8]:
```
from catboost import CatBoostClassifier
model = CatBoostClassifier(iterations=100)
model.fit(X, y, cat_features=cat_features, verbose=10)
```

Out [8]:
```
Learning rate set to 0.349945
0:      learn: 0.4668712        total: 69ms     remaining:
6.83s
10:     learn: 0.1791505        total: 262ms    remaining:
2.12s
20:     learn: 0.1684338        total: 474ms    remaining:
1.78s
30:     learn: 0.1650651        total: 685ms    remaining:
1.52s
40:     learn: 0.1640099        total: 893ms    remaining:
1.28s
50:     learn: 0.1619138        total: 1.1s     remaining:
1.06s
60:     learn: 0.1606667        total: 1.31s    remaining:
840ms
70:     learn: 0.1595925        total: 1.53s    remaining:
624ms
80:     learn: 0.1578885        total: 1.74s    remaining:
408ms
90:     learn: 0.1566430        total: 1.95s    remaining:
193ms
99:     learn: 0.1556847        total: 2.14s    remaining: 0us
```

Out [8]:
```
<catboost.core.CatBoostClassifier at 0x7f1ff5e50438>
```

In []:
```
model.predict_proba(X)
```

In [9]:
```
model.classes_
```

Out [9]:
```
[]
```

# Working with dataset

There are several ways of passing dataset to training - using X,y (the initial matrix) or using Pool class. Pool class is the class for storing the dataset. In the next few blocks we'll explore the ways to create a Pool object.

You can use Pool class if the dataset has more than just X and y (for example, it has sample weights or groups) or if the dataset is large and it takes long time to read it into python.

In [10]:
```
from catboost import Pool
pool = Pool(data=X, label=y, cat_features=cat_features)
```

# Split your data into train and validation

In [11]:
```
from sklearn.model_selection import train_test_split

data = train_test_split(X, y, test_size=0.2, random_sta
X_train, X_validation, y_train, y_validation = data

train_pool = Pool(
    data=X_train,
    label=y_train,
    cat_features=cat_features
)

validation_pool = Pool(
    data=X_validation,
    label=y_validation,
    cat_features=cat_features
)
```

# Selecting the objective function

Possible options for binary classification:

`Logloss` for binary target.

`CrossEntropy` for probabilities in target.

Out [14]:
```
model = CatBoostClassifier(
    iterations=5,
    learning_rate=0.1,
    loss_function='Logloss'
)
model.fit(train_pool, eval_set=validation_pool, verbose

print('Model is fitted: {}'.format(model.is_fitted()))
print('Model params:\n{}'.format(model.get_params()))
```

```
Model is fitted: True
Model params:
{'learning_rate': 0.1, 'iterations': 5, 'loss_function':
'Logloss'}
```

In [15]:

```
model.classes_
```

Out [15]:

```
[]
```

# Stdout of the training

In [16]:

```
model = CatBoostClassifier(
    iterations=15,
#     verbose=5,
)
model.fit(train_pool, eval_set=validation_pool);
```

Out [16]:

```
Learning rate set to 0.5
0:      learn: 0.3971379        test: 0.3960691 best:
0.3960691 (0)   total: 21.9ms   remaining: 307ms
1:      learn: 0.2948071        test: 0.2924021 best:
0.2924021 (1)   total: 44.2ms   remaining: 287ms
2:      learn: 0.2485015        test: 0.2455237 best:
0.2455237 (2)   total: 64.1ms   remaining: 256ms
3:      learn: 0.2234262        test: 0.2192359 best:
0.2192359 (3)   total: 86.4ms   remaining: 238ms
4:      learn: 0.2003506        test: 0.1938956 best:
0.1938956 (4)   total: 108ms    remaining: 215ms
5:      learn: 0.1916473        test: 0.1831990 best:
0.1831990 (5)   total: 125ms    remaining: 187ms
6:      learn: 0.1842038        test: 0.1759780 best:
0.1759780 (6)   total: 144ms    remaining: 164ms
7:      learn: 0.1808767        test: 0.1722588 best:
0.1722588 (7)   total: 161ms    remaining: 141ms
8:      learn: 0.1783738        test: 0.1678080 best:
0.1678080 (8)   total: 180ms    remaining: 120ms
9:      learn: 0.1769061        test: 0.1658153 best:
0.1658153 (9)   total: 202ms    remaining: 101ms
10:     learn: 0.1761268        test: 0.1653031 best:
0.1653031 (10)  total: 220ms    remaining: 79.9ms
11:     learn: 0.1752620        test: 0.1645631 best:
0.1645631 (11)  total: 237ms    remaining: 59.3ms
12:     learn: 0.1749475        test: 0.1643973 best:
0.1643973 (12)  total: 256ms    remaining: 39.4ms
13:     learn: 0.1746794        test: 0.1643479 best:
0.1643479 (13)  total: 275ms    remaining: 19.7ms
14:     learn: 0.1739538        test: 0.1632657 best:
0.1632657 (14)  total: 293ms    remaining: 0us

bestTest = 0.1632656889
```

```
bestIteration = 14
```

# Metrics calculation and graph plotting

In [20]:

```
model = CatBoostClassifier(
    iterations=50,
    learning_rate=0.5,
    loss_function='Logloss',
    custom_loss=['AUC', 'Accuracy']
)

model.fit(
    train_pool,
    eval_set=validation_pool,
    verbose=False,
    plot=True
);
```

Out [20]:

Out [20]:

```
MetricVisualizer(layout=Layout(align_self='stretch',
height='500px'))
```

In [21]:

```
model.classes_
```

Out [21]:

```
[]
```

# Model comparison

In []:

```
model1 = CatBoostClassifier(
    learning_rate=0.7,
    iterations=100,
    train_dir='learing_rate_0.7'
)

model2 = CatBoostClassifier(
    learning_rate=0.01,
    iterations=100,
    train_dir='learing_rate_0.01'
)

model1.fit(train_pool, eval_set=validation_pool, verbos
model2.fit(train_pool, eval_set=validation_pool, verbos
```

In []:
```
from catboost import MetricVisualizer
MetricVisualizer(['learing_rate_0.7', 'learing_rate_0.0
```

## Best iteration

In []:
```
model = CatBoostClassifier(
    iterations=100,
#     use_best_model=False
)
model.fit(
    train_pool,
    eval_set=validation_pool,
    verbose=False,
    plot=True
);
```

In []:
```
print('Tree count: ' + str(model.tree_count_))
```

## Cross-validation

In []:
```
from catboost import cv

params = {
    'loss_function': 'Logloss',
    'iterations': 80,
    'custom_loss': 'AUC',
    'learning_rate': 0.5,
}

cv_data = cv(
    params = params,
    pool = train_pool,
    fold_count=5,
    shuffle=True,
    partition_random_seed=0,
    plot=True,
    verbose=False
)
```

In []:
```
cv_data.head(10)
```

In []:
```
best_value = np.min(cv_data['test-Logloss-mean'])
best_iter = np.argmin(cv_data['test-Logloss-mean'])

print('Best validation Logloss score, not stratified: {
    best_value,
    cv_data['test-Logloss-std'][best_iter],
    best_iter)
)
```

In []:
```
from catboost import cv

params = {
    'loss_function': 'Logloss',
    'iterations': 80,
    'custom_loss': 'AUC',
    'learning_rate': 0.5,
}

cv_data = cv(
    params = params,
    pool = train_pool,
    fold_count=5,
    shuffle=True,
    partition_random_seed=0,
    plot=True,
    stratified=False,
    verbose=False
)
```

In []:
```
best_value = cv_data['test-Logloss-mean'].min()
best_iter = cv_data['test-Logloss-mean'].values.argmin(

print('Best validation Logloss score, stratified: {:.4f
    best_value,
    cv_data['test-Logloss-std'][best_iter],
    best_iter)
)
```

## Sklearn Grid Search

In []:
```
from sklearn.model_selection import GridSearchCV

param_grid = {
```

```
    "learning_rate": [0.001, 0.01, 0.5],
}

clf = CatBoostClassifier(
    iterations=20,
    cat_features=cat_features,
    verbose=20
)
grid_search = GridSearchCV(clf, param_grid=param_grid,
results = grid_search.fit(X_train, y_train)
results.best_estimator_.get_params()
```

# Overfitting Detector

In []:

```
model_with_early_stop = CatBoostClassifier(
    iterations=200,
    learning_rate=0.5,
    early_stopping_rounds=20
)

model_with_early_stop.fit(
    train_pool,
    eval_set=validation_pool,
    verbose=False,
    plot=True
);
```

In []:

```
print(model_with_early_stop.tree_count_)
```

## Overfitting Detector with eval metric

In []:

```
model_with_early_stop = CatBoostClassifier(
    eval_metric='AUC',
    iterations=200,
    learning_rate=0.5,
    early_stopping_rounds=20
)
model_with_early_stop.fit(
    train_pool,
    eval_set=validation_pool,
    verbose=False,
    plot=True
);
```

In []:
```
print(model_with_early_stop.tree_count_)
```

## Model predictions

In []:
```
model = CatBoostClassifier(iterations=200, learning_rat
model.fit(train_pool, verbose=50);
```

In []:
```
print(model.predict(X_validation))
```

In []:
```
print(model.predict_proba(X_validation))
```

In []:
```
raw_pred = model.predict(
    X_validation,
    prediction_type='RawFormulaVal'
)

print(raw_pred)
```

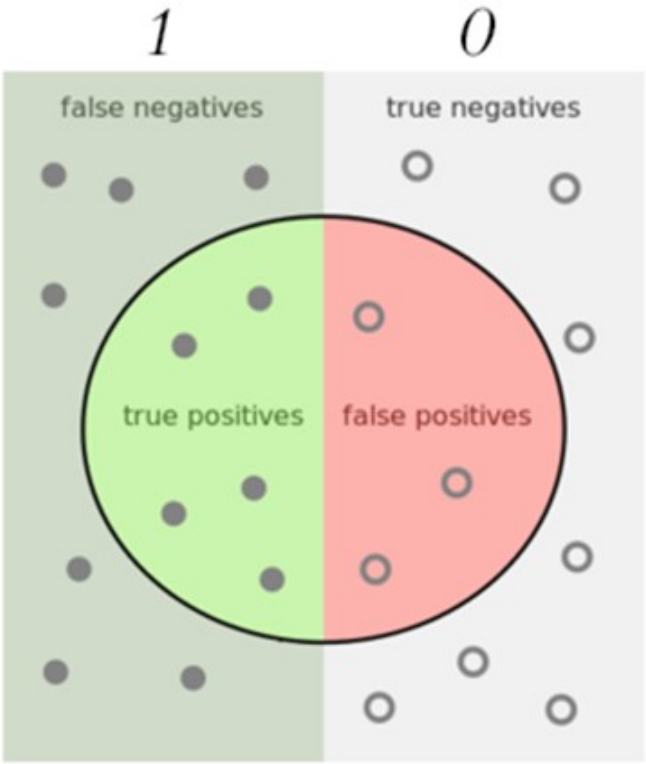
In []:
```
from numpy import exp

sigmoid = lambda x: 1 / (1 + exp(-x))

probabilities = sigmoid(raw_pred)

print(probabilities)
```

## Select decision boundary





In []:

```
import matplotlib.pyplot as plt
from catboost.utils import get_roc_curve
from catboost.utils import get_fpr_curve
from catboost.utils import get_fnr_curve

curve = get_roc_curve(model, validation_pool)
(fpr, tpr, thresholds) = curve

(thresholds, fpr) = get_fpr_curve(curve=curve)
(thresholds, fnr) = get_fnr_curve(curve=curve)
```

In []:

```
plt.figure(figsize=(16, 8))
style = {'alpha':0.5, 'lw':2}

plt.plot(thresholds, fpr, color='blue', label='FPR', **
plt.plot(thresholds, fnr, color='green', label='FNR', *

plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xticks(fontsize=16)
plt.yticks(fontsize=16)
plt.grid(True)
plt.xlabel('Threshold', fontsize=16)
plt.ylabel('Error Rate', fontsize=16)
plt.title('FPR-FNR curves', fontsize=20)
```

```
plt.legend(loc="lower left", fontsize=16);
```

In []:
```
from catboost.utils import select_threshold

print(select_threshold(model, validation_pool, FNR=0.01
print(select_threshold(model, validation_pool, FPR=0.01
```

# Metric evaluation on a new dataset

In []:
```
metrics = model.eval_metrics(
    data=validation_pool,
    metrics=['Logloss','AUC'],
    ntree_start=0,
    ntree_end=0,
    eval_period=1,
    plot=True
)
```

In []:
```
print('AUC values:\n{}'.format(np.array(metrics['AUC'])
```

# Feature importances

## Prediction values change

Default feature importances for binary classification is PredictionValueChange - how much on average does the model change when the feature value changes. These feature importances are non negative. They are normalized and sum to 1, so you can look on these values like percentage of importance.

In []:
```
np.array(model.get_feature_importance(prettified=True))
```

## Loss function change

The non default feature importance approximates how much the optimized loss function will change if the value of the feature changes. This importances might be negative if the feature has bad influence on the loss function. The importances are not normalized, the absolute value of the importance has the same scale as the optimized loss value. To calculate this importance value you need to pass train_pool as an argument.

In []:
```
np.array(model.get_feature_importance(
```

```
    train_pool,
    'LossFunctionChange',
    prettified=True
))
```

### Shap values

In []:
```
print(model.predict_proba([X.iloc[1,:]]))
print(model.predict_proba([X.iloc[91,:]]))
```

In []:
```
shap_values = model.get_feature_importance(
    validation_pool,
    'ShapValues'
)
expected_value = shap_values[0,-1]
shap_values = shap_values[:,:-1]
print(shap_values.shape)
```

In []:
```
proba = model.predict_proba([X.iloc[1,:]])[0]
raw = model.predict([X.iloc[1,:]], prediction_type='Raw
print('Probabilities', proba)
print('Raw formula value %.4f' % raw)
print('Probability from raw value %.4f' % sigmoid(raw))
```

In []:
```
import shap

shap.initjs()
shap.force_plot(expected_value, shap_values[1,:], X_val
```

In []:
```
proba = model.predict_proba([X.iloc[91,:]])[0]
raw = model.predict([X.iloc[91,:]], prediction_type='Ra
print('Probabilities', proba)
print('Raw formula value %.4f' % raw)
print('Probability from raw value %.4f' % sigmoid(raw))
```

In []:
```
import shap
shap.initjs()
shap.force_plot(expected_value, shap_values[91,:], X_va
```

In []:
```
shap.summary_plot(shap_values, X_validation)
```

## Snapshotting

In []:
```
#!rm 'catboost_info/snapshot.bkp'

model = CatBoostClassifier(
    iterations=100,
    save_snapshot=True,
    snapshot_file='snapshot.bkp',
    snapshot_interval=1
)

model.fit(train_pool, eval_set=validation_pool, verbose
```

## Saving the model

In []:
```
model = CatBoostClassifier(iterations=10)
model.fit(train_pool, eval_set=validation_pool, verbose
model.save_model('catboost_model.bin')
model.save_model('catboost_model.json', format='json')
```

In []:
```
model.load_model('catboost_model.bin')
print(model.get_params())
print(model.learning_rate_)
```

## Hyperparameter tunning

In []:
```
tunned_model = CatBoostClassifier(
    iterations=1000,
    learning_rate=0.03,
    depth=6,
    l2_leaf_reg=3,
    random_strength=1,
    bagging_temperature=1
)

tunned_model.fit(
    X_train, y_train,
    cat_features=cat_features,
    verbose=False,
    eval_set=(X_validation, y_validation),
    plot=True
);
```

# Speeding up the training

In []:

```
fast_model = CatBoostClassifier(
    boosting_type='Plain',
    rsm=0.5,
    one_hot_max_size=50,
    leaf_estimation_iterations=1,
    max_ctr_complexity=1,
    iterations=100,
    learning_rate=0.3,
    bootstrap_type='Bernoulli',
    subsample=0.5
)
fast_model.fit(
    X_train, y_train,
    cat_features=cat_features,
    verbose=False,
    eval_set=(X_validation, y_validation),
    plot=True
);
```

# Reducing model size

In []:

```
small_model = CatBoostClassifier(
    learning_rate=0.03,
    iterations=500,
    model_size_reg=50,
    max_ctr_complexity=1,
    ctr_leaf_count_limit=100
)
small_model.fit(
    X_train, y_train,
    cat_features=cat_features,
    verbose=False,
    eval_set=(X_validation, y_validation),
    plot=True
);
```

---