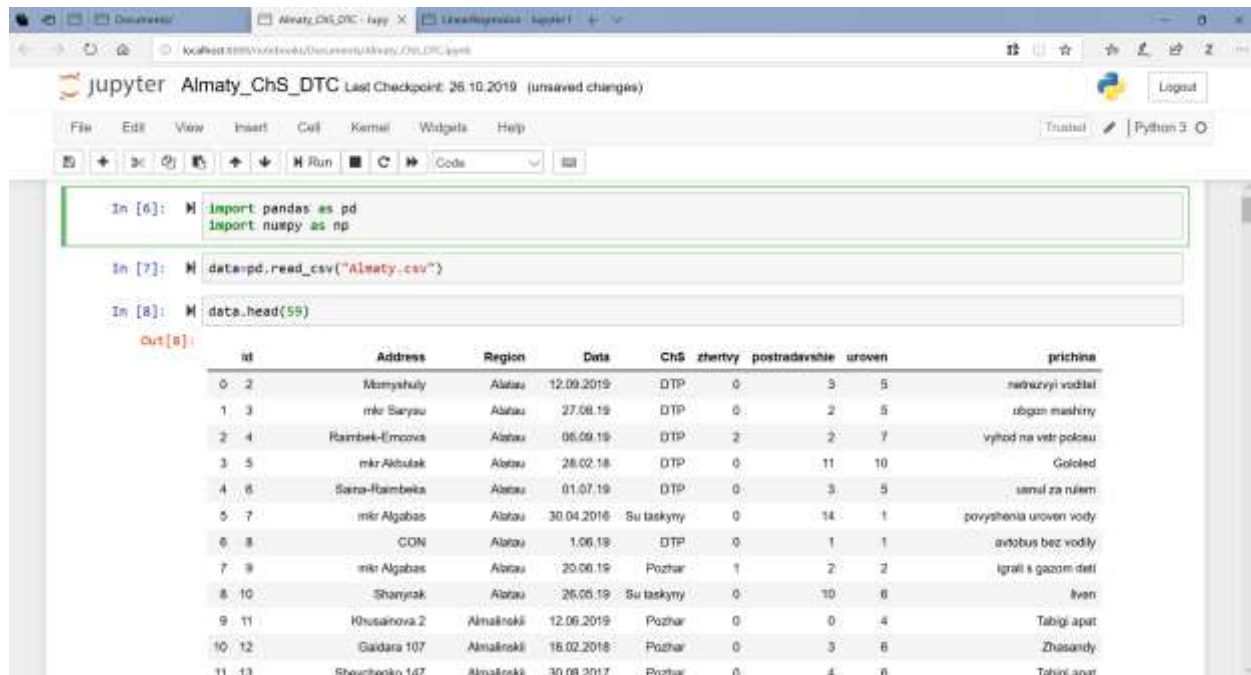


Лабораторное занятие 12. Градиентный бустинг.

12-зертханалық сабақ. Градиентті бустинг.

Python және Anaconda да градиенттік бустинг әдісін жүзеге асыру. Мысал. Сіз өз деректер қорына жасаныз!

На Python и Anaconde реализовать метод градиентного бустинга. Ниже приведен пример. Каждый должен сделать для своего датасета.



The screenshot shows a Jupyter Notebook interface with the following code and output:

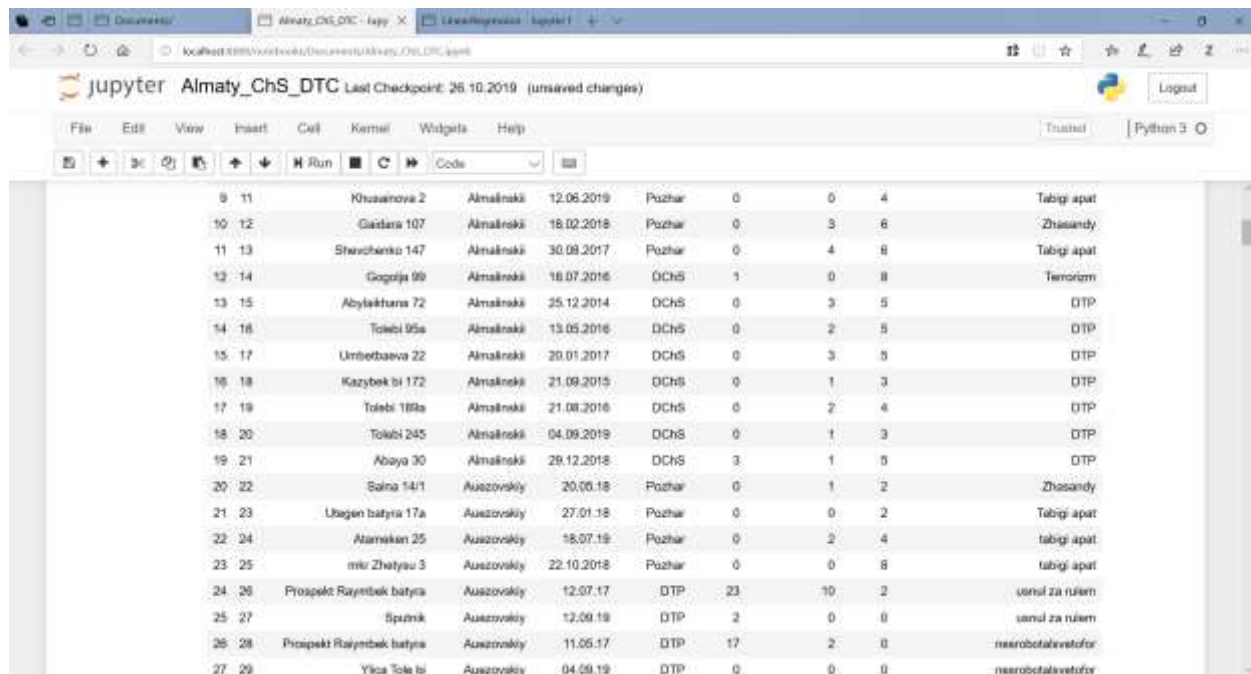
```
In [6]: import pandas as pd
import numpy as np

In [7]: data = pd.read_csv("Almaty.csv")

In [8]: data.head(59)
```

Out[8]:

	id	Address	Region	Date	ChS	zhertyy	postradavshie	uroven	prichina
0	2	Momyshuly	Alatau	12.09.2019	DTP	0	3	5	neduznyi vozhel
1	3	mkr Saryau	Alatau	27.08.19	DTP	0	2	5	zbgan mashiny
2	4	Raimbek-Emova	Alatau	06.09.19	DTP	2	2	7	vyhod na vstr pokou
3	5	mkr Akbulak	Alatau	28.02.18	DTP	0	11	10	Golosed
4	6	Saina-Raimbek	Alatau	01.07.19	DTP	0	3	5	samul za rukem
5	7	mkr Algabaz	Alatau	30.04.2016	Su taskyny	0	14	1	porvsheniya uroven vody
6	8	CON	Alatau	1.06.18	DTP	0	1	1	avtobus bez vodily
7	9	mkr Algabaz	Alatau	20.06.19	Pozhar	1	2	2	igrati s gazom deli
8	10	Shanyrak	Alatau	26.06.19	Su taskyny	0	10	6	liven
9	11	Khusainova 2	Almatinskii	12.06.2019	Pozhar	0	0	4	Tabigi apat
10	12	Gaidara 107	Almatinskii	18.02.2018	Pozhar	0	3	6	Zhasandy
11	13	Shovchenko 147	Almatinskii	30.08.2017	Pozhar	0	4	6	Tabigi apat



The screenshot shows a Jupyter Notebook interface with the following code and output:

```
In [9]: data.tail(20)
```

Out[9]:

9	11	Khusainova 2	Almatinskii	12.06.2019	Pozhar	0	0	4	Tabigi apat
10	12	Gaidara 107	Almatinskii	18.02.2018	Pozhar	0	3	6	Zhasandy
11	13	Shovchenko 147	Almatinskii	30.08.2017	Pozhar	0	4	6	Tabigi apat
12	14	Gogolja 99	Almatinskii	18.07.2016	OChS	1	0	8	Temorizm
13	15	Abylekhanova 72	Almatinskii	25.12.2014	OChS	0	3	5	DTP
14	16	Tokeli 95a	Almatinskii	13.05.2016	OChS	0	2	5	DTP
15	17	Umbetbaeva 22	Almatinskii	20.01.2017	OChS	0	3	5	DTP
16	18	Kazybek bi 172	Almatinskii	21.09.2015	OChS	0	1	3	DTP
17	19	Tokeli 188a	Almatinskii	21.08.2016	OChS	0	2	4	DTP
18	20	Tokeli 245	Almatinskii	04.09.2019	OChS	0	1	3	DTP
19	21	Abaya 30	Almatinskii	29.12.2018	OChS	3	1	0	DTP
20	22	Saina 14/1	Aussovskiy	20.06.18	Pozhar	0	1	2	Zhasandy
21	23	Ubaev batyr 17a	Aussovskiy	27.01.18	Pozhar	0	0	2	Tabigi apat
22	24	Atameken 25	Aussovskiy	18.07.19	Pozhar	0	2	4	tabigi apat
23	25	mkr Zhatyau 3	Aussovskiy	22.10.2018	Pozhar	0	0	8	tabigi apat
24	26	Prospekt Raimbek batyr	Aussovskiy	12.07.17	DTP	23	10	2	usnul za rukem
25	27	Bpdnrik	Aussovskiy	12.09.19	DTP	2	0	0	samul za rukem
26	28	Prospekt Raimbek batyr	Aussovskiy	11.05.17	DTP	17	2	0	neisobotalevetofo
27	29	Ylka Tokel	Aussovskiy	04.09.19	DTP	0	0	0	neisobotalevetofo

Almaty_ChS_DTC - jupyter X LiveKernelSession - jupyter1

localhost:8888/vscode?token=Almaty_ChS_DTC.ipynb

jupyter Almaty_ChS_DTC Last Checkpoint: 26.10.2019 (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

```

57 68 ul. Chernyshevskogo 35 Turksibskii raion 21 iul' 2019 Pozhar 0 0 2 zaryskanie provodov
58 69 ul. Melnikova Turksibskii raion 13 iun' 2019 Pozhar 0 0 8 ne izvestno

```

In [9]: `data.info()`

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 113 entries, 0 to 112
Data columns (total 9 columns):
id                113 non-null int64
Address           113 non-null object
Region            113 non-null object
Data              113 non-null object
ChS               113 non-null object
zhertvy           113 non-null int64
postradavshie     113 non-null int64
uroven            113 non-null int64
prichina          113 non-null object
dtypes: int64(4), object(5)
memory usage: 8.1+ KB

```

In [10]: `data.shape`

```
Out[10]: (113, 9)
```

In [11]: `data['Region'].value_counts()`

```
Out[11]: Nauryzbayski    30
```

Almaty_ChS_DTC - jupyter X LiveKernelSession - jupyter1

localhost:8888/vscode?token=Almaty_ChS_DTC.ipynb

jupyter Almaty_ChS_DTC Last Checkpoint: 26.10.2019 (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

```
Out[10]: (113, 9)
```

In [11]: `data['Region'].value_counts()`

```
Out[11]: Nauryzbayski    30
Turksibskii raion    19
Auezovskiy           17
Zhetisu              15
Bostandykyski        12
AleaIinskii          11
Alatau               9
Name: Region, dtype: int64

```

In [12]: `data.describe()`

```
Out[12]:
```

	id	zhertvy	postradavshie	uroven
count	113.000000	113.000000	113.000000	113.000000
mean	65.681416	2.274336	4.920354	3.893805
std	37.856484	9.900609	14.170607	4.162680
min	2.000000	0.000000	0.000000	0.000000
25%	30.000000	0.000000	0.000000	1.000000
50%	67.000000	0.000000	1.000000	3.000000
75%	97.000000	1.000000	3.000000	5.000000

```
Almaty_ChS_DTC - jupyter X LiveRegistrator - jupyter X
localhost:8888/vnotebooks/Documents/Almaty_ChS_DTC.ipynb
jupyter Almaty_ChS_DTC Last Checkpoint: 26.10.2019 (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3
In [12]: data.describe()
Out[12]:
```

	id	zheroty	postredavshie	uroven
count	113.000000	113.000000	113.000000	113.000000
mean	65.681416	2.274336	4.920254	3.803805
std	37.858464	9.960859	14.170607	4.162680
min	2.000000	0.000000	0.000000	0.000000
25%	30.000000	0.000000	0.000000	1.000000
50%	67.000000	0.000000	1.000000	3.000000
75%	87.000000	1.000000	3.000000	5.000000
max	130.000000	100.000000	90.000000	37.000000

```
In [13]: data["ChS"].value_counts()
Out[13]: Pozhar: 44
DTP: 40
Su taskyny: 13
DChS: 9
zemletresenie: 5
ChS: 1
podtoplenie: 1
Name: ChS, dtype: int64
```

```
Almaty_ChS_DTC - jupyter X LiveRegistrator - jupyter X
localhost:8888/vnotebooks/Documents/Almaty_ChS_DTC.ipynb
jupyter Almaty_ChS_DTC Last Checkpoint: 26.10.2019 (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3
In [14]: data["Region"].value_counts()
Out[14]: Nauryzbayski: 38
Turksibskii raion: 19
Auezovskiy: 17
Zhetisu: 15
Bostandyyski: 12
Almalinskii: 11
Alatau: 9
Name: Region, dtype: int64
In [15]: data["ChS"] = data["ChS"].map({'DTP':1, 'Pozhar':2, 'Su taskyny':3, 'DChS':4, 'ChS':5, 'podtoplenie':6, 'zemletresenie':7})
In [16]: data["Region"] = data["Region"].map({'Alatau':1, 'Almalinskii':2, 'Auezovskiy':3, 'Zhetisu':4, 'Turksibskii raion':5, 'Nauryzbayski':6, 'Nauryzbayski':7, 'Bostandyyski':8})
In [17]: data.head(113)
Out[17]:
```

	id	Address	Region	Date	ChS	zheroty	postredavshie	uroven	prichina
0	2	Mornyshuly	1	12.09.2019	1	0	3	5	netreznyi voditel
1	3	mkr Saryau	1	27.08.19	1	0	2	5	obgon mashiny
2	4	Raimbek-Emcova	1	06.09.19	1	2	2	7	vyhod na vstr pokosu
3	5	mkr Akbulak	1	28.02.18	1	0	11	10	Golosed

```
Almaty_ChS_DTC - jupyter X LiveKernel: jupyter1
localhost:8888/vnotebooks/Documents/Almaty_ChS_DTC.ipynb

jupyter Almaty_ChS_DTC Last Checkpoint: 26.10.2019 (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

In [17]: data.head(113)

Out[17]:
```

	id	Address	Region	Data	ChS	zherty	postradavshie	uroven	prichina
0	2	Momyshuly	1	12.09.2019	1	0	3	5	netrezyl voditel
1	3	mkr Sarysu	1	27.08.19	1	0	2	5	obgon mashiny
2	4	Raimbek-Encova	1	06.09.19	1	2	2	7	vyhod na vstr polosy
3	5	mkr Akbulak	1	28.02.18	1	0	11	10	Goloked
4	6	Saina-Raimbeka	1	01.07.19	1	0	3	5	vsnul za rukom
...	...	...	...	...	...	...	...	...	...
108	125	Rozybakiwa 235	8	12.04.2016	2	0	0	0	Gaz tutumy
109	126	Tauqul	8	14.01.2015	1	1	2	2	Algodik kui
110	128	Gagarina 95	8	06.04.2018	1	1	3	2	Algodik kui
111	129	Orbita 1	8	03.11.2014	2	0	1	2	Kaupstizdik saktamau
112	130	Rozybakiwa 271	8	17.01.2019	1	0	3	3	Algodik kui

113 rows x 9 columns

```
In [18]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 113 entries, 0 to 112
```

```
Almaty_ChS_DTC - jupyter X LiveKernel: jupyter1
localhost:8888/vnotebooks/Documents/Almaty_ChS_DTC.ipynb

jupyter Almaty_ChS_DTC Last Checkpoint: 26.10.2019 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

In [17]: data.head(113)

Out[17]:
```

	id	Address	Region	Data	ChS	zherty	postradavshie	uroven	prichina
0	2	Momyshuly	1	12.09.2019	1	0	3	5	netrezyl voditel
1	3	mkr Sarysu	1	27.08.19	1	0	2	5	obgon mashiny
2	4	Raimbek-Encova	1	06.09.19	1	2	2	7	vyhod na vstr polosy
3	5	mkr Akbulak	1	28.02.18	1	0	11	10	Goloked
4	6	Saina-Raimbeka	1	01.07.19	1	0	3	5	vsnul za rukom
...	...	...	...	...	...	...	...	...	...
108	125	Rozybakiwa 235	8	12.04.2016	2	0	0	0	Gaz tutumy
109	126	Tauqul	8	14.01.2015	1	1	2	2	Algodik kui
110	128	Gagarina 95	8	06.04.2018	1	1	3	2	Algodik kui
111	129	Orbita 1	8	03.11.2014	2	0	1	2	Kaupstizdik saktamau
112	130	Rozybakiwa 271	8	17.01.2019	1	0	3	3	Algodik kui

113 rows x 9 columns

```
In [18]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 113 entries, 0 to 112
```

```
Almaty_ChS_DTC - jupyter X LiveKernel: jupyter1
localhost:8888/vnotebooks/Documents/Almaty_ChS_DTC.ipynb

jupyter Almaty_ChS_DTC Last Checkpoint: 26.10.2019 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

In [18]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 113 entries, 0 to 112
Data columns (total 9 columns):
id          113 non-null int64
Address     113 non-null object
Region      113 non-null int64
Data        113 non-null object
ChS         113 non-null int64
zherty      113 non-null int64
postradavshie 113 non-null int64
uroven      113 non-null int64
prichina    113 non-null object
dtypes: int64(6), object(3)
memory usage: 8.1+ KB

In [19]: data.describe()

Out[19]:
```

	id	Region	ChS	zherty	postradavshie	uroven
count	113.000000	113.000000	113.000000	113.000000	113.000000	113.000000
mean	63.661416	4.890310	2.20354	2.274336	4.920354	3.893609
std	37.859464	2.215443	1.44642	9.900659	14.170007	4.162680
min	2.000000	1.000000	1.000000	0.000000	0.000000	0.000000

```
Almaty_ChS_DTC - jupyter X LiveKernel: jupyter1
localhost:8888/vscode/Document/Almaty_ChS_DTC.ipynb

jupyter Almaty_ChS_DTC Last Checkpoint: 26.10.2019 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

75% 97.000000 7.000000 3.000000 1.000000 3.000000 5.000000
max 130.000000 8.000000 7.000000 100.000000 90.000000 37.000000

In [20]: M y=data['Ch5'].astype('int')
#nepe0e0e0e run int

In [21]: M sum(data['Ch5']==1)
Out[21]: 48

In [22]: M X=data.drop(['Ch5','Address','Data','prichina'], axis=1)

In [23]: M data['Region'].value_counts()
Out[23]:
7    30
5    19
3    17
4    15
8    12
2    11
1     9
Name: Region, dtype: int64

In [24]: M X.shape, y.shape
Out[24]: ((113, 5), (113,))
```

```
Almaty_ChS_DTC - jupyter X LiveKernel: jupyter1
localhost:8888/vscode/Document/Almaty_ChS_DTC.ipynb

jupyter Almaty_ChS_DTC Last Checkpoint: 26.10.2019 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Out[24]: ((113, 5), (113,))

In [25]: M data['Region'].astype('int')
Out[25]:
0     1
1     1
2     1
3     1
4     1
..
108    8
109    8
110    8
111    8
112    8
Name: Region, Length: 113, dtype: int32

In [26]: M from sklearn.model_selection import train_test_split

In [27]: M X_train, X_valid, y_train, y_valid = train_test_split(X, y,
test_size=0.22, random_state=1)

In [28]: M X_train.shape, X_valid.shape
Out[28]: ((88, 5), (25, 5))
```

```
Almaty_ChS_DTC - jupyter X LiveKernel: jupyter1
localhost:8888/vscode/Document/Almaty_ChS_DTC.ipynb

jupyter Almaty_ChS_DTC Last Checkpoint: 26.10.2019 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

In [28]: M X_train.shape, X_valid.shape
Out[28]: ((88, 5), (25, 5))

In [29]: M ! pip install xgboost
Requirement already satisfied: xgboost in c:\programdata\anaconda3\lib\site-packages (0.90)
Requirement already satisfied: scipy in c:\programdata\anaconda3\lib\site-packages (from xgboost) (1.3.1)
Requirement already satisfied: numpy in c:\programdata\anaconda3\lib\site-packages (from xgboost) (1.16.5)

In [30]: M #!pip install sklearn.model_selection
from sklearn.model_selection import KFold, train_test_split
from xgboost import XGBClassifier
from xgboost import XGBRegressor

In [31]: M xgb= XGBClassifier(max_depth=5,
random_state = 2,
learning_rate=0.03333)

In [32]: M xgb.fit(X_train, y_train)
Out[32]: XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
colsample_bynode=1, colsample_bytree=1, gamma=0,
learning_rate=0.03333, max_delta_step=0, max_depth=5,
min_child_weight=1, missing=None, n_estimators=100, n_jobs=1,
nthread=None, objective='multi:softprob', random_state=2,
reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
silent=None, subsample=1, verbosity=1)
```



```
Almaty_ChS_DTC - Ipy X | LinearRegression - Ipy X | +
localhost:8888/vscode/Document/Almaty_ChS_DTC.ipynb
jupyter Almaty_ChS_DTC Last Checkpoint: 26.10.2019 (autosaved)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3
+ + + + + Run + Code
learning_rate=0.03333)

In [32]: Xgb.fit(X_train, y_train)

Out[32]: XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
colsample_bynode=1, colsample_bytree=1, gamma=0,
learning_rate=0.03333, max_delta_step=0, max_depth=5,
min_child_weight=1, missing=None, n_estimators=100, n_jobs=1,
nthread=None, objective='multi:softprob', random_state=2,
reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
silent=None, subsample=1, verbosity=1)

In [35]: from sklearn import metrics
metrics.accuracy_score(xgb.predict(X_valid), y_valid)

Out[35]: 0.68

In [36]: xgbr=XGBRegressor()

In [37]: xgbr.fit(X_train, y_train)

C:\ProgramData\Anaconda3\lib\site-packages\xgboost\core.py:587: FutureWarning: Series.base is deprecated and will be removed in a future version
if getattr(data, 'base', None) is not None and \

[21:23:50] WARNING: C:/Jenkins/workspace/xgboost-win64_release_0.90/src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.
```

```
Almaty_ChS_DTC - Ipy X | LinearRegression - Ipy X | +
localhost:8888/vscode/Document/Almaty_ChS_DTC.ipynb
jupyter Almaty_ChS_DTC Last Checkpoint: 26.10.2019 (autosaved)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3
+ + + + + Run + Code

In [37]: xgbr.fit(X_train, y_train)

C:\ProgramData\Anaconda3\lib\site-packages\xgboost\core.py:587: FutureWarning: Series.base is deprecated and will be removed in a future version
if getattr(data, 'base', None) is not None and \

[21:23:50] WARNING: C:/Jenkins/workspace/xgboost-win64_release_0.90/src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.

Out[37]: XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
colsample_bynode=1, colsample_bytree=1, gamma=0,
importance_type='gain', learning_rate=0.1, max_delta_step=0,
max_depth=3, min_child_weight=1, missing=None, n_estimators=100,
n_jobs=1, nthread=None, objective='reg:linear', random_state=0,
reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
silent=None, subsample=1, verbosity=1)

In [42]: y_pred=xgbr.predict(X_valid)

In [43]: mae=metrics.mean_absolute_error(y_valid,y_pred)
r2=metrics.r2_score(y_valid, y_pred)
mse=metrics.mean_squared_error(y_valid, y_pred)

In [45]: print("mae=",mae)
print("mse=",mse)
print("r2=",r2)
```

Almaty_ChS_DTC · Jupyter X

localhost:8888/vscode44/Documents/Almaty_ChS_DTC.ipynb

jupyter Almaty_ChS_DTC Last Checkpoint: 26.10.2019 (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

if getattr(data, 'base', None) is not None and \

[21:23:50] WARNING: C:/Jenkins/workspace/xgboost-win64_release_0.90/src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.

Out[37]: XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1, gamma=0, importance_type='gain', learning_rate=0.1, max_delta_step=0, max_depth=3, min_child_weight=1, missing=None, n_estimators=100, n_jobs=1, nthread=None, objective='reg:linear', random_state=0, reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None, silent=None, subsample=1, verbosity=1)

In [42]: `y_pred=xghr.predict(X_valid)`

In [43]: `mse=metrics.mean_absolute_error(y_valid,y_pred)
r2=metrics.r2_score(y_valid, y_pred)
mse=metrics.mean_squared_error(y_valid, y_pred)`

In [45]: `print("mae=",mae)
print("mse=",mse)
print("r2=",r2)`

mae= 1.0776214265823365
mse= 2.9792717366385837
r2= -0.7508421385330023