Hack_n_change

December 1, 2019

```python
[738]: import warnings
       import numpy as np
       import pandas as pd
       from sklearn.model_selection import train_test_split
       import matplotlib.pyplot as plt
       import warnings
       import numpy as np
       import pandas as pd
       from catboost import CatBoostClassifier
       from catboost import *
       SEED = 17
```

```python
[739]: df_hours = pd.read_excel('abs_hours_df.xlsx', sep=',')
       df_hours = df_hours.drop(['Unnamed: 0'], axis=1)
       df_hours.rename(columns={0: "work_hours"}, inplace=True)
       df_hours.fillna((df_hours['work_hours'].median()), inplace=True)
```

```python
[740]: df = pd.read_excel('hackaton_data_5post.xlsx', sep=',')
       df = pd.concat([df,df_hours], axis=1)
       df.head()
```

```
[740]:    Unnamed: 0                    ADDRESS    CITY          REGION  \
       0           0    .,   ., 25/15
       1           1     .,   ., 1
       2           2    .,   ., 25
       3           3    .,   ., 60/49
       4           4      .,   ., 81

          Postamat_daily  cashbox_daily  Postamat_trend  cashbox_MAX  cashbox_trend  \
       0             NaN            3.3             NaN         13.0            1.0
       1             NaN            NaN             NaN          NaN            NaN
       2             NaN            3.1             NaN         12.0            1.0
       3             3.0            2.2             0.0         12.0            1.0
       4             1.0            2.0             0.0          9.0            1.0

          hh_500  value1  value2  POPULATION  NEAR_Malls  NEAR_Business_centers  \
       0  7696.0   10099    3382      324698           0                      0
       1  6369.0   10068    3236      324698           1                      1
```

```
2   5553.0   10100   3436       324698              1                    0
3   7162.0   10099   3257       324698              0                    0
4   6269.0   10096   3195       324698              1                    0


    NEAR_metro_rjd  NEAR_Stations  macro_salary_avg_yearly  work_hours
0             NaN                                    423321        24.0
1             NaN                                    423321        24.0
2             NaN            NaN                      423321        15.0
3             NaN            NaN                      423321        15.0
4             NaN                                    423321        15.0
```

[741]:
```python
df=df.drop(['Unnamed: 0'],axis=1)
df=df.drop(['ADDRESS'],axis=1)
df.head()
```

[741]:
```
     CITY       REGION  Postamat_daily  cashbox_daily  Postamat_trend  \
0             NaN              3.3             NaN
1             NaN              NaN             NaN
2             NaN              3.1             NaN
3             3.0              2.2             0.0
4             1.0              2.0             0.0


    cashbox_MAX  cashbox_trend  hh_500  value1  value2  POPULATION  NEAR_Malls  \
0          13.0            1.0  7696.0   10099    3382      324698           0
1           NaN            NaN  6369.0   10068    3236      324698           1
2          12.0            1.0  5553.0   10100    3436      324698           1
3          12.0            1.0  7162.0   10099    3257      324698           0
4           9.0            1.0  6269.0   10096    3195      324698           1


    NEAR_Business_centers  NEAR_metro_rjd  NEAR_Stations  \
0                       0             NaN
1                       1             NaN
2                       0             NaN            NaN
3                       0             NaN            NaN
4                       0             NaN


    macro_salary_avg_yearly  work_hours
0                    423321        24.0
1                    423321        24.0
2                    423321        15.0
3                    423321        15.0
4                    423321        15.0
```

[742]:
```python
#
df.isna().mean()
```

[742]:
```
CITY                       0.000000
REGION                     0.000000
Postamat_daily             0.827338
```

```
cashbox_daily              0.633094
Postamat_trend             0.827338
cashbox_MAX                0.633094
cashbox_trend              0.664269
hh_500                     0.002398
value1                     0.000000
value2                     0.000000
POPULATION                 0.000000
NEAR_Malls                 0.000000
NEAR_Business_centers      0.000000
NEAR_metro_rjd             0.976019
NEAR_Stations              0.779376
macro_salary_avg_yearly    0.000000
work_hours                 0.000000
dtype: float64
```

[743]:
```python
df.fillna(value = {'NEAR_metro_rjd':0, 'NEAR_Stations':0}, inplace=True)
df.replace(to_replace=[''], value=int(1), inplace=True)
df.replace(to_replace=[''], value=int(2), inplace=True)
df.head()
```

[743]:
```
    CITY          REGION  Postamat_daily  cashbox_daily  Postamat_trend  \
0    NaN             3.3             NaN
1    NaN             NaN             NaN
2    NaN             3.1             NaN
3    3.0             2.2             0.0
4    1.0             2.0             0.0

   cashbox_MAX  cashbox_trend  hh_500  value1  value2  POPULATION  NEAR_Malls  \
0         13.0            1.0  7696.0   10099    3382      324698           0
1          NaN            NaN  6369.0   10068    3236      324698           1
2         12.0            1.0  5553.0   10100    3436      324698           1
3         12.0            1.0  7162.0   10099    3257      324698           0
4          9.0            1.0  6269.0   10096    3195      324698           1

   NEAR_Business_centers  NEAR_metro_rjd  NEAR_Stations  \
0                      0             0.0              1
1                      1             0.0              1
2                      0             0.0              0
3                      0             0.0              0
4                      0             0.0              1

   macro_salary_avg_yearly  work_hours
0                   423321        24.0
1                   423321        24.0
2                   423321        15.0
3                   423321        15.0
4                   423321        15.0
```

```
[744]: df.describe()
```

```
[744]:        Postamat_daily  cashbox_daily  Postamat_trend  cashbox_MAX  \
       count       72.000000     153.000000       72.000000   153.000000
       mean         2.416667       1.362745        0.347222     7.156863
       std          1.275445       0.899699        0.479428     3.509606
       min          1.000000       0.200000        0.000000     1.000000
       25%          2.000000       0.700000        0.000000     5.000000
       50%          2.000000       1.200000        0.000000     7.000000
       75%          3.000000       1.900000        1.000000     9.000000
       max          7.000000       4.800000        1.000000    19.000000

              cashbox_trend        hh_500        value1       value2      POPULATION  \
       count     140.000000    416.000000    417.000000   417.000000      417.000000
       mean        0.892857   2551.600962  10006.038369  2808.601918   147402.803357
       std         0.310405   1963.959893     57.084102   202.017409   175342.903452
       min         0.000000      0.000000   9911.000000  2231.000000      119.000000
       25%         1.000000   1001.500000   9956.000000  2670.000000    11288.000000
       50%         1.000000   2165.000000   9999.000000  2781.000000    55282.000000
       75%         1.000000   3558.750000  10056.000000  2928.000000   324698.000000
       max         1.000000   9775.000000  10100.000000  3484.000000   482873.000000

              NEAR_Malls  NEAR_Business_centers  NEAR_metro_rjd  NEAR_Stations  \
       count  417.000000             417.000000      417.000000     417.000000
       mean     0.103118               0.035971        0.023981       0.225420
       std      0.304478               0.186442        0.153173       0.429698
       min      0.000000               0.000000        0.000000       0.000000
       25%      0.000000               0.000000        0.000000       0.000000
       50%      0.000000               0.000000        0.000000       0.000000
       75%      0.000000               0.000000        0.000000       0.000000
       max      1.000000               1.000000        1.000000       2.000000

              macro_salary_avg_yearly  work_hours
       count               417.000000  417.000000
       mean             404089.705036   15.672062
       std               17786.448797    2.520811
       min              387679.000000   13.000000
       25%              387679.000000   15.000000
       50%              387679.000000   15.000000
       75%              423321.000000   15.000000
       max              423321.000000   24.000000
```

```
[745]: num_cols = [
           'Postamat_trend',
           'cashbox_MAX',
           'cashbox_trend',
           'hh_500',
           'value1',
```

```
        'value2',
        'POPULATION',
        'NEAR_Malls',
        'NEAR_Business_centers',
        'NEAR_metro_rjd',
        'NEAR_Stations',
        'macro_salary_avg_yearly'
]

cat_cols = [
        'CITY',
        'REGION',
]

target_col = ['Postamat_daily',
              'cashbox_daily']

cols = num_cols + cat_cols + target_col
```

```
[746]: df['NEAR_Stations']=df['NEAR_Stations'].astype('int64')
        df['work_hours']=df['work_hours'].astype('int64')
        df['NEAR_metro_rjd']=df['NEAR_metro_rjd'].astype('int64')
        df['CITY']=df['CITY'].astype('object')
        df['REGION']=df['REGION'].astype('object')
```

```
[747]: df.corr().style.background_gradient(cmap='coolwarm').set_precision(2)
```

```
[747]: <pandas.io.formats.style.Styler at 0x197050e8ef0>
```

```
[837]: import seaborn as sns
        # `pairplot()` may become very slow with the SVG format
        %config InlineBackend.figure_format = 'png'
        sns.pairplot(df.dropna());
```
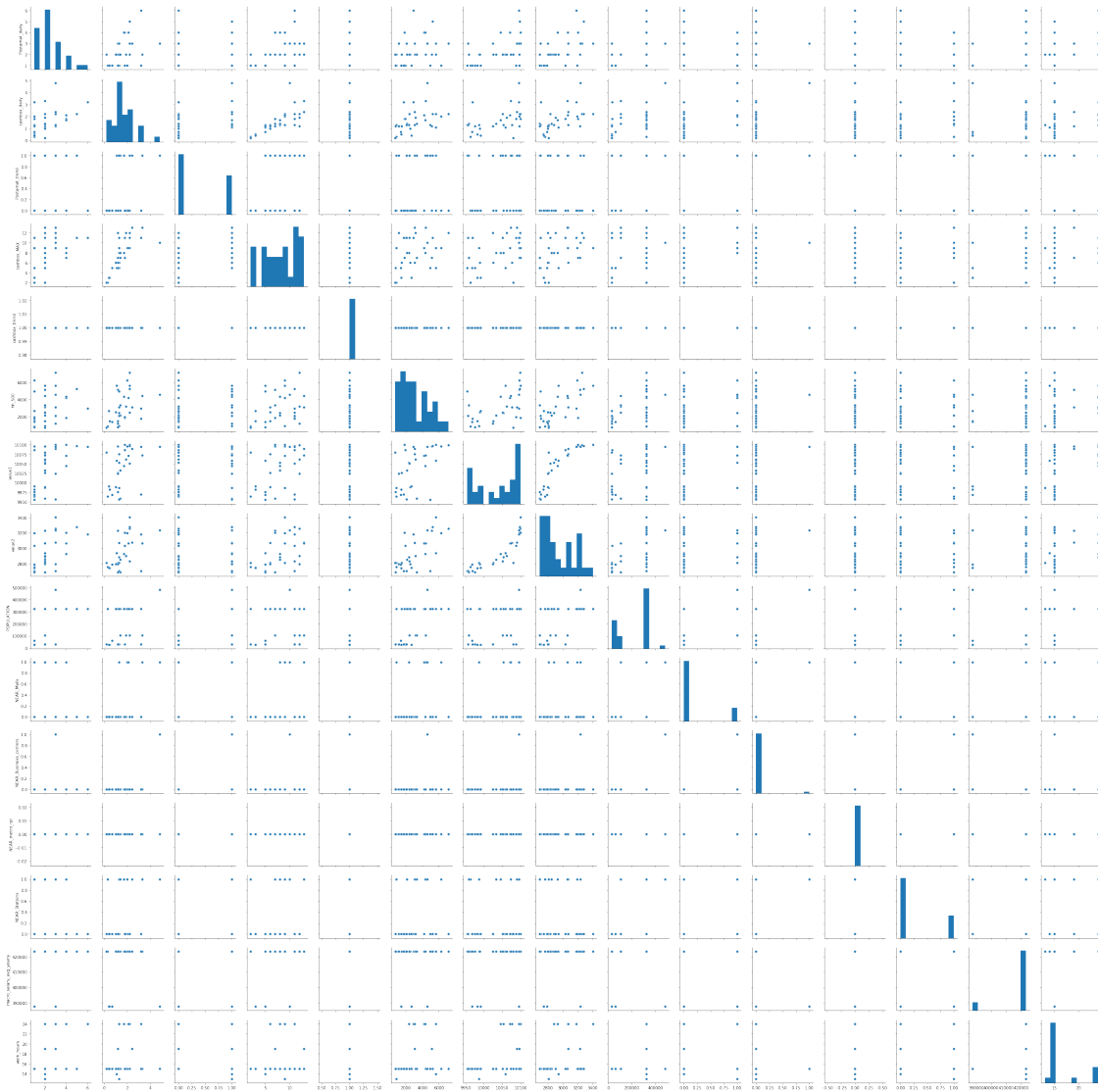
[749]: `df.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 417 entries, 0 to 416
Data columns (total 17 columns):
CITY                 417 non-null object
REGION               417 non-null object
Postamat_daily        72 non-null float64
cashbox_daily        153 non-null float64
Postamat_trend        72 non-null float64
cashbox_MAX          153 non-null float64
cashbox_trend        140 non-null float64
hh_500               416 non-null float64
value1               417 non-null int64
```

```
value2                    417 non-null int64
POPULATION                417 non-null int64
NEAR_Malls                417 non-null int64
NEAR_Business_centers     417 non-null int64
NEAR_metro_rjd            417 non-null int64
NEAR_Stations             417 non-null int64
macro_salary_avg_yearly   417 non-null int64
work_hours                417 non-null int64
dtypes: float64(6), int64(9), object(2)
memory usage: 55.5+ KB
```

[750]:
```python
dff = df.copy()
dff = pd.concat([dff, pd.get_dummies(dff.REGION).rename(columns={" ": "K", " ":␣
  →"T"})], axis=1).copy()
dff = dff.drop(columns=['REGION'], axis=1)
dff = dff.drop(columns=['CITY'], axis=1)
dff['values'] = dff['value1']/dff['value2']

dff = dff.drop(columns=['value2'], axis=1)
dff = dff.drop(columns=['value1'], axis=1)
dff.head()
```

[750]:
```
   Postamat_daily  cashbox_daily  Postamat_trend  cashbox_MAX  cashbox_trend  \
0             NaN            3.3             NaN         13.0            1.0
1             NaN            NaN             NaN          NaN            NaN
2             NaN            3.1             NaN         12.0            1.0
3             3.0            2.2             0.0         12.0            1.0
4             1.0            2.0             0.0          9.0            1.0

    hh_500  POPULATION  NEAR_Malls  NEAR_Business_centers  NEAR_metro_rjd  \
0   7696.0      324698           0                      0               0
1   6369.0      324698           1                      1               0
2   5553.0      324698           1                      0               0
3   7162.0      324698           0                      0               0
4   6269.0      324698           1                      0               0

   NEAR_Stations  macro_salary_avg_yearly  work_hours  K  T     values
0              1                   423321          24  1  0  2.986103
1              1                   423321          24  1  0  3.111248
2              0                   423321          15  1  0  2.939464
3              0                   423321          15  1  0  3.100706
4              1                   423321          15  1  0  3.159937
```

0.1 Postamat_daily

```
[850]: df_post_purpose = dff[dff.isna().Postamat_daily == True].copy()
       df_post_train = dff[dff.isna().Postamat_daily == False].copy()
       print('post purpose:', df_post_purpose.shape[0])
       print('post train:', df_post_train.shape[0])
```

```
post purpose: 345
post train: 72
```

0.2 Postamat_daily

```
[851]: df_post_train.head()
```

```
[851]:     Postamat_daily  cashbox_daily  Postamat_trend  cashbox_MAX  cashbox_trend  \
       3              3.0            2.2             0.0         12.0            1.0
       4              1.0            2.0             0.0          9.0            1.0
       6              3.0            2.2             0.0          9.0            1.0
       8              2.0            1.2             0.0          7.0            1.0
       10             5.0            2.2             1.0         11.0            1.0

            hh_500  POPULATION  NEAR_Malls  NEAR_Business_centers  NEAR_metro_rjd  \
       3    7162.0      324698           0                      0               0
       4    6269.0      324698           1                      0               0
       6    5632.0      324698           0                      0               0
       8    5156.0      324698           0                      0               0
       10   5232.0      324698           0                      0               0

            NEAR_Stations  macro_salary_avg_yearly  work_hours   K  T     values
       3                0                   423321          15   1  0   3.100706
       4                1                   423321          15   1  0   3.159937
       6                0                   423321          24   1  0   2.967969
       8                0                   423321          19   1  0   3.127633
       10               0                   423321          15   1  0   3.080232
```

```
[852]: df_post_train = df_post_train.
        ↪drop(columns=['cashbox_daily','cashbox_MAX','cashbox_trend',␣
        ↪'macro_salary_avg_yearly'])
```

```
[853]: df_post_train.head()
```

```
[853]:     Postamat_daily  Postamat_trend  hh_500  POPULATION  NEAR_Malls  \
       3              3.0             0.0  7162.0      324698           0
       4              1.0             0.0  6269.0      324698           1
       6              3.0             0.0  5632.0      324698           0
       8              2.0             0.0  5156.0      324698           0
       10             5.0             1.0  5232.0      324698           0

            NEAR_Business_centers  NEAR_metro_rjd  NEAR_Stations  work_hours  K  T  \
```

```
3                     0              0              0        15  1  0
4                     0              0              1        15  1  0
6                     0              0              0        24  1  0
8                     0              0              0        19  1  0
10                    0              0              0        15  1  0

      values
3   3.100706
4   3.159937
6   2.967969
8   3.127633
10  3.080232
```

[854]:
```python
df_post_purpose=df_post_purpose.
 ↪drop(columns=['Postamat_daily','cashbox_daily','cashbox_MAX','cashbox_trend',
 ↪'macro_salary_avg_yearly'])
df_post_purpose['hh_500'].fillna((df_post_purpose['hh_500'].mean()),
 ↪inplace=True)
```

[855]:
```python
df_post_purpose.head()
```

[855]:
```
   Postamat_trend  hh_500  POPULATION  NEAR_Malls  NEAR_Business_centers  \
0             NaN  7696.0      324698           0                      0
1             NaN  6369.0      324698           1                      1
2             NaN  5553.0      324698           1                      0
5             NaN  7703.0      324698           0                      0
7             NaN  6487.0      324698           0                      0

   NEAR_metro_rjd  NEAR_Stations  work_hours   K  T    values
0               0              1          24   1  0  2.986103
1               0              1          24   1  0  3.111248
2               0              0          15   1  0  2.939464
5               0              1          19   1  0  3.110700
7               0              1          24   1  0  3.160301
```

[856]:
```python
print(df_post_train.shape)
print(df_post_purpose.shape)
```

```
(72, 12)
(345, 11)
```

0.2.1 X_POST

[857]:
```python
X_post = df_post_train.drop(['Postamat_daily'] , axis=1)
y_post = df_post_train['Postamat_daily']
print(X_post.shape, y_post.shape)
```

```
(72, 11) (72,)
```

0.3 CATBOOST

```
[858]: X_post.head()
```

```
[858]:     Postamat_trend  hh_500  POPULATION  NEAR_Malls  NEAR_Business_centers  \
       3             0.0  7162.0      324698           0                      0
       4             0.0  6269.0      324698           1                      0
       6             0.0  5632.0      324698           0                      0
       8             0.0  5156.0      324698           0                      0
       10            1.0  5232.0      324698           0                      0

           NEAR_metro_rjd  NEAR_Stations  work_hours   K  T    values
       3                0              0          15   1  0  3.100706
       4                0              1          15   1  0  3.159937
       6                0              0          24   1  0  2.967969
       8                0              0          19   1  0  3.127633
       10               0              0          15   1  0  3.080232
```

```
[859]: X_post.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 72 entries, 3 to 336
Data columns (total 11 columns):
Postamat_trend         72 non-null float64
hh_500                 72 non-null float64
POPULATION             72 non-null int64
NEAR_Malls             72 non-null int64
NEAR_Business_centers  72 non-null int64
NEAR_metro_rjd         72 non-null int64
NEAR_Stations          72 non-null int64
work_hours             72 non-null int64
K                      72 non-null uint8
T                      72 non-null uint8
values                 72 non-null float64
dtypes: float64(3), int64(6), uint8(2)
memory usage: 5.8 KB
```

```
[860]: X_post['Postamat_trend'] = X_post['Postamat_trend'].astype('int64').
       ↪astype('str')
       X_post['NEAR_Malls'] = X_post['NEAR_Malls'].astype('str')
       X_post['NEAR_Business_centers'] = X_post['NEAR_Business_centers'].astype('str')
       X_post['NEAR_metro_rjd'] = X_post['NEAR_metro_rjd'].astype('str')
       X_post['NEAR_Stations'] = X_post['NEAR_Stations'].astype('str')
       X_post['K'] = X_post['K'].astype('str')
       X_post['T'] = X_post['T'].astype('str')

       X_post['hh_500'] = X_post['hh_500'].astype('int64')
```

```
[861]: X_post.head()
```

```
[861]:     Postamat_trend  hh_500  POPULATION NEAR_Malls NEAR_Business_centers  \
       3               0    7162       324698          0                      0
       4               0    6269       324698          1                      0
       6               0    5632       324698          0                      0
       8               0    5156       324698          0                      0
       10              1    5232       324698          0                      0

           NEAR_metro_rjd NEAR_Stations  work_hours  K  T    values
       3                0             0          15  1  0  3.100706
       4                0             1          15  1  0  3.159937
       6                0             0          24  1  0  2.967969
       8                0             0          19  1  0  3.127633
       10               0             0          15  1  0  3.080232
```

```python
[862]: X_post_train_part, X_post_valid, y_post_train_part, y_post_valid =␣
       ↪train_test_split(X_post, y_post,

                                                                test_size=0.2,

                                                                    ␣

       ↪random_state=SEED)
       print(X_post_train_part.shape[0], X_post_valid.shape[0], y_post_train_part.
       ↪shape[0], y_valid.shape[0])
```

```
57 15 57 15
```

```python
[863]: categ_feat_idx = ['Postamat_trend', 'NEAR_Malls', 'NEAR_Business_centers',␣
       ↪'NEAR_metro_rjd', 'NEAR_Stations', 'K', 'T']
```

```python
[864]: # RMSE, MAE, Quantile, LogLinQuantile, Poisson, MAPE, Lq
       params = {'loss_function':'MAE',
                 'eval_metric':'MAE',
                 'cat_features': categ_feat_idx,
                 'verbose': 200,
                 'random_seed': SEED,
                 'early_stopping_rounds': 100,
                 'boosting_type': 'Ordered', #Ordered
                 'bootstrap_type':'Bayesian',
                 'iterations': 1000,
                 'task_type':'GPU'
               }
       ctb_par_post = CatBoostRegressor(**params)
       ctb_par_post.fit(X_post_train_part, y_post_train_part,
                     eval_set=(X_post_valid, y_post_valid),
                     use_best_model=True,
                     plot=True
                     );
```

```
<IPython.core.display.HTML object>
```

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

```
0:      learn: 2.4249755       test: 2.3196833 best: 2.3196833 (0)      total:
22.7ms   remaining: 22.7s
200:    learn: 0.7727980       test: 0.8638288 best: 0.8559601 (175)   total:
4.59s    remaining: 18.2s
bestTest = 0.855960083
bestIteration = 175
Shrink model to first 176 iterations.
```

[865]:
```python
ctb_par_post.get_feature_importance(prettified=True)
feature_importance_df = pd.DataFrame(ctb_par_post.
 →get_feature_importance(prettified=True), columns=['Feature Id',␣
 →'Importances'])
feature_importance_df
```

[865]:

	Feature Id	Importances
0	hh_500	51.463399
1	POPULATION	23.920782
2	values	8.399194
3	NEAR_Business_centers	5.343277
4	work_hours	4.346571
5	NEAR_metro_rjd	3.446845
6	NEAR_Malls	1.005310
7	Postamat_trend	0.850929
8	T	0.799959
9	NEAR_Stations	0.383039
10	K	0.040695

[866]:
```python
plt.scatter(ctb_par_post.predict(X_post_valid), y_post_valid)
```

[866]: <matplotlib.collections.PathCollection at 0x19711191dd8>
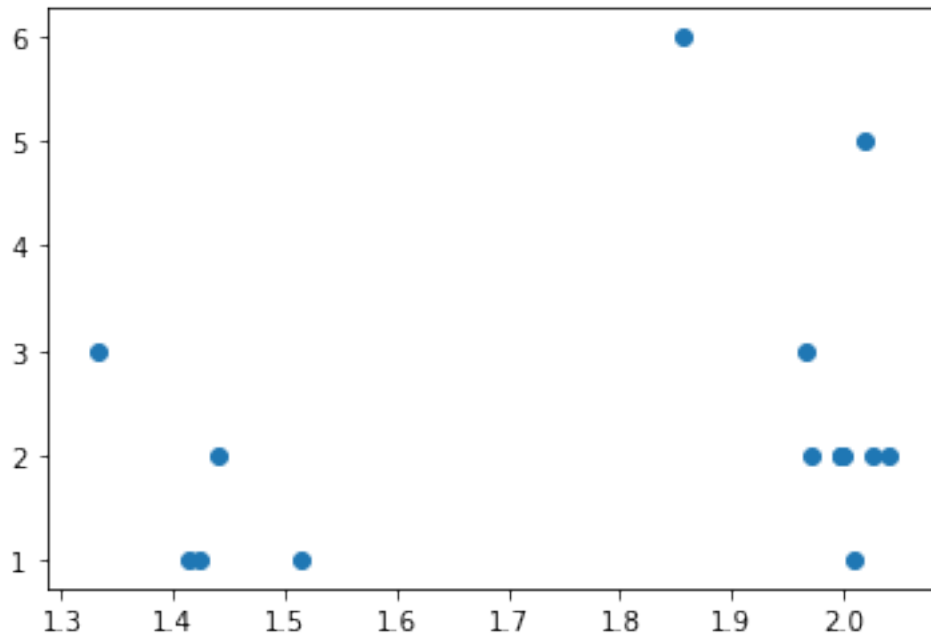
```
[867]: df_pr = pd.DataFrame({'Actual': y_post_valid, 'Predicted': ctb_par_post.
        ↪predict(X_post_valid)})
       df_pr.head()
```

```
[867]:      Actual  Predicted
       64      1.0   1.422452
       209     2.0   1.998622
       274     3.0   1.966932
       35      2.0   1.972224
       227     2.0   1.998913
```

```
[868]: df_pr.plot(kind='bar',figsize=(10,5))
       plt.grid(which='major', linestyle='-', linewidth='0.5', color='green')
       plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
       plt.show()
```
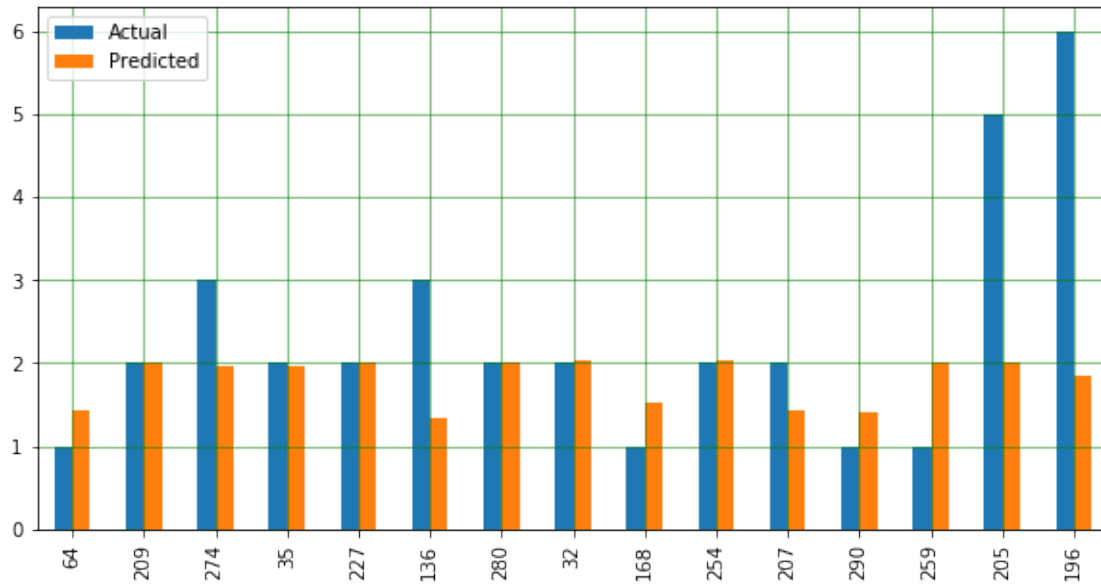
[869]: `X_post_valid.head()`

[869]:

	Postamat_trend	hh_500	POPULATION	NEAR_Malls	NEAR_Business_centers	\
64	0	2127	324698	0	0	
209	0	4632	58139	0	0	
274	0	2115	482873	0	0	
35	1	4412	104739	1	0	
227	0	3365	61732	0	0	

	NEAR_metro_rjd	NEAR_Stations	work_hours	K	T	values
64	1	1	15	1	0	3.707463
209	0	0	15	0	1	3.153342
274	0	0	15	0	1	3.597627
35	0	1	15	1	0	3.287533
227	0	0	15	0	1	3.287064

[870]: `df_post_purpose.head()`

[870]:

	Postamat_trend	hh_500	POPULATION	NEAR_Malls	NEAR_Business_centers	\
0	NaN	7696.0	324698	0	0	
1	NaN	6369.0	324698	1	1	
2	NaN	5553.0	324698	1	0	
5	NaN	7703.0	324698	0	0	
7	NaN	6487.0	324698	0	0	

	NEAR_metro_rjd	NEAR_Stations	work_hours	K	T	values
0	0	1	24	1	0	2.986103
1	0	1	24	1	0	3.111248
2	0	0	15	1	0	2.939464

14

```
5               0               1          19  1  0  3.110700
7               0               1          24  1  0  3.160301
```

```
[895]:  df_post_purpose.head()
        df_post_purpose['Postamat_trend'].fillna(1, inplace=True)
        df_post_purpose['Postamat_trend'] = df_post_purpose['Postamat_trend'].
        ↪astype('int64')
        df_post_purpose['NEAR_Malls'] = df_post_purpose['NEAR_Malls'].astype('str')
        df_post_purpose['NEAR_Business_centers'] =␣
        ↪df_post_purpose['NEAR_Business_centers'].astype('str')
        df_post_purpose['NEAR_metro_rjd'] = df_post_purpose['NEAR_metro_rjd'].
        ↪astype('str')
        df_post_purpose['NEAR_Stations'] = df_post_purpose['NEAR_Stations'].
        ↪astype('str')
        df_post_purpose['K'] = df_post_purpose['K'].astype('str')
        df_post_purpose['T'] = df_post_purpose['T'].astype('str')

        df_post_purpose['hh_500'] = df_post_purpose['hh_500'].astype('int64')
```

```
[896]:  df_post_purpose.isna().mean()
```

```
[896]:  level_0                 0.0
        index                   0.0
        Postamat_trend          0.0
        hh_500                  0.0
        POPULATION              0.0
        NEAR_Malls              0.0
        NEAR_Business_centers   0.0
        NEAR_metro_rjd          0.0
        NEAR_Stations           0.0
        work_hours              0.0
        K                       0.0
        T                       0.0
        values                  0.0
        dtype: float64
```

```
[901]:  df_post_purpose.head()
```

```
[901]:     level_0  index  Postamat_trend  hh_500  POPULATION NEAR_Malls  \
        0        0      0               1    7696      324698          0
        1        1      1               1    6369      324698          1
        2        2      2               1    5553      324698          1
        3        3      5               1    7703      324698          0
        4        4      7               1    6487      324698          0

           NEAR_Business_centers NEAR_metro_rjd NEAR_Stations  work_hours  K  T  \
        0                      0              0             0           1  24  1  0
        1                      1              1             0           1  24  1  0
        2                                     0             0           0  15  1  0
```

```
3                    0            0            1        19  1  0
4                    0            0            1        24  1  0

      values
0  2.986103
1  3.111248
2  2.939464
3  3.110700
4  3.160301
```

[898]:
```python
df_post_purpose.reset_index(inplace=True)
```

```
    ␣
 ↪---------------------------------------------------------------------------

      ValueError                                Traceback (most recent call␣
 ↪last)

      <ipython-input-898-d17e335badaa> in <module>
 ----> 1 df_post_purpose.reset_index(inplace=True)


      D:\Anaconda3\lib\site-packages\pandas\core\frame.py in reset_index(self,␣
 ↪level, drop, inplace, col_level, col_fill)
    4429                    # to ndarray and maybe infer different dtype
    4430                    level_values = _maybe_casted_values(lev, lab)
 -> 4431                    new_obj.insert(0, name, level_values)
    4432
    4433          new_obj.index = new_index


      D:\Anaconda3\lib\site-packages\pandas\core\frame.py in insert(self, loc,␣
 ↪column, value, allow_duplicates)
    3471          value = self._sanitize_column(column, value, broadcast=False)
    3472          self._data.insert(loc, column, value,
 -> 3473                            allow_duplicates=allow_duplicates)
    3474
    3475      def assign(self, **kwargs):


      D:\Anaconda3\lib\site-packages\pandas\core\internals\managers.py in␣
 ↪insert(self, loc, item, value, allow_duplicates)
    1147          if not allow_duplicates and item in self.items:
    1148              # Should this be a different kind of error??
 -> 1149              raise ValueError('cannot insert {}, already exists'.
 ↪format(item))
```

```
1150
1151              if not isinstance(loc, int):


     ValueError: cannot insert level_0, already exists
```

[900]:
```python
df_post_purpose_new = df_post_purpose.copy()
df_post_purpose_new = pd.concat([df_post_purpose, pd.
 ↪Series(Postamat_daily_new_v1).rename(columns={0: "Postamat_daily"},␣
 ↪inplace=True)], axis=1)
df_post_purpose_new.rename(columns={0: "Postamat_daily"}, inplace=True)
df_post_purpose_new = df_post_purpose_new.drop(['index'], axis=1)
df_post_purpose_new.head()
```

[900]:
```
   level_0  Postamat_trend  hh_500  POPULATION  NEAR_Malls  \
0        0               1    7696      324698           0
1        1               1    6369      324698           1
2        2               1    5553      324698           1
3        3               1    7703      324698           0
4        4               1    6487      324698           0

   NEAR_Business_centers  NEAR_metro_rjd  NEAR_Stations  work_hours  K  T  \
0                      0               0              1          24  1  0
1                      1               0              1          24  1  0
2                      0               0              0          15  1  0
3                      0               0              1          19  1  0
4                      0               0              1          24  1  0

      values  Postamat_daily
0   2.986103        1.926006
1   3.111248        1.403362
2   2.939464        1.883915
3   3.110700        2.030870
4   3.160301        2.004101
```

[894]:
```python
cols = df_post_purpose_new.columns.tolist()
cols = cols[-1:] + cols[:-1]
print(cols)
df_post_purpose_new = df_post_purpose_new[cols]
df_post_purpose_new.head()
```

```
['Postamat_daily', 'level_0', 'Postamat_trend', 'hh_500', 'POPULATION',
'NEAR_Malls', 'NEAR_Business_centers', 'NEAR_metro_rjd', 'NEAR_Stations',
'work_hours', 'K', 'T', 'values']
```

[894]:
```
   Postamat_daily  level_0  Postamat_trend  hh_500  POPULATION NEAR_Malls  \
0        1.926006        0               1    7696      324698          0
```

1	1.403362	1	1	6369	324698	1
2	1.883915	2	1	5553	324698	1
3	2.030870	3	1	7703	324698	0
4	2.004101	4	1	6487	324698	0

	NEAR_Business_centers	NEAR_metro_rjd	NEAR_Stations	work_hours	K	T \
0	0	0	0	1	24	1	0
1	1	1	0	1	24	1	0
2	2	0	0	0	15	1	0
3	3	0	0	1	19	1	0
4	4	0	0	1	24	1	0

	values
0	2.986103
1	3.111248
2	2.939464
3	3.110700
4	3.160301

```
[889]: # print(df_post_train.shape)
       # print(df_post_purpose.shape)
       df_post_train
```

[889]:

	Postamat_daily	Postamat_trend	hh_500	POPULATION	NEAR_Malls \
3	3.0	0.0	7162.0	324698	0
4	1.0	0.0	6269.0	324698	1
6	3.0	0.0	5632.0	324698	0
8	2.0	0.0	5156.0	324698	0
10	5.0	1.0	5232.0	324698	0
20	4.0	0.0	1834.0	324698	0
21	4.0	0.0	4219.0	324698	1
24	6.0	0.0	2944.0	324698	0
25	3.0	1.0	3093.0	324698	0
32	2.0	1.0	3105.0	324698	0
34	2.0	1.0	5619.0	324698	0
35	2.0	1.0	4412.0	104739	1
43	3.0	0.0	4573.0	324698	0
44	4.0	1.0	4357.0	324698	0
45	2.0	0.0	5602.0	104739	1
55	2.0	0.0	2413.0	324698	1
56	2.0	0.0	3170.0	104739	0
60	2.0	0.0	1493.0	324698	0
61	2.0	0.0	2164.0	324698	0
62	2.0	1.0	887.0	324698	1
63	1.0	1.0	4937.0	324698	0
64	1.0	0.0	2127.0	324698	0
74	2.0	0.0	2544.0	324698	0
75	1.0	0.0	1950.0	31882	0

78	2.0	1.0	2455.0	104739	0
82	1.0	0.0	1692.0	324698	0
94	1.0	0.0	859.0	324698	0
107	2.0	1.0	3316.0	104739	0
126	2.0	0.0	723.0	30392	0
136	3.0	1.0	1167.0	30392	0
..	...	...	...	...	...
207	2.0	1.0	8114.0	482873	0
209	2.0	0.0	4632.0	58139	0
210	2.0	0.0	2166.0	482873	1
213	2.0	0.0	5903.0	58139	1
214	3.0	1.0	3552.0	482873	0
215	2.0	1.0	5137.0	482873	0
218	3.0	1.0	5423.0	482873	0
219	3.0	1.0	9775.0	482873	0
224	4.0	1.0	6226.0	482873	0
225	2.0	0.0	6274.0	482873	0
227	2.0	0.0	3365.0	61732	0
232	2.0	0.0	6772.0	482873	0
235	3.0	0.0	2042.0	482873	1
244	2.0	0.0	3991.0	55282	0
247	2.0	1.0	3900.0	482873	0
248	1.0	0.0	3648.0	482873	0
249	2.0	0.0	3338.0	482873	0
250	4.0	0.0	4081.0	482873	0
254	2.0	0.0	2831.0	482873	0
259	1.0	0.0	2639.0	131386	0
265	4.0	0.0	5457.0	482873	0
274	3.0	0.0	2115.0	482873	0
279	3.0	1.0	3834.0	131386	0
280	2.0	0.0	3852.0	131386	0
290	1.0	0.0	1042.0	482873	0
297	2.0	0.0	3109.0	55282	1
318	2.0	1.0	42.0	482873	0
332	1.0	0.0	1389.0	61732	0
335	1.0	0.0	1457.0	28485	0
336	1.0	0.0	2691.0	28485	0

	NEAR_Business_centers	NEAR_metro_rjd	NEAR_Stations	work_hours	K	T	\
3	0	0	0	15	1	0	
4	0	0	1	15	1	0	
6	0	0	0	24	1	0	
8	0	0	0	19	1	0	
10	0	0	0	15	1	0	
20	0	0	0	15	1	0	
21	0	0	0	24	1	0	
24	0	0	0	24	1	0	

25	0	0	1	19	1	0
32	0	0	0	24	1	0
34	0	0	0	14	1	0
35	0	0	1	15	1	0
43	0	0	1	15	1	0
44	0	0	1	24	1	0
45	0	0	0	24	1	0
55	0	0	0	24	1	0
56	0	0	0	15	1	0
60	0	0	0	15	1	0
61	0	0	1	15	1	0
62	0	0	1	13	1	0
63	0	0	0	15	1	0
64	0	1	1	15	1	0
74	0	0	0	15	1	0
75	0	0	0	15	1	0
78	0	0	0	15	1	0
82	0	0	1	15	1	0
94	0	0	1	15	1	0
107	0	0	1	15	1	0
126	0	0	0	15	1	0
136	0	0	0	15	1	0
..	...	...	...	...	..	..
207	1	0	0	15	0	1
209	0	0	0	15	0	1
210	1	0	1	15	0	1
213	0	0	0	15	0	1
214	0	0	1	15	0	1
215	0	0	0	15	0	1
218	0	0	0	15	0	1
219	0	0	1	15	0	1
224	0	0	0	15	0	1
225	0	0	0	14	0	1
227	0	0	0	15	0	1
232	0	0	0	15	0	1
235	0	1	1	15	0	1
244	0	0	0	15	0	1
247	0	0	1	14	0	1
248	0	0	0	15	0	1
249	0	0	1	15	0	1
250	0	0	1	15	0	1
254	0	0	0	15	0	1
259	0	0	1	15	0	1
265	0	0	0	15	0	1
274	0	0	0	15	0	1
279	0	0	0	15	0	1
280	0	0	0	15	0	1

290	0	0	1	15	0	1
297	0	0	1	15	0	1
318	0	0	1	15	0	1
332	0	0	0	15	0	1
335	0	0	0	15	0	1
336	0	0	0	15	0	1

	values
3	3.100706
4	3.159937
6	2.967969
8	3.127633
10	3.080232
20	3.152310
21	3.283665
24	3.174528
25	3.281301
32	3.281016
34	3.422593
35	3.287533
43	3.374455
44	3.430669
45	3.469796
55	3.486993
56	3.468459
60	3.569242
61	3.519116
62	3.555358
63	3.682945
64	3.707463
74	3.571326
75	3.323558
78	3.547829
82	3.689822
94	3.610145
107	3.702230
126	3.585912
136	3.583840
..	...
207	3.581120
209	3.153342
210	3.285339
213	3.257762
214	3.223251
215	3.360481
218	3.428864
219	3.609641

```
224  3.696409
225  3.586691
227  3.287064
232  3.689425
235  3.687709
244  3.311740
247  3.519339
248  3.528149
249  3.533779
250  3.568060
254  3.547096
259  3.444025
265  3.760121
274  3.597627
279  3.513399
280  3.669248
290  3.599349
297  3.629197
318  3.658029
332  3.572760
335  3.641737
336  3.638383

[72 rows x 12 columns]
```

[ ]:
```python
df_new_ver1 = df
```

0.4   cashbox_daily

[602]:
```python
df_cashb_purpose = dff[dff.isna().cashbox_daily == True].copy()
df_cashb_train = dff[dff.isna().cashbox_daily == False].copy()
print('cashb purpose:', df_cashb_purpose.shape[0])
print('cashb train:', df_cashb_train.shape[0])
```

```
cashb purpose: 264
cashb train: 153
```

0.5   cashbox_daily

[603]:
```python
df_cashb_train.head()
```

[603]:
```
    Postamat_daily  cashbox_daily  Postamat_trend  cashbox_MAX  cashbox_trend  \
0            NaN            3.3             NaN          13.0            1.0
2            NaN            3.1             NaN          12.0            1.0
3            3.0            2.2             0.0          12.0            1.0
4            1.0            2.0             0.0           9.0            1.0
5            NaN            1.2             NaN           6.0            1.0
```

```
     hh_500  POPULATION  NEAR_Malls  NEAR_Business_centers  NEAR_metro_rjd  \
0    7696.0      324698           0                      0               0
2    5553.0      324698           1                      0               0
3    7162.0      324698           0                      0               0
4    6269.0      324698           1                      0               0
5    7703.0      324698           0                      0               0

   NEAR_Stations  macro_salary_avg_yearly  work_hours  K  T    values
0              1                   423321          24  1  0  2.986103
2              0                   423321          15  1  0  2.939464
3              0                   423321          15  1  0  3.100706
4              1                   423321          15  1  0  3.159937
5              1                   423321          19  1  0  3.110700
```

[604]:
```python
df_cashb_train.isna().mean()
```

[604]:
```
Postamat_daily           0.764706
cashbox_daily            0.000000
Postamat_trend           0.764706
cashbox_MAX              0.000000
cashbox_trend            0.084967
hh_500                   0.000000
POPULATION               0.000000
NEAR_Malls               0.000000
NEAR_Business_centers    0.000000
NEAR_metro_rjd           0.000000
NEAR_Stations            0.000000
macro_salary_avg_yearly  0.000000
work_hours               0.000000
K                        0.000000
T                        0.000000
values                   0.000000
dtype: float64
```

[605]:
```python
df_cashb_train = df_cashb_train.
 ↪drop(columns=['Postamat_daily','Postamat_trend','macro_salary_avg_yearly'])
```

[606]:
```python
df_cashb_train.dropna(subset=['cashbox_trend'], inplace=True)
```

[607]:
```python
print(df_cashb_train.shape)
print(df_cashb_purpose.shape)
```

```
(140, 13)
(264, 16)
```

0.5.1 X_CASHB

```
[608]: X_cashb = df_cashb_train.drop('cashbox_daily', axis=1)
       y_cashb = df_cashb_train['cashbox_daily']
       print(X_cashb.shape, y_cashb.shape)
```

```
(140, 12) (140,)
```

```
[609]: X_train_part.head()
```

[609]:		Postamat_trend	hh_500	POPULATION	NEAR_Malls	NEAR_Business_centers	\
	56	0	3170	104739	0	0	
	232	0	6772	482873	0	0	
	61	0	2164	324698	0	0	
	10	1	5232	324698	0	0	
	126	0	723	30392	0	0	

	NEAR_metro_rjd	NEAR_Stations	work_hours	K	T	values
56	0	0	15	1	0	3.468459
232	0	0	15	0	1	3.689425
61	0	1	15	1	0	3.519116
10	0	0	15	1	0	3.080232
126	0	0	15	1	0	3.585912

0.6 CATBOOST

```
[610]: X_cashb.head()
```

[610]:		cashbox_MAX	cashbox_trend	hh_500	POPULATION	NEAR_Malls	\
	0	13.0	1.0	7696.0	324698	0	
	2	12.0	1.0	5553.0	324698	1	
	3	12.0	1.0	7162.0	324698	0	
	4	9.0	1.0	6269.0	324698	1	
	5	6.0	1.0	7703.0	324698	0	

	NEAR_Business_centers	NEAR_metro_rjd	NEAR_Stations	work_hours	K	T	\
0	0	0	1	24	1	0	
2	0	0	0	15	1	0	
3	0	0	0	15	1	0	
4	0	0	1	15	1	0	
5	0	0	1	19	1	0	

	values
0	2.986103
2	2.939464
3	3.100706
4	3.159937
5	3.110700

```
[611]: X_cashb.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 140 entries, 0 to 375
Data columns (total 12 columns):
cashbox_MAX             140 non-null float64
cashbox_trend           140 non-null float64
hh_500                  140 non-null float64
POPULATION              140 non-null int64
NEAR_Malls              140 non-null int64
NEAR_Business_centers   140 non-null int64
NEAR_metro_rjd          140 non-null int64
NEAR_Stations           140 non-null int64
work_hours              140 non-null int64
K                       140 non-null uint8
T                       140 non-null uint8
values                  140 non-null float64
dtypes: float64(4), int64(6), uint8(2)
memory usage: 12.3 KB
```

```
[612]: X_cashb['cashbox_MAX'] = X_cashb['cashbox_MAX'].astype('int64')
       X_cashb['cashbox_trend'] = X_cashb['cashbox_trend'].astype('int64')
       X_cashb['NEAR_Malls'] = X_cashb['NEAR_Malls'].astype('str')
       X_cashb['NEAR_Business_centers'] = X_cashb['NEAR_Business_centers'].
        ↪astype('str')
       X_cashb['NEAR_metro_rjd'] = X_cashb['NEAR_metro_rjd'].astype('str')
       X_cashb['NEAR_Stations'] = X_cashb['NEAR_Stations'].astype('str')
       X_cashb['K'] = X_cashb['K'].astype('str')
       X_cashb['T'] = X_cashb['T'].astype('str')


       X_cashb['hh_500'] = X_cashb['hh_500'].astype('int64')
```

```
[613]: X_cashb.head()
```

```
[613]:    cashbox_MAX   cashbox_trend   hh_500   POPULATION  NEAR_Malls  \
       0           13               1     7696       324698           0
       2           12               1     5553       324698           1
       3           12               1     7162       324698           0
       4            9               1     6269       324698           1
       5            6               1     7703       324698           0

          NEAR_Business_centers NEAR_metro_rjd NEAR_Stations  work_hours  K  T  \
       0                      0              0             0          24  1  0
       2                      0              0             0          15  1  0
       3                      0              0             0          15  1  0
       4                      0              0             1          15  1  0
       5                      0              0             1          19  1  0
```

```
          values
0   2.986103
2   2.939464
3   3.100706
4   3.159937
5   3.110700
```

[614]:
```
X_cashb_train_part, X_cashb_valid, y_cashb_train_part, y_cashb_valid =␣
 ↪train_test_split(X_cashb, y_cashb,
                                                          test_size=0.3,
                                                     ␣
 ↪random_state=SEED)
print(X_cashb_train_part.shape[0], X_cashb_valid.shape[0], y_cashb_train_part.
 ↪shape[0], y_cashb_valid.shape[0])
```

```
98 42 98 42
```

[615]:
```
X_cashb_train_part.head()
```

[615]:
```
     cashbox_MAX  cashbox_trend  hh_500  POPULATION  NEAR_Malls  \
65             4              1    3595      324698           0
117            5              1    3044       40530           0
52            12              1    2957      324698           0
136           12              1    1167       30392           0
50            10              0    6045      324698           0

     NEAR_Business_centers  NEAR_metro_rjd  NEAR_Stations  work_hours  K  T  \
65                       0               0              0          15  1  0
117                      0               0              0          15  1  0
52                       0               0              0          15  1  0
136                      0               0              0          15  1  0
50                       0               0              0          15  1  0

        values
65    3.608397
117   3.633138
52    3.401080
136   3.583840
50    3.584831
```

[616]:
```
categ_cashb_feat_idx = ['cashbox_trend', 'NEAR_Malls', 'NEAR_Business_centers',␣
 ↪'NEAR_metro_rjd', 'NEAR_Stations', 'K', 'T']
```

[617]:
```
params = {'loss_function':'MAE',
          'eval_metric':'MAE',
          'cat_features': categ_cashb_feat_idx,
          'verbose': 200,
          'random_seed': SEED,
          'early_stopping_rounds': 100,
```

```
            'boosting_type': 'Ordered',
            'bootstrap_type':'Bayesian',
            'iterations': 1000,
            'task_type':'GPU',
            'learning_rate':0.01,
            'leaf_estimation_iterations':5,
            'max_ctr_complexity':4
        }
ctb_par_cashb = CatBoostRegressor(**params)
ctb_par_cashb.fit(X_cashb_train_part, y_cashb_train_part,
        eval_set=(X_cashb_valid, y_cashb_valid),
        use_best_model=True,
        plot=True
        );
```

<IPython.core.display.HTML object>

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


```
0:      learn: 1.3640289        test: 1.5264791 best: 1.5264791 (0)     total:
37.5ms   remaining: 37.5s
200:    learn: 0.4002868        test: 0.5928939 best: 0.5928939 (200)   total:
7.92s    remaining: 31.5s
400:    learn: 0.2336266        test: 0.4289981 best: 0.4289981 (400)   total:
15.8s    remaining: 23.6s
600:    learn: 0.1922992        test: 0.3956473 best: 0.3956473 (600)   total:
24.5s    remaining: 16.3s
800:    learn: 0.1744510        test: 0.3908632 best: 0.3905067 (769)   total:
33.3s    remaining: 8.27s
999:    learn: 0.1587092        test: 0.3902684 best: 0.3901249 (985)   total:
42.6s    remaining: 0us
bestTest = 0.3901249114
bestIteration = 985
Shrink model to first 986 iterations.
```

[618]:
```
ctb_par_cashb.get_feature_importance(prettified=True)
feature_importance_df = pd.DataFrame(ctb_par_cashb.
 →get_feature_importance(prettified=True), columns=['Feature Id',␣
 →'Importances'])
feature_importance_df
```

[618]:
```
              Feature Id  Importances
0            cashbox_MAX    61.101304
1                hh_500    12.223296
2                values     5.392169
3            POPULATION     5.375006
```

```
4              work_hours      3.663616
5                       T      3.520352
6                       K      1.964651
7    NEAR_Business_centers      1.950239
8              NEAR_Malls      1.702288
9            cashbox_trend      1.677882
10           NEAR_Stations      0.738123
11           NEAR_metro_rjd      0.691073
```
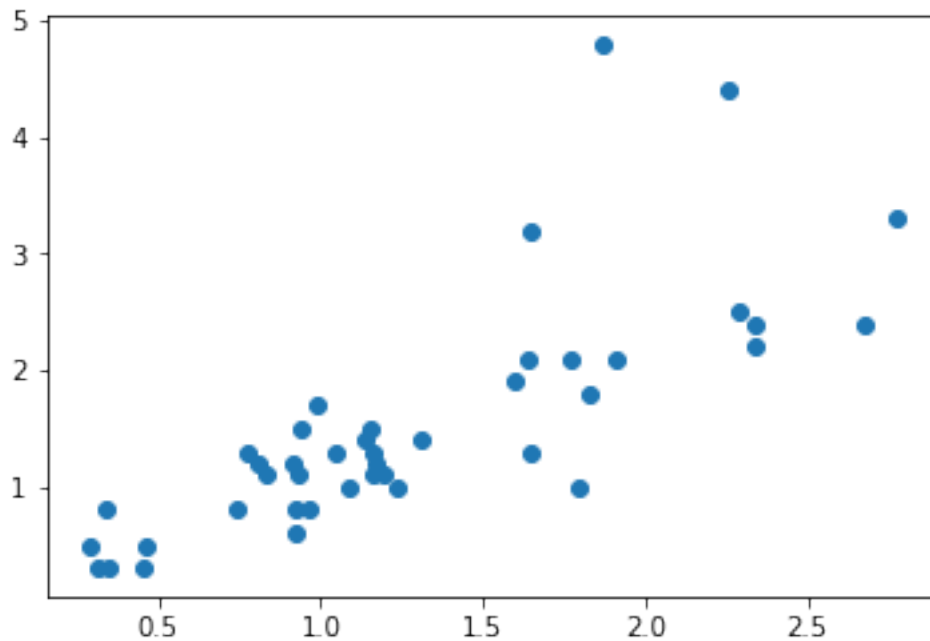
[619]: `plt.scatter(ctb_par_cashb.predict(X_cashb_valid), y_cashb_valid)`

[619]: `<matplotlib.collections.PathCollection at 0x19704e43cc0>`

[620]:
```
df_pr = pd.DataFrame({'Actual': y_cashb_valid, 'Predicted': ctb_par_cashb.
 ↪predict(X_cashb_valid)})
df_pr.head()
```

[620]:
```
     Actual  Predicted
95      0.3   0.450659
288     1.5   1.149937
122     1.9   1.593992
93      1.5   0.936794
251     1.1   0.932991
```

[621]:
```
df_pr.plot(kind='bar',figsize=(10,5))
plt.grid(which='major', linestyle='-', linewidth='0.5', color='green')
plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
plt.show()
```
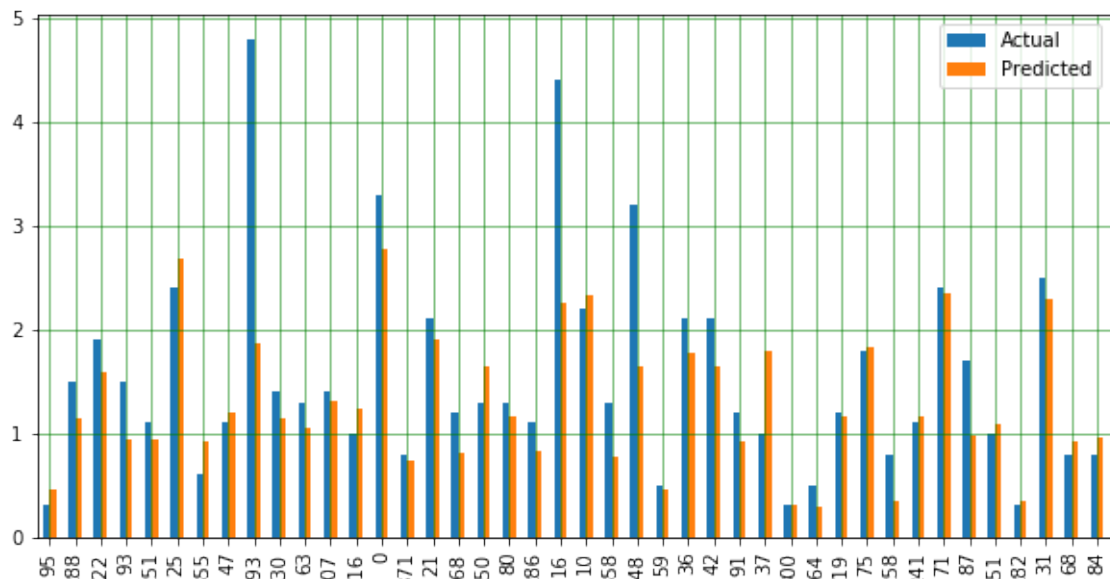
[622]: `X_cashb_valid.head()`

[622]:

	cashbox_MAX	cashbox_trend	hh_500	POPULATION	NEAR_Malls	\
95	3	1	420	324698	0	
288	7	1	4692	25557	0	
122	11	1	1663	104739	0	
93	5	1	3190	104739	0	
251	6	1	3295	131386	0	

	NEAR_Business_centers	NEAR_metro_rjd	NEAR_Stations	work_hours	K	T	\
95	0	0	1	15	1	0	
288	0	0	0	15	0	1	
122	0	0	0	15	1	0	
93	0	0	0	15	1	0	
251	0	0	0	15	0	1	

	values
95	3.639303
288	3.590093
122	3.786636
93	3.756318
251	3.488680

0.7 predict

[623]: `X_cashb_train_part.head()`

```
[623]:      cashbox_MAX  cashbox_trend  hh_500  POPULATION NEAR_Malls  \
      65              4              1    3595      324698           0
      117             5              1    3044       40530           0
      52             12              1    2957      324698           0
      136            12              1    1167       30392           0
      50             10              0    6045      324698           0

           NEAR_Business_centers NEAR_metro_rjd NEAR_Stations  work_hours  K  T  \
      65                       0              0             0          15  1  0
      117                      0              0             0          15  1  0
      52                       0              0             0          15  1  0
      136                      0              0             0          15  1  0
      50                       0              0             0          15  1  0

             values
      65    3.608397
      117   3.633138
      52    3.401080
      136   3.583840
      50    3.584831
```

```python
[624]: df_cashb_purpose = df_cashb_purpose.drop(['Postamat_daily', 'Postamat_trend',
       ↪'cashbox_daily',
                                           'macro_salary_avg_yearly'], axis=1)
       df_cashb_purpose['cashbox_trend'].fillna(1, inplace=True)
       df_cashb_purpose.fillna((df_cashb_purpose['hh_500'].median()), inplace=True)
       df_cashb_purpose['cashbox_MAX'].fillna(df.cashbox_MAX.median(), inplace=True)
       df_cashb_purpose.head()
```

```
[624]:     cashbox_MAX  cashbox_trend  hh_500  POPULATION  NEAR_Malls  \
      1        1464.0            1.0  6369.0      324698           1
      11       1464.0            1.0  7552.0      324698           0
      12       1464.0            1.0  3773.0      324698           0
      13       1464.0            1.0  4982.0      104739           0
      14       1464.0            1.0  7836.0      104739           0

           NEAR_Business_centers  NEAR_metro_rjd  NEAR_Stations  work_hours  K  T  \
      1                        1               0              1          24  1  0
      11                       1               0              0          24  1  0
      12                       1               0              1          20  1  0
      13                       1               0              1          15  1  0
      14                       0               0              1          24  1  0

             values
      1     3.111248
      11    3.407810
      12    3.241490
      13    3.193099
```

```
14  3.269593
```

[625]:
```python
df_cashb_purpose['cashbox_MAX'] = df_cashb_purpose['cashbox_MAX'].
 ↪astype('int64')
df_cashb_purpose['cashbox_trend'] = df_cashb_purpose['cashbox_trend'].
 ↪astype('int64')
df_cashb_purpose['NEAR_Malls'] = df_cashb_purpose['NEAR_Malls'].astype('str')
df_cashb_purpose['NEAR_Business_centers'] =␣
 ↪df_cashb_purpose['NEAR_Business_centers'].astype('str')
df_cashb_purpose['NEAR_metro_rjd'] = df_cashb_purpose['NEAR_metro_rjd'].
 ↪astype('str')
df_cashb_purpose['NEAR_Stations'] = df_cashb_purpose['NEAR_Stations'].
 ↪astype('str')
df_cashb_purpose['K'] = df_cashb_purpose['K'].astype('str')
df_cashb_purpose['T'] = df_cashb_purpose['T'].astype('str')

df_cashb_purpose['hh_500'] = df_cashb_purpose['hh_500'].astype('int64')
```

[626]:
```python
df_cashb_purpose.head()
```

[626]:
```
    cashbox_MAX  cashbox_trend  hh_500  POPULATION NEAR_Malls  \
1          1464              1    6369      324698          1
11         1464              1    7552      324698          0
12         1464              1    3773      324698          0
13         1464              1    4982      104739          0
14         1464              1    7836      104739          0

   NEAR_Business_centers NEAR_metro_rjd NEAR_Stations  work_hours  K  T  \
1                      1              0             1          24  1  0
11                     1              0             0          24  1  0
12                     1              0             1          20  1  0
13                     1              0             1          15  1  0
14                     0              0             1          24  1  0

      values
1   3.111248
11  3.407810
12  3.241490
13  3.193099
14  3.269593
```

[630]:
```python
cashb_predict_ver1 = ctb_par_cashb.predict(df_cashb_purpose)
```

1 2 -

```
[631]: df_post_purpose = dff[dff.isna().Postamat_daily == True].copy()
       df_post_train = dff[dff.isna().Postamat_daily == False].copy()
       print('post purpose:', df_post_purpose.shape[0])
       print('post train:', df_post_train.shape[0])
```

post purpose: 345
post train: 72

1.1 Postamat_daily

```
[632]: df_post_purpose = dff[dff.isna().Postamat_daily == True].copy()
       df_post_train = dff[dff.isna().Postamat_daily == False].copy()
       print('post purpose:', df_post_purpose.shape[0])
       print('post train:', df_post_train.shape[0])
```

post purpose: 345
post train: 72

1.2 Postamat_daily

```
[633]: df_post_train.head()
```

```
[633]:    Postamat_daily  cashbox_daily  Postamat_trend  cashbox_MAX  cashbox_trend  \
       3             3.0            2.2             0.0         12.0            1.0
       4             1.0            2.0             0.0          9.0            1.0
       6             3.0            2.2             0.0          9.0            1.0
       8             2.0            1.2             0.0          7.0            1.0
       10            5.0            2.2             1.0         11.0            1.0

           hh_500  POPULATION  NEAR_Malls  NEAR_Business_centers  NEAR_metro_rjd  \
       3   7162.0      324698           0                      0               0
       4   6269.0      324698           1                      0               0
       6   5632.0      324698           0                      0               0
       8   5156.0      324698           0                      0               0
       10  5232.0      324698           0                      0               0

           NEAR_Stations  macro_salary_avg_yearly  work_hours   K  T    values
       3               0                   423321          15   1  0  3.100706
       4               1                   423321          15   1  0  3.159937
       6               0                   423321          24   1  0  2.967969
       8               0                   423321          19   1  0  3.127633
       10              0                   423321          15   1  0  3.080232
```

```
[634]: df_post_train = df_post_train.
       ↪drop(columns=['Postamat_trend','cashbox_daily','cashbox_MAX','cashbox_trend',␣
       ↪'macro_salary_avg_yearly'])
```

32

```
[635]: df_post_train.head()
```

```
[635]:     Postamat_daily  hh_500  POPULATION  NEAR_Malls  NEAR_Business_centers  \
       3              3.0  7162.0      324698           0                      0
       4              1.0  6269.0      324698           1                      0
       6              3.0  5632.0      324698           0                      0
       8              2.0  5156.0      324698           0                      0
       10             5.0  5232.0      324698           0                      0

           NEAR_metro_rjd  NEAR_Stations  work_hours   K  T     values
       3                0              0          15   1  0  3.100706
       4                0              1          15   1  0  3.159937
       6                0              0          24   1  0  2.967969
       8                0              0          19   1  0  3.127633
       10               0              0          15   1  0  3.080232
```

```
[636]: df_post_purpose=df_post_purpose.drop(columns=['Postamat_trend',
       ↪'Postamat_daily','cashbox_daily','cashbox_MAX','cashbox_trend',
       ↪'macro_salary_avg_yearly'])
       df_post_purpose['hh_500'].fillna((df_post_purpose['hh_500'].mean()),
       ↪inplace=True)
```

```
[637]: df_post_purpose.head()
```

```
[637]:    hh_500  POPULATION  NEAR_Malls  NEAR_Business_centers  NEAR_metro_rjd  \
       0  7696.0      324698           0                      0               0
       1  6369.0      324698           1                      1               0
       2  5553.0      324698           1                      0               0
       5  7703.0      324698           0                      0               0
       7  6487.0      324698           0                      0               0

          NEAR_Stations  work_hours   K  T     values
       0              1          24   1  0  2.986103
       1              1          24   1  0  3.111248
       2              0          15   1  0  2.939464
       5              1          19   1  0  3.110700
       7              1          24   1  0  3.160301
```

```
[638]: print(df_post_train.shape)
       print(df_post_purpose.shape)
```

```
(72, 11)
(345, 10)
```

1.2.1 X_POST

```
[639]: X_post = df_post_train.drop(['Postamat_daily'] , axis=1)
       y_post = df_post_train['Postamat_daily']
       print(X_post.shape, y_post.shape)
```

33

```
(72, 10) (72,)
```

1.3 CATBOOST

```
[640]: X_post.head()
```

```
[640]:      hh_500  POPULATION  NEAR_Malls  NEAR_Business_centers  NEAR_metro_rjd  \
       3    7162.0      324698           0                      0               0
       4    6269.0      324698           1                      0               0
       6    5632.0      324698           0                      0               0
       8    5156.0      324698           0                      0               0
       10   5232.0      324698           0                      0               0

            NEAR_Stations  work_hours  K  T     values
       3                 0          15  1  0   3.100706
       4                 1          15  1  0   3.159937
       6                 0          24  1  0   2.967969
       8                 0          19  1  0   3.127633
       10                0          15  1  0   3.080232
```

```
[641]: X_post.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 72 entries, 3 to 336
Data columns (total 10 columns):
hh_500                 72 non-null float64
POPULATION             72 non-null int64
NEAR_Malls             72 non-null int64
NEAR_Business_centers  72 non-null int64
NEAR_metro_rjd         72 non-null int64
NEAR_Stations          72 non-null int64
work_hours             72 non-null int64
K                      72 non-null uint8
T                      72 non-null uint8
values                 72 non-null float64
dtypes: float64(2), int64(6), uint8(2)
memory usage: 5.2 KB
```

```
[642]: X_post['NEAR_Malls'] = X_post['NEAR_Malls'].astype('str')
       X_post['NEAR_Business_centers'] = X_post['NEAR_Business_centers'].astype('str')
       X_post['NEAR_metro_rjd'] = X_post['NEAR_metro_rjd'].astype('str')
       X_post['NEAR_Stations'] = X_post['NEAR_Stations'].astype('str')
       X_post['K'] = X_post['K'].astype('str')
       X_post['T'] = X_post['T'].astype('str')

       X_post['hh_500'] = X_post['hh_500'].astype('int64')
```

```
[643]: X_post.head()
```

```
[643]:      hh_500  POPULATION NEAR_Malls NEAR_Business_centers NEAR_metro_rjd  \
       3    7162      324698           0                     0              0
       4    6269      324698           1                     0              0
       6    5632      324698           0                     0              0
       8    5156      324698           0                     0              0
       10   5232      324698           0                     0              0

           NEAR_Stations  work_hours  K  T    values
       3                0          15  1  0  3.100706
       4                1          15  1  0  3.159937
       6                0          24  1  0  2.967969
       8                0          19  1  0  3.127633
       10               0          15  1  0  3.080232
```

```python
[644]: X_post_train_part, X_post_valid, y_post_train_part, y_post_valid =␣
       ↪train_test_split(X_post, y_post,
                                                              test_size=0.2,
                                                         ␣
       ↪random_state=SEED)
       print(X_post_train_part.shape[0], X_post_valid.shape[0], y_post_train_part.
       ↪shape[0], y_valid.shape[0])
```

```
57 15 57 15
```

```python
[645]: categ_feat_idx = ['NEAR_Malls', 'NEAR_Business_centers', 'NEAR_metro_rjd',␣
       ↪'NEAR_Stations', 'K', 'T']
```

```python
[646]: params = {'loss_function':'MAE',
                 'eval_metric':'MAE',
                 'cat_features': categ_feat_idx,
                 'verbose': 200,
                 'random_seed': SEED,
                 'early_stopping_rounds': 100,
                 'boosting_type': 'Ordered', #Ordered
                 'bootstrap_type':'Bayesian',
                 'iterations': 1000,
                 'task_type':'GPU'
                }
       ctb_par_post = CatBoostRegressor(**params)
       ctb_par_post.fit(X_post_train_part, y_post_train_part,
                        eval_set=(X_post_valid, y_post_valid),
                            use_best_model=True,
                            plot=True
                            );
```

```
<IPython.core.display.HTML object>
```

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

```
0:      learn: 2.4254148          test: 2.3207204 best: 2.3207204 (0)      total:
22.2ms   remaining: 22.2s
200:     learn: 0.7741228          test: 0.8432397 best: 0.8432397 (200)   total:
4.94s    remaining: 19.6s
bestTest = 0.8432397207
bestIteration = 200
Shrink model to first 201 iterations.
```

[647]:
```python
ctb_par_post.get_feature_importance(prettified=True)
feature_importance_df = pd.DataFrame(ctb_par_post.
 ↪get_feature_importance(prettified=True), columns=['Feature Id',␣
 ↪'Importances'])
feature_importance_df
```

[647]:
```
            Feature Id  Importances
0                hh_500    38.219595
1            POPULATION    29.123362
2                values    16.966854
3  NEAR_Business_centers     4.356315
4          NEAR_metro_rjd     3.425293
5            work_hours     3.208375
6                     T     2.043189
7            NEAR_Malls     1.194684
8                     K     1.036324
9         NEAR_Stations     0.426010
```

[648]:
```python
plt.scatter(ctb_par_post.predict(X_post_valid), y_post_valid)
```

[648]: <matplotlib.collections.PathCollection at 0x19704fa65c0>
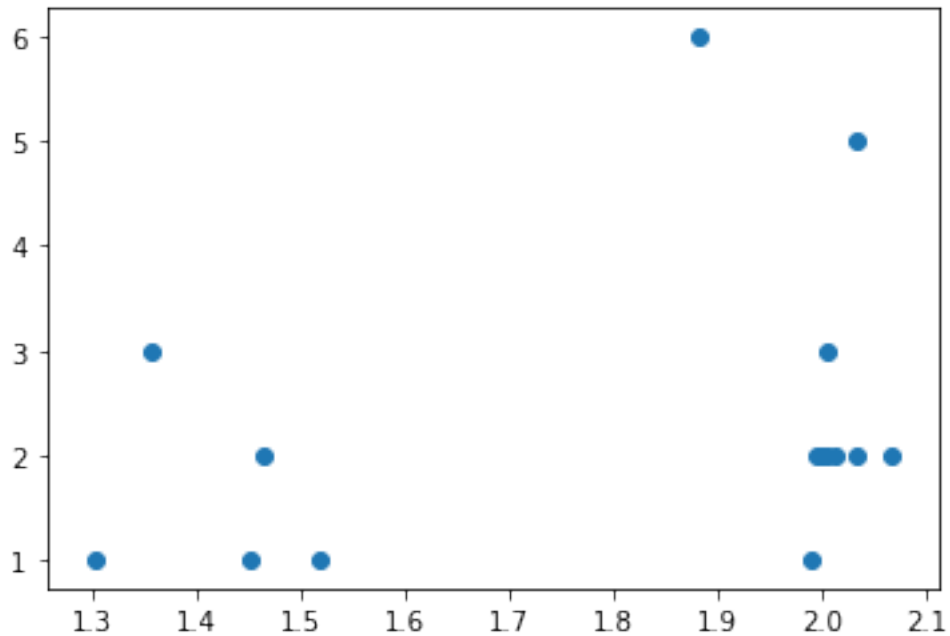
```
[649]: df_pr = pd.DataFrame({'Actual': y_post_valid, 'Predicted': ctb_par_post.
       ↪predict(X_post_valid)})
       df_pr.head()
```

```
[649]:      Actual  Predicted
       64      1.0   1.304006
       209     2.0   2.013751
       274     3.0   2.004724
       35      2.0   1.994939
       227     2.0   2.005648
```

```
[650]: df_pr.plot(kind='bar',figsize=(10,5))
       plt.grid(which='major', linestyle='-', linewidth='0.5', color='green')
       plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
       plt.show()
```
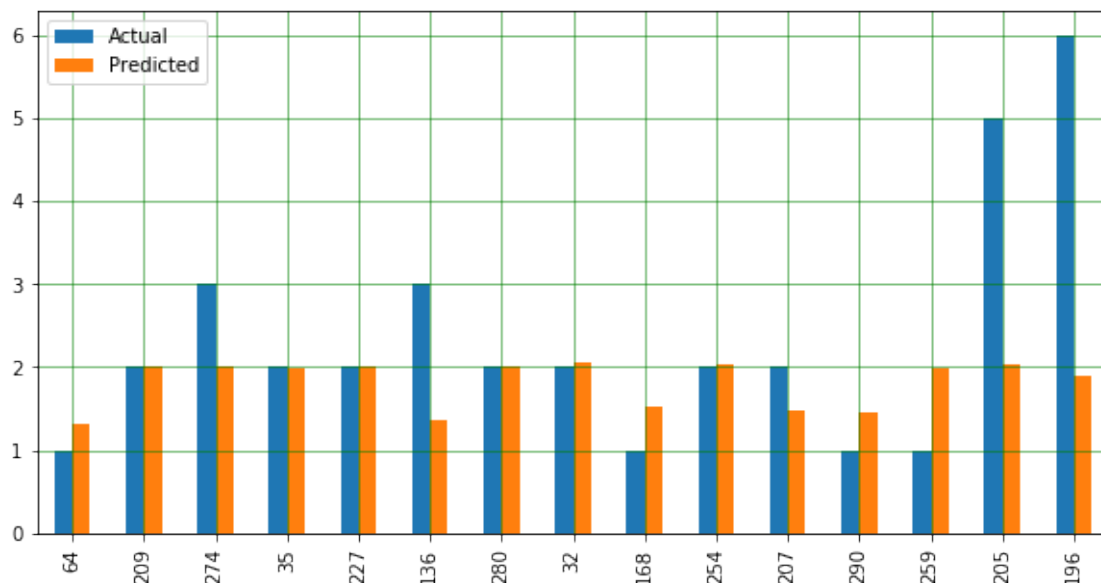
1.4 cashbox_daily

```
[651]: df_cashb_purpose = dff[dff.isna().cashbox_daily == True].copy()
       df_cashb_train = dff[dff.isna().cashbox_daily == False].copy()
       print('cashb purpose:', df_cashb_purpose.shape[0])
       print('cashb train:', df_cashb_train.shape[0])
```

```
cashb purpose: 264
cashb train: 153
```

1.5 cashbox_daily

```
[652]: df_cashb_train.head()
```

```
[652]:    Postamat_daily  cashbox_daily  Postamat_trend  cashbox_MAX  cashbox_trend  \
       0            NaN            3.3             NaN         13.0            1.0
       2            NaN            3.1             NaN         12.0            1.0
       3            3.0            2.2             0.0         12.0            1.0
       4            1.0            2.0             0.0          9.0            1.0
       5            NaN            1.2             NaN          6.0            1.0

          hh_500  POPULATION  NEAR_Malls  NEAR_Business_centers  NEAR_metro_rjd  \
       0  7696.0      324698           0                      0               0
       2  5553.0      324698           1                      0               0
       3  7162.0      324698           0                      0               0
       4  6269.0      324698           1                      0               0
       5  7703.0      324698           0                      0               0
```

```
     NEAR_Stations   macro_salary_avg_yearly   work_hours  K  T     values
0                1                    423321           24  1  0   2.986103
2                0                    423321           15  1  0   2.939464
3                0                    423321           15  1  0   3.100706
4                1                    423321           15  1  0   3.159937
5                1                    423321           19  1  0   3.110700
```

[653]: `df_cashb_train.isna().mean()`

[653]:
```
Postamat_daily              0.764706
cashbox_daily               0.000000
Postamat_trend              0.764706
cashbox_MAX                 0.000000
cashbox_trend               0.084967
hh_500                      0.000000
POPULATION                  0.000000
NEAR_Malls                  0.000000
NEAR_Business_centers       0.000000
NEAR_metro_rjd              0.000000
NEAR_Stations               0.000000
macro_salary_avg_yearly     0.000000
work_hours                  0.000000
K                           0.000000
T                           0.000000
values                      0.000000
dtype: float64
```

[654]:
```
df_cashb_train = df_cashb_train.
 ↪drop(columns=['cashbox_MAX','cashbox_trend','Postamat_daily','Postamat_trend','macro_salary
```

[655]:
```
df_cashb_purpose = df_cashb_purpose.
 ↪drop(columns=['cashbox_daily','cashbox_MAX','cashbox_trend','Postamat_daily','Postamat_tren
df_cashb_purpose['hh_500'].fillna((df_cashb_purpose['hh_500'].mean()),␣
 ↪inplace=True)
```

[656]: `df_cashb_train.shape`

[656]: `(153, 11)`

1.5.1 X_CASHB

[657]:
```
X_cashb = df_cashb_train.drop('cashbox_daily', axis=1)
y_cashb = df_cashb_train['cashbox_daily']
print(X_cashb.shape, y_cashb.shape)
```

```
(153, 10) (153,)
```

[658]: `X_train_part.head()`

```
[658]:     Postamat_trend  hh_500  POPULATION  NEAR_Malls  NEAR_Business_centers  \
     56                0    3170      104739           0                      0
     232               0    6772      482873           0                      0
     61                0    2164      324698           0                      0
     10                1    5232      324698           0                      0
     126               0     723       30392           0                      0

         NEAR_metro_rjd  NEAR_Stations  work_hours   K  T    values
     56                0              0          15   1  0  3.468459
     232               0              0          15   0  1  3.689425
     61                0              1          15   1  0  3.519116
     10                0              0          15   1  0  3.080232
     126               0              0          15   1  0  3.585912
```

1.6 CATBOOST

```
[659]: X_cashb.head()
```

```
[659]:    hh_500  POPULATION  NEAR_Malls  NEAR_Business_centers  NEAR_metro_rjd  \
     0  7696.0      324698           0                      0               0
     2  5553.0      324698           1                      0               0
     3  7162.0      324698           0                      0               0
     4  6269.0      324698           1                      0               0
     5  7703.0      324698           0                      0               0

        NEAR_Stations  work_hours   K  T    values
     0              1          24   1  0  2.986103
     2              0          15   1  0  2.939464
     3              0          15   1  0  3.100706
     4              1          15   1  0  3.159937
     5              1          19   1  0  3.110700
```

```
[660]: X_cashb.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 153 entries, 0 to 396
Data columns (total 10 columns):
hh_500                  153 non-null float64
POPULATION              153 non-null int64
NEAR_Malls              153 non-null int64
NEAR_Business_centers   153 non-null int64
NEAR_metro_rjd          153 non-null int64
NEAR_Stations           153 non-null int64
work_hours              153 non-null int64
K                       153 non-null uint8
T                       153 non-null uint8
values                  153 non-null float64
```

40

```
dtypes: float64(2), int64(6), uint8(2)
memory usage: 11.1 KB
```

[661]:
```python
X_cashb['NEAR_Malls'] = X_cashb['NEAR_Malls'].astype('str')
X_cashb['NEAR_Business_centers'] = X_cashb['NEAR_Business_centers'].
 ↪astype('str')
X_cashb['NEAR_metro_rjd'] = X_cashb['NEAR_metro_rjd'].astype('str')
X_cashb['NEAR_Stations'] = X_cashb['NEAR_Stations'].astype('str')
X_cashb['K'] = X_cashb['K'].astype('str')
X_cashb['T'] = X_cashb['T'].astype('str')


X_cashb['hh_500'] = X_cashb['hh_500'].astype('int64')
```

[662]:
```python
X_cashb.head()
```

[662]:
```
   hh_500  POPULATION NEAR_Malls NEAR_Business_centers NEAR_metro_rjd  \
0    7696      324698          0                     0              0
2    5553      324698          1                     0              0
3    7162      324698          0                     0              0
4    6269      324698          1                     0              0
5    7703      324698          0                     0              0

  NEAR_Stations  work_hours  K  T     values
0             1          24  1  0  2.986103
2             0          15  1  0  2.939464
3             0          15  1  0  3.100706
4             1          15  1  0  3.159937
5             1          19  1  0  3.110700
```

[663]:
```python
X_cashb_train_part, X_cashb_valid, y_cashb_train_part, y_cashb_valid =
 ↪train_test_split(X_cashb, y_cashb,
                                                          test_size=0.3,
                                                                    ␣
 ↪random_state=SEED)
print(X_cashb_train_part.shape[0], X_cashb_valid.shape[0], y_cashb_train_part.
 ↪shape[0], y_cashb_valid.shape[0])
```

```
107 46 107 46
```

[664]:
```python
X_cashb_train_part.head()
```

[664]:
```
     hh_500  POPULATION NEAR_Malls NEAR_Business_centers NEAR_metro_rjd  \
130    1590       12283          0                     0              0
210    2166      482873          1                     1              0
356    2368       16165          0                     0              0
136    1167       30392          0                     0              0
31     6143      104739          0                     0              0

    NEAR_Stations  work_hours  K  T     values
```

```
130              0      15  1  0  3.637251
210              1      15  0  1  3.285339
356              0      15  0  1  3.652747
136              0      15  1  0  3.583840
31               0      15  1  0  3.449743
```

[665]:
```python
categ_cashb_feat_idx = ['NEAR_Malls', 'NEAR_Business_centers',
 ↪'NEAR_metro_rjd', 'NEAR_Stations', 'K', 'T']
```

[666]:
```python
params = {'loss_function':'MAE',
          'eval_metric':'MAE',
          'cat_features': categ_cashb_feat_idx,
          'verbose': 200,
          'random_seed': SEED,
          'early_stopping_rounds': 100,
          'boosting_type': 'Ordered',
          'bootstrap_type':'Bayesian',
          'iterations': 1000,
          'task_type':'GPU',
          'learning_rate':0.01,
          'leaf_estimation_iterations':5,
          'max_ctr_complexity':4
         }
ctb_par_cashb = CatBoostRegressor(**params)
ctb_par_cashb.fit(X_cashb_train_part, y_cashb_train_part,
        eval_set=(X_cashb_valid, y_cashb_valid),
        use_best_model=True,
        plot=True
        );
```

<IPython.core.display.HTML object>

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


```
0:      learn: 1.3552795        test: 1.3430857 best: 1.3430857 (0)     total:
36.1ms   remaining: 36.1s
200:    learn: 0.4983826        test: 0.5987883 best: 0.5987883 (200)   total:
7.81s    remaining: 31s
400:    learn: 0.3796129        test: 0.5204967 best: 0.5204967 (400)   total:
16.3s    remaining: 24.4s
600:    learn: 0.3517222        test: 0.5052420 best: 0.5052420 (600)   total:
24.7s    remaining: 16.4s
800:    learn: 0.3312451        test: 0.4973848 best: 0.4973051 (796)   total:
33.5s    remaining: 8.31s
999:    learn: 0.3150220        test: 0.4945101 best: 0.4940847 (931)   total:
42.6s    remaining: 0us
bestTest = 0.4940847314
```

```
bestIteration = 931
Shrink model to first 932 iterations.
```

[667]:
```
ctb_par_cashb.get_feature_importance(prettified=True)
feature_importance_df = pd.DataFrame(ctb_par_cashb.
 ↪get_feature_importance(prettified=True), columns=['Feature Id',␣
 ↪'Importances'])
feature_importance_df
```

[667]:
```
              Feature Id  Importances
0                 hh_500    36.964862
1                 values    17.925999
2             POPULATION    15.836146
3                      T    10.634080
4                      K     6.568805
5             work_hours     5.837535
6             NEAR_Malls     2.942241
7  NEAR_Business_centers     2.016953
8          NEAR_Stations     1.046858
9         NEAR_metro_rjd     0.226522
```

[668]:
```
plt.scatter(ctb_par_cashb.predict(X_cashb_valid), y_cashb_valid)
```

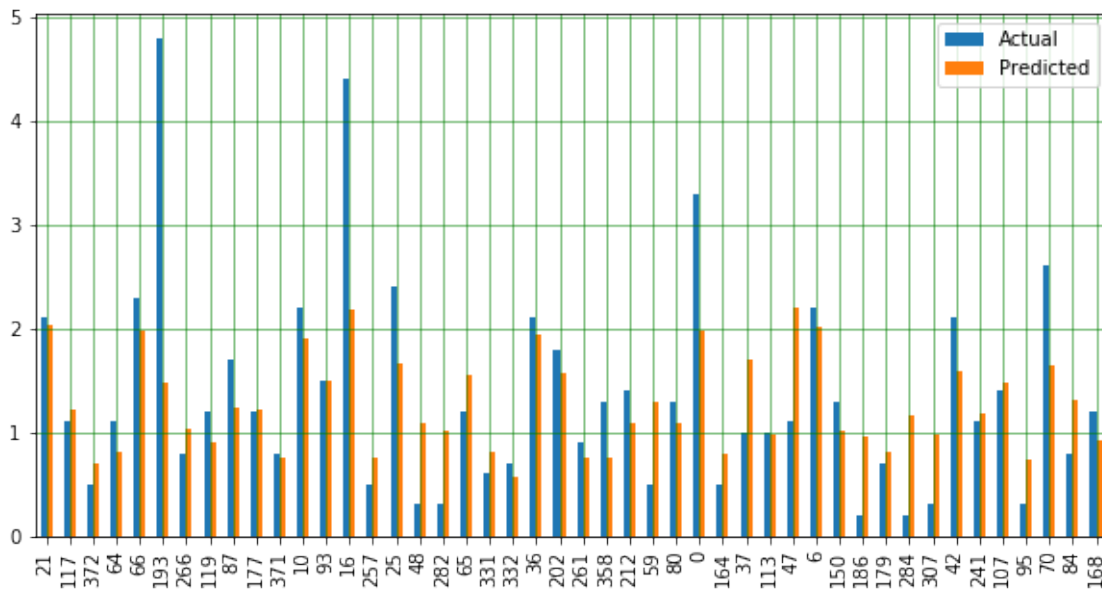[668]: <matplotlib.collections.PathCollection at 0x19704fc4390>
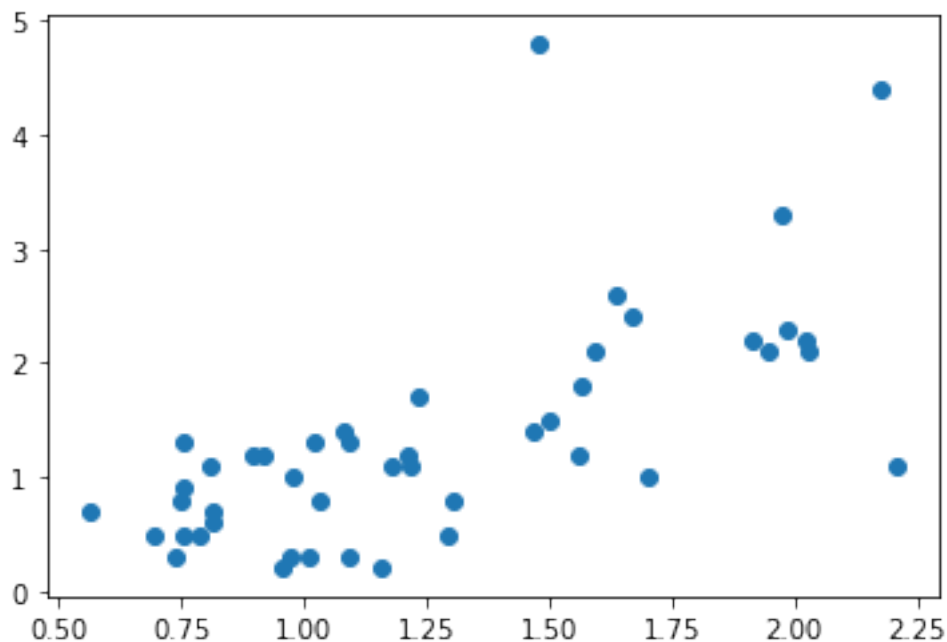
[669]:
```
df_pr = pd.DataFrame({'Actual': y_cashb_valid, 'Predicted': ctb_par_cashb.
 ↪predict(X_cashb_valid)})
df_pr.head()
```

```
[669]:        Actual   Predicted
        21      2.1    2.028805
        117     1.1    1.218512
        372     0.5    0.694230
        64      1.1    0.810241
        66      2.3    1.987257
```

```
[670]: df_pr.plot(kind='bar',figsize=(10,5))
       plt.grid(which='major', linestyle='-', linewidth='0.5', color='green')
       plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
       plt.show()
```



```
[671]: X_cashb_valid.head()
```

```
[671]:      hh_500  POPULATION NEAR_Malls NEAR_Business_centers NEAR_metro_rjd  \
       21     4219      324698          1                     0               0
       117    3044       40530          0                     0               0
       372    2457       42350          0                     0               0
       64     2127      324698          0                     0               1
       66     3870      104739          1                     0               0

            NEAR_Stations  work_hours  K  T   values
       21               0          24  1  0  3.283665
       117              0          15  1  0  3.633138
       372              0          15  0  1  3.819022
       64               1          15  1  0  3.707463
       66               1          15  1  0  3.568361
```

44

2

```
[672]: df_post_purpose.head()
```

```
[672]:    hh_500  POPULATION  NEAR_Malls  NEAR_Business_centers  NEAR_metro_rjd  \
       0  7696.0      324698           0                      0               0
       1  6369.0      324698           1                      1               0
       2  5553.0      324698           1                      0               0
       5  7703.0      324698           0                      0               0
       7  6487.0      324698           0                      0               0

          NEAR_Stations  work_hours  K  T   values
       0              1          24  1  0  2.986103
       1              1          24  1  0  3.111248
       2              0          15  1  0  2.939464
       5              1          19  1  0  3.110700
       7              1          24  1  0  3.160301
```

```
[673]: df_post_purpose['NEAR_Malls'] = df_post_purpose['NEAR_Malls'].astype('str')
       df_post_purpose['NEAR_Business_centers'] =␣
        ↪df_post_purpose['NEAR_Business_centers'].astype('str')
       df_post_purpose['NEAR_metro_rjd'] = df_post_purpose['NEAR_metro_rjd'].
        ↪astype('str')
       df_post_purpose['NEAR_Stations'] = df_post_purpose['NEAR_Stations'].
        ↪astype('str')
       df_post_purpose['K'] = df_post_purpose['K'].astype('str')
       df_post_purpose['T'] = df_post_purpose['T'].astype('str')

       df_post_purpose['hh_500'] = df_post_purpose['hh_500'].astype('int64')
```

```
[678]: post_predict_ver2 = ctb_par_post.predict(df_post_purpose)
```

```
[536]: df_cashb_purpose.head()
```

```
[536]:     hh_500  POPULATION  NEAR_Malls  NEAR_Business_centers  NEAR_metro_rjd  \
       1   6369.0      324698           1                      1               0
       11  7552.0      324698           0                      1               0
       12  3773.0      324698           0                      1               0
       13  4982.0      104739           0                      1               0
       14  7836.0      104739           0                      0               0

           NEAR_Stations  work_hours  K  T   values
       1               1          24  1  0  3.111248
       11              0          24  1  0  3.407810
       12              1          20  1  0  3.241490
       13              1          15  1  0  3.193099
       14              1          24  1  0  3.269593
```

```
[537]: df_cashb_purpose.isna().mean()
```

```
[537]: hh_500                   0.0
       POPULATION               0.0
       NEAR_Malls               0.0
       NEAR_Business_centers    0.0
       NEAR_metro_rjd           0.0
       NEAR_Stations            0.0
       work_hours               0.0
       K                        0.0
       T                        0.0
       values                   0.0
       dtype: float64
```

```python
[538]: df_cashb_purpose['NEAR_Malls'] = df_cashb_purpose['NEAR_Malls'].astype('str')
       df_cashb_purpose['NEAR_Business_centers'] =␣
        ↪df_cashb_purpose['NEAR_Business_centers'].astype('str')
       df_cashb_purpose['NEAR_metro_rjd'] = df_cashb_purpose['NEAR_metro_rjd'].
        ↪astype('str')
       df_cashb_purpose['NEAR_Stations'] = df_cashb_purpose['NEAR_Stations'].
        ↪astype('str')
       df_cashb_purpose['K'] = df_cashb_purpose['K'].astype('str')
       df_cashb_purpose['T'] = df_cashb_purpose['T'].astype('str')

       df_cashb_purpose['hh_500'] = df_cashb_purpose['hh_500'].astype('int64')
```

```python
[674]: cashb_predict_ver2 = ctb_par_cashb.predict(df_cashb_purpose)
```

```python
[675]: cashb_predict_ver2
```

```
[675]: array([1.87555734, 1.8638911 , 2.01993026, 2.050162  , 2.36734602,
              1.70796601, 0.90987514, 0.86861085, 1.60642016, 2.27033533,
              1.70320358, 0.95174631, 1.19776063, 0.96358283, 1.62987034,
              1.41880251, 0.99902978, 1.09371016, 0.79876455, 0.59517392,
              0.73461224, 0.94645162, 0.76733129, 1.07368198, 1.0987746 ,
              0.8297903 , 1.09023394, 1.03102583, 0.93971039, 1.02218908,
              1.15439281, 1.09823632, 1.2225367 , 0.88209353, 1.1001581 ,
              1.19856535, 1.08832683, 1.25192167, 1.08446357, 0.82226815,
              1.01566412, 1.02244065, 1.10540678, 1.12162358, 1.12162358,
              1.02120215, 1.04533377, 0.68444729, 0.82859305, 0.82817114,
              0.98693433, 0.98851126, 0.84626614, 0.98688131, 1.07110466,
              1.08088114, 1.03320694, 0.7734208 , 0.89086209, 0.92086537,
              1.01254966, 1.04356497, 0.74450021, 0.7377153 , 0.81837542,
              0.7357153 , 0.82800179, 1.02383252, 0.90615606, 0.95810563,
              0.74876166, 0.95213511, 0.90212736, 1.05821869, 0.76059655,
              0.96251424, 0.96354654, 1.02467788, 0.70868932, 0.92441529,
              0.84918835, 1.03160507, 0.70868932, 0.79801903, 0.81291009,
              0.91526449, 0.70868932, 0.76618597, 0.64277727, 0.76200029,
              0.64277727, 1.14843889, 1.13230527, 1.39977597, 1.28933613,
              1.29140488, 1.37247769, 1.49319415, 1.31669454, 1.29195597,
              1.34317798, 1.26675751, 1.26023208, 1.10581487, 1.22149067,
```

```
1.21008263, 1.20979296, 1.40755286, 1.28766228, 0.97892577,
0.98076532, 1.16683794, 1.08667551, 1.10472   , 1.09615437,
1.20732227, 0.47351416, 0.58644397, 1.15010172, 0.94429128,
1.11031337, 1.24790396, 1.19863767, 1.15232645, 0.97571294,
1.11363237, 1.22364168, 0.56881911, 0.70918697, 0.89637814,
1.11852814, 1.16868797, 1.11195023, 1.26339504, 0.48060162,
0.91000594, 1.11755335, 1.12307232, 1.04007855, 0.89721403,
0.81486899, 1.03933855, 1.11770621, 0.80158582, 0.93636683,
0.75217519, 1.16374265, 0.98889313, 1.22578924, 1.0873379 ,
1.16428356, 0.98763733, 0.79035297, 0.72626856, 0.70194207,
0.97527629, 0.83586996, 0.66416678, 0.70706153, 0.48855316,
1.11996186, 1.16601292, 0.849873  , 0.76758455, 0.78600587,
0.67136353, 0.93536575, 0.89944291, 0.87542687, 0.75591351,
0.82863403, 0.70361575, 0.60901298, 0.64540629, 0.38443788,
0.53952116, 0.89944291, 0.74507915, 0.80854608, 0.47569258,
0.98749841, 0.97086303, 0.68623948, 0.4629938 , 0.59987629,
0.59927761, 0.39070527, 0.50843309, 0.50312592, 0.60109425,
0.38879131, 0.72184977, 0.5769873 , 0.74358397, 0.88697971,
0.63080971, 0.50199691, 0.51468125, 0.54254371, 0.55364911,
0.64373636, 0.7501295 , 0.52523605, 0.36690323, 0.34991601,
0.73709078, 0.53476353, 0.4665099 , 0.50147053, 0.39084251,
0.69554095, 0.76653532, 0.62642288, 0.51281264, 0.62854931,
0.59554908, 0.46021638, 0.60475492, 0.60462753, 0.54065354,
0.52298386, 0.39433321, 0.40128667, 0.58179149, 0.38629128,
0.46812473, 0.43422285, 0.42883396, 0.51468149, 0.53998605,
0.5510391 , 0.71782117, 0.56378986, 0.38538024, 0.43125313,
0.62766467, 0.58388482, 0.37226168, 0.36995059, 0.42570628,
0.44105489, 0.36932559, 0.39941813, 0.45428136, 0.38137234,
0.43682726, 0.41117875, 0.43125313, 0.47285683, 0.60197164,
0.40828014, 0.38316449, 0.39372831, 0.49000653, 0.46050612,
0.48088294, 0.40287556, 0.53343801, 0.40287556, 0.46829484,
0.42091753, 0.53190571, 0.38253949, 0.32965857])
```