combined_features

December 11, 2019

https://github.com/QuantCS109/TrumpTweets/blob/master/notebooks_features/combined_features.ipynb

[1]:
```
import sys
sys.path.append('..') #to add top-level to path

import pickle
import pandas as pd
import numpy as np
from datetime import timedelta
from sklearn.model_selection import train_test_split
from modules.project_helper import FuturesCloseData, VolFeatures,␣
↪TweetReturnsFeatures, TradeModel, MarketFeatures
import matplotlib.pyplot as plt
#import graphviz
```

0.0.1 Response Variable (One Day Log-Returns)

[2]:
```
fc = FuturesCloseData()
fc.single_log_returns('ES').head()
```

[2]:
```
date
2014-01-02   -0.000570
2014-01-03   -0.002710
2014-01-06    0.005697
2014-01-07    0.000994
2014-01-08    0.000284
Name: ES, dtype: float64
```

0.0.2 1) Tweet topics

[3]:
```
topics = pd.read_csv('../data/features/topic_features_clusters=25.csv').
↪set_index('date')
topics.columns = ["topic_"+column for column in topics.columns.tolist()]
```

[4]:
```
topics.head()
```

[4]:
```
            topic_0   topic_1  topic_2   topic_3   topic_4   topic_5  topic_6  \
date
2017-01-01      0.0  0.000000      0.0  0.000000  0.000000  0.000000      0.0
2017-01-02      0.0  1.000000      0.0  0.000000  0.000000  0.000000      0.0
2017-01-03      0.0  0.222222      0.0  0.111111  0.222222  0.000000      0.0
2017-01-04      0.0  0.214286      0.0  0.214286  0.142857  0.000000      0.0
2017-01-05      0.0  0.000000      0.0  0.000000  0.333333  0.166667      0.0

            topic_7   topic_8  topic_9  …  topic_15  topic_16  topic_17  \
date                                    …
2017-01-01      0.0  0.000000      0.0  …  0.000000  0.000000       0.5
2017-01-02      0.0  0.000000      0.0  …  0.000000  0.000000       0.0
2017-01-03      0.0  0.111111      0.0  …  0.000000  0.111111       0.0
2017-01-04      0.0  0.000000      0.0  …  0.071429  0.071429       0.0
2017-01-05      0.0  0.000000      0.0  …  0.166667  0.166667       0.0

            topic_18  topic_19  topic_20  topic_21  topic_22  topic_23  \
date
2017-01-01       0.0       0.0       0.0       0.0       0.0       0.0
2017-01-02       0.0       0.0       0.0       0.0       0.0       0.0
2017-01-03       0.0       0.0       0.0       0.0       0.0       0.0
2017-01-04       0.0       0.0       0.0       0.0       0.0       0.0
2017-01-05       0.0       0.0       0.0       0.0       0.0       0.0

            topic_24
date
2017-01-01       0.0
2017-01-02       0.0
2017-01-03       0.0
2017-01-04       0.0
2017-01-05       0.0

[5 rows x 25 columns]
```

0.0.3 2) TF-IDF (First Two Components) Features

[5]:
```
svd_df_daily = pd.read_csv('../data/features/combined_svd_df.
 ↪csv',names=['index','svd_1','svd_2','date'], index_col = 0, skiprows = 1)
svd_df_daily.set_index('date', inplace = True)
svd_df_daily.index = pd.to_datetime(svd_df_daily.index)
svd_df_daily.head()
```

[5]:
```
              svd_1     svd_2
date
2009-05-05  0.229959  0.195915
2009-05-08  0.052085  0.062540
2009-05-09  0.079564  0.035554
```

```
2009-05-12  0.101352  0.043649
2009-05-13  0.068212  0.062037
```

0.0.4 3) Trump Tweet Returns Features

[6]:
```
tr = TweetReturnsFeatures()
tr.features('ES').head()
```

[6]:
```
            ES_min_tweet  ES_max_tweet  ES_daily_tweet
date
2017-02-01      0.001327      0.003084        0.000107
2017-02-02      0.000208      0.004080        0.000087
2017-02-03      0.001417      0.005195        0.000137
2017-02-06      0.001328      0.003935        0.000134
2017-02-07      0.001821      0.004572        0.000142
```

[7]:
```
tr.features('ES').dtypes
```

[7]:
```
ES_min_tweet      float64
ES_max_tweet      float64
ES_daily_tweet    float64
dtype: object
```

0.0.5 4) Futures Market Features

[8]:
```
market = pd.read_csv('../data/features/market_features.csv')
market = MarketFeatures()
market.features('ES').head()
```

[8]:
```
            ES_volume_chg  ES_opening_down  ES_opening_unch  ES_opening_up
date
2015-11-17      -2.869230                0                0              1
2015-11-18      -2.553725                0                0              1
2015-11-19      -5.156960                1                0              0
2015-11-20      13.056758                1                0              0
2015-11-23      -8.501065                0                0              1
```

0.0.6 5) S&P500 Intraday Features

[9]:
```
intraday = pd.read_csv('../data/features/intra_sp_features.csv')
intraday.set_index('date', inplace = True)
intraday.index = pd.to_datetime(intraday.index)
intraday.head()
```

[9]:
```
            intra_ret_1  intra_ret_5  intra_ret_15  intra_diff_15_5  \
date
```

```
2016-11-14    -0.000343    -0.000343     -0.000343        -0.000343
2016-11-15    -0.000114    -0.000571     -0.001141         0.000114
2016-11-16    -0.000457    -0.000228     -0.000800        -0.000228
2016-11-17     0.000908     0.000908      0.000908         0.000908
2016-11-18    -0.000114    -0.000114     -0.000114        -0.000114

            intra_blend
date
2016-11-14    -0.000343
2016-11-15    -0.000608
2016-11-16    -0.000495
2016-11-17     0.000908
2016-11-18    -0.000114
```

0.0.7 6) Volatility Curve Features

[10]:
```
vol = VolFeatures()
vol.features('ES').head()
```

[10]:
```
            ES_1M_atm_vol  ES_1M_RR25  ES_1M_Fly25  ES_2M_RR25  ES_2M_Fly25  \
Date
2016-11-14       0.125167   -0.046660     0.004511   -0.054350     0.005055
2016-11-15       0.115973   -0.037389     0.004496   -0.047791     0.005024
2016-11-16       0.117994   -0.039737     0.004298   -0.050144     0.004259
2016-11-17       0.114572   -0.037683     0.003737   -0.047440     0.004492
2016-11-18       0.110470   -0.037870     0.002963   -0.049350     0.004098

            ES_2M_1M_atm_vol
Date
2016-11-14          0.005187
2016-11-15          0.004549
2016-11-16          0.005215
2016-11-17          0.005008
2016-11-18          0.006858
```

0.0.8 7) Agricultural Gamma Features

[11]:
```
corn_gamma = pd.read_csv('../data/features/corn_gamma_features.csv')
corn_gamma.set_index('date', inplace = True)
corn_gamma.index = pd.to_datetime(corn_gamma.index)
corn_gamma.head()
```

[11]:
```
            C_up_gamma_5  C_down_gamma_5   C_up_diff_5  C_down_diff_5
date
2016-09-21  53219.375632    53384.321274  11268.575199   11471.420535
2016-09-22  52737.562141    54080.579964  11594.644211   11644.373967
```

```
2016-09-27  56031.645597    37727.193425  12965.800732    7437.719089
2016-09-29  56912.627474    39498.507310  13003.249796    8306.498080
2016-09-30  56946.454785    40410.808361  12994.168946    8664.380661
```

[12]:
```
wheat_gamma = pd.read_csv('../data/features/wheat_gamma_features.csv')
wheat_gamma.set_index('date', inplace = True)
wheat_gamma.index = pd.to_datetime(wheat_gamma.index)
wheat_gamma.head()
```

[12]:
```
            W_up_gamma_5  W_down_gamma_5  W_up_diff_5  W_down_diff_5
date
2016-09-27  11281.695503     5119.843342  2705.436002    1032.187814
2016-09-29  11350.174510     6931.755883  2601.233781    1962.597329
2016-10-04  11245.962151     6619.575422  2693.882756    1672.955456
2016-10-11  14704.655914     9338.077715  2955.022047    2552.378465
2016-10-13  18393.989866     8111.147772  3938.700607    2212.199049
```

[13]:
```
soybeans_gamma = pd.read_csv('../data/features/soybeans_gamma_features.csv')
soybeans_gamma.set_index('date', inplace = True)
soybeans_gamma.index = pd.to_datetime(soybeans_gamma.index)
soybeans_gamma.head()
```

[13]:
```
            S_up_gamma_5  S_down_gamma_5   S_up_diff_5  S_down_diff_5
date
2016-09-21  34465.710071    39120.529046  13183.794488   16165.099994
2016-09-22  37548.504509    38822.637529  13185.095326   15769.076624
2016-09-23  29833.167794    39086.796216  13758.593327   16592.201574
2016-09-26  39720.842812    33082.313487  16241.849329   13876.210431
2016-09-27  38721.781147    34020.341945  16677.953367   12458.728470
```

0.0.9 7) Agricultural Gamma Features

[14]:
```
sentiment_features = pd.read_csv("../data/features/daily_sentiment.csv")
sentiment_features = sentiment_features.set_index('date')
sentiment_features.head()
```

[14]:
```
            negative_proportion_min  negative_proportion_max  \
date
2009-05-05                    0.000                    0.000
2009-05-08                    0.000                    0.000
2009-05-09                    0.000                    0.000
2009-05-12                    0.000                    0.000
2009-05-13                    0.075                    0.075

            negative_proportion_mean  positive_proportion_min  \
date
2009-05-05                     0.000                    0.163
```

```
2009-05-08                    0.000                    0.277
2009-05-09                    0.000                    0.000
2009-05-12                    0.000                    0.000
2009-05-13                    0.075                    0.222

            positive_proportion_max  positive_proportion_mean  \
date
2009-05-05                    0.252                    0.2075
2009-05-08                    0.277                    0.2770
2009-05-09                    0.000                    0.0000
2009-05-12                    0.000                    0.0000
2009-05-13                    0.222                    0.2220

            neutral_proportion_min  neutral_proportion_max  \
date
2009-05-05                   0.748                   0.837
2009-05-08                   0.723                   0.723
2009-05-09                   1.000                   1.000
2009-05-12                   1.000                   1.000
2009-05-13                   0.703                   0.703

            neutral_proportion_mean  combined_score_min  combined_score_max  \
date
2009-05-05                   0.7925              0.4767              0.7506
2009-05-08                   0.7230              0.6115              0.6115
2009-05-09                   1.0000              0.0000              0.0000
2009-05-12                   1.0000              0.0000              0.0000
2009-05-13                   0.7030              0.4809              0.4809

            combined_score_mean
date
2009-05-05              0.61365
2009-05-08              0.61150
2009-05-09              0.00000
2009-05-12              0.00000
2009-05-13              0.48090
```

1 Combine All Features

[15]:
```
instrument_list = ['ES', 'NQ', 'CD', 'EC', 'JY', 'MP', 'TY', 'US', 'C', 'S',␣
 ↪'W', 'CL', 'GC']
start_date = pd.to_datetime('2017-02-01')
end_date = pd.to_datetime('2019-11-07')
full_features ={}

for inst in instrument_list:
```

```
    date_filter = fc.single_log_returns(inst,1,2).index[(fc.
 ↪single_log_returns(inst,1,2).index >= start_date) &
                                                        (fc.
 ↪single_log_returns(inst,1,2).index <= end_date)]

    full_features[inst] = pd.DataFrame(fc.single_log_returns(inst,1,2).
 ↪loc[date_filter])\
                            .join(tr.features(inst))

    full_features[inst] = full_features[inst].fillna(full_features[inst].mean())
    full_features[inst] = full_features[inst]\
                            .join(vol.features(inst))\
                            .join(topics)\
                            .join(svd_df_daily)\
                            .fillna(0)\
                            .join(market.features(inst))\
                            .join(intraday)\
                            .fillna(0)\
                            .join(sentiment_features)\
                            .fillna(0)

    if inst=='C':
         full_features[inst] = full_features[inst]\
                                .join(corn_gamma)\
                                .fillna(method='ffill')
    if inst=='W':
             full_features[inst] = full_features[inst]\
                                    .join(wheat_gamma)\
                                    .fillna(method='ffill')
    if inst=='S':
         full_features[inst] = full_features[inst]\
                                .join(soybeans_gamma)\
                                .fillna(method='ffill')




#vol.features(inst).loc[features_index].join(fc.returns(inst).
 ↪loc[features_index])import pickle
filehandler = open("../data/features/full_features.pkl","wb")
pickle.dump(full_features,filehandler)
filehandler.close()
```