feature_engineering_ablation_study

May 10, 2024

1 Feature Engineering Ablation Study

[ ]:
```
import pandas as pd
import sys

sys.path.insert(1, "/Users/simon/Documents/II/Dissertation/")
from src.evaluate import get_results_df

%load_ext autoreload
%autoreload 2
```

```
The autoreload extension is already loaded. To reload it, use:
  %reload_ext autoreload
```

[ ]:
```
models = ["Linear", "ARIMA", "RandomForest", "CNN", "LSTM", "ConvLSTM"]
stocks = ["NVDA", "JPM", "HD", "UNH"]

yes_df = []
no_df = []
for m in models:
    for s in stocks:
        yes_df.append(get_results_df(f"{m}_{s}_Yes-Feature-Engineering"))
        no_df.append(get_results_df(f"{m}_{s}_No-Feature-Engineering"))

yes_df = pd.concat(yes_df)
no_df = pd.concat(no_df)
```

```
Loading Linear_NVDA_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 45.0199203187251. Run completed at 2024-04-29
14:16:50.763116
Loading Linear_NVDA_No-Feature-Engineering.
Rank 1: trial no. 0, value: 48.20717131474104. Run completed at 2024-04-29
14:16:49.600029
Loading Linear_JPM_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 46.613545816733065. Run completed at 2024-04-29
14:16:53.084257
Loading Linear_JPM_No-Feature-Engineering.
Rank 1: trial no. 0, value: 46.613545816733065. Run completed at 2024-04-29
```

14:16:51.934299
Loading Linear_HD_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 52.589641434262944. Run completed at 2024-04-29
14:16:55.412646
Loading Linear_HD_No-Feature-Engineering.
Rank 1: trial no. 0, value: 49.40239043824701. Run completed at 2024-04-29
14:16:54.284397
Loading Linear_UNH_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 47.808764940239044. Run completed at 2024-04-29
14:16:57.722188
Loading Linear_UNH_No-Feature-Engineering.
Rank 1: trial no. 0, value: 55.77689243027888. Run completed at 2024-04-29
14:16:56.589737
Loading ARIMA_NVDA_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 54.980079681274894. Run completed at 2024-04-29
14:17:12.819213
Loading ARIMA_NVDA_No-Feature-Engineering.
Rank 1: trial no. 0, value: 46.21513944223107. Run completed at 2024-04-29
14:17:00.772048
Loading ARIMA_JPM_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 51.39442231075697. Run completed at 2024-04-29
14:17:28.279760
Loading ARIMA_JPM_No-Feature-Engineering.
Rank 1: trial no. 0, value: 46.613545816733065. Run completed at 2024-04-29
14:17:15.813764
Loading ARIMA_HD_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 52.589641434262944. Run completed at 2024-04-29
14:17:42.183274
Loading ARIMA_HD_No-Feature-Engineering.
Rank 1: trial no. 0, value: 48.60557768924303. Run completed at 2024-04-29
14:17:31.028398
Loading ARIMA_UNH_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 46.613545816733065. Run completed at 2024-04-29
14:17:56.707862
Loading ARIMA_UNH_No-Feature-Engineering.
Rank 1: trial no. 0, value: 50.199203187250994. Run completed at 2024-04-29
14:17:45.339249
Loading RandomForest_NVDA_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 47.410358565737056. Run completed at 2024-04-29
14:19:23.968185
Loading RandomForest_NVDA_No-Feature-Engineering.
Rank 1: trial no. 0, value: 50.59760956175299. Run completed at 2024-04-29
14:18:04.329617
Loading RandomForest_JPM_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 51.39442231075697. Run completed at 2024-04-29
14:21:16.050882
Loading RandomForest_JPM_No-Feature-Engineering.
Rank 1: trial no. 0, value: 47.808764940239044. Run completed at 2024-04-29

14:19:32.183616
Loading RandomForest_HD_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 52.98804780876494. Run completed at 2024-04-29
14:22:59.787532
Loading RandomForest_HD_No-Feature-Engineering.
Rank 1: trial no. 0, value: 43.02788844621514. Run completed at 2024-04-29
14:21:24.810764
Loading RandomForest_UNH_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 51.79282868525896. Run completed at 2024-04-29
14:24:31.640818
Loading RandomForest_UNH_No-Feature-Engineering.
Rank 1: trial no. 0, value: 50.199203187250994. Run completed at 2024-04-29
14:23:07.511479
Loading CNN_NVDA_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 0.5378485918045044. Run completed at 2024-04-29
14:26:28.496674
Loading CNN_NVDA_No-Feature-Engineering.
Rank 1: trial no. 0, value: 0.5179283022880554. Run completed at 2024-04-29
14:25:57.501574
Loading CNN_JPM_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 0.5458167195320129. Run completed at 2024-04-29
14:27:40.619409
Loading CNN_JPM_No-Feature-Engineering.
Rank 1: trial no. 0, value: 0.5458167195320129. Run completed at 2024-04-29
14:27:10.325449
Loading CNN_HD_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 0.5498008131980896. Run completed at 2024-04-29
14:28:43.077136
Loading CNN_HD_No-Feature-Engineering.
Rank 1: trial no. 0, value: 0.5737051963806152. Run completed at 2024-04-29
14:28:09.833524
Loading CNN_UNH_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 0.5577689409255981. Run completed at 2024-04-29
14:29:47.895458
Loading CNN_UNH_No-Feature-Engineering.
Rank 1: trial no. 0, value: 0.5657370686531067. Run completed at 2024-04-29
14:29:24.932915
Loading LSTM_NVDA_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 0.5537848472595215. Run completed at 2024-04-29
14:34:44.074320
Loading LSTM_NVDA_No-Feature-Engineering.
Rank 1: trial no. 0, value: 0.5458167195320129. Run completed at 2024-04-29
14:34:27.434376
Loading LSTM_JPM_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 0.525896430015564. Run completed at 2024-04-29
14:35:35.519313
Loading LSTM_JPM_No-Feature-Engineering.
Rank 1: trial no. 0, value: 0.5657370686531067. Run completed at 2024-04-29

```
14:35:04.394477
Loading LSTM_HD_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 0.5378485918045044. Run completed at 2024-04-29
14:36:28.609972
Loading LSTM_HD_No-Feature-Engineering.
Rank 1: trial no. 0, value: 0.5219123363494873. Run completed at 2024-04-29
14:35:57.649011
Loading LSTM_UNH_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 0.5577689409255981. Run completed at 2024-04-29
14:25:30.109564
Loading LSTM_UNH_No-Feature-Engineering.
Rank 1: trial no. 0, value: 0.5458167195320129. Run completed at 2024-04-29
14:25:00.538455
Loading ConvLSTM_NVDA_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 0.5179283022880554. Run completed at 2024-04-29
14:30:37.460949
Loading ConvLSTM_NVDA_No-Feature-Engineering.
Rank 1: trial no. 0, value: 0.5179283022880554. Run completed at 2024-04-29
14:30:05.073017
Loading ConvLSTM_JPM_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 0.525896430015564. Run completed at 2024-04-29
14:31:21.965994
Loading ConvLSTM_JPM_No-Feature-Engineering.
Rank 1: trial no. 0, value: 0.525896430015564. Run completed at 2024-04-29
14:30:58.192919
Loading ConvLSTM_HD_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 0.5139442086219788. Run completed at 2024-04-29
14:32:02.569686
Loading ConvLSTM_HD_No-Feature-Engineering.
Rank 1: trial no. 0, value: 0.5139442086219788. Run completed at 2024-04-29
14:31:40.298160
Loading ConvLSTM_UNH_Yes-Feature-Engineering.
Rank 1: trial no. 0, value: 0.5577689409255981. Run completed at 2024-04-29
14:32:56.464764
Loading ConvLSTM_UNH_No-Feature-Engineering.
Rank 1: trial no. 0, value: 0.5577689409255981. Run completed at 2024-04-29
14:32:28.216694

/var/folders/d7/ktx3dym91yjgj_gpmnfs0rh00000gn/T/ipykernel_49809/548548786.py:11
: FutureWarning: The behavior of DataFrame concatenation with empty or all-NA
entries is deprecated. In a future version, this will no longer exclude empty or
all-NA columns when determining the result dtypes. To retain the old behavior,
exclude the relevant entries before the concat operation.
  yes_df = pd.concat(yes_df)
/var/folders/d7/ktx3dym91yjgj_gpmnfs0rh00000gn/T/ipykernel_49809/548548786.py:12
: FutureWarning: The behavior of DataFrame concatenation with empty or all-NA
entries is deprecated. In a future version, this will no longer exclude empty or
all-NA columns when determining the result dtypes. To retain the old behavior,
```

```
exclude the relevant entries before the concat operation.
  no = pd.concat(no_df)
```

[ ]:
```
# Aggregating by stock
df = yes_df.copy()
df = (df["Validation set"]) / 2
df["Stock"] = df.index.str.split("_").str[1]
df = df.groupby("Stock").mean()
after = df

df = no_df.copy()
df = (df["Validation set"]) / 2
df["Stock"] = df.index.str.split("_").str[1]
df = df.groupby("Stock").mean()
before = df

(after - before).mean()
```

[ ]:
```
R2                  -20.21329884
MSE                   0.00940316
RMSE                  0.02153411
MAE                   0.02188086
p                     0.00893545
Accuracy              0.03320053
Avg. daily return     0.00000542
Std. daily return     0.00003772
Risk adj. return     -0.00014600
dtype: float64
```