# Machine Learning for financial time series

Liqiu MA

Data Analytics, Capital Fund Management, Paris

May 14, 2020

# Outline

# Introduction

# Introduction

**Goal**

Find the linear dynamics of **daily** liquidity metrics in the **future** market to provide a benchmark for further analyses.

- The liquidity metrics we consider are daily bid-ask spread, volume and volatility .
- Find a linear model using past values of liquidity metrics and some exogenous(Open interest, return) and deterministic parameters(day of week,etc) as regressors, we wish the model works for all future contract.
- In this presentation, we focus on linear model as benchmark. We will go to non-linear model(deep neural network) in the future.

# Notation

- $F_t^i(T_j)$ represents price of a future contract $j$ over underlying $i$ at time t which has expiry $T_j$, and suppose that we have different contracts under a same underlying $T_0 < T_1 < T_2 < ...$ .
- $F_t^i(\delta) := F_t^i(T)$ where $\delta$ represents time to expiry $T - t$.
- The same notation for these variables: bid-ask spread $S_t^i(T)$, volume $V_t^i(T)$, volatility $\sigma_t^i(T)$, open interest $O_t^i(T)$ and return $R_t^i(T)$.

# Database

## Data Description

- We have 4276 future contracts over 78 different underlyings.
- The underlyings contain different asset class, such as commodity and equity index, etc.
- Our dataset covers period of January 2011 to March 2020.

| | |
|---|---|
| total numbers of underlying ($i$) | 78 |
| total numbers of future contract ($i, T_j$) | 4276 |
| total numbers of prodgen | 302 |
| avg numbers of observation per underlying | 7654.69 |
| avg numbers of observation per future contract | 139.63 |
| avg numbers of observation per prodgen | 1951.20 |

**Table 1:** Dataset numbers

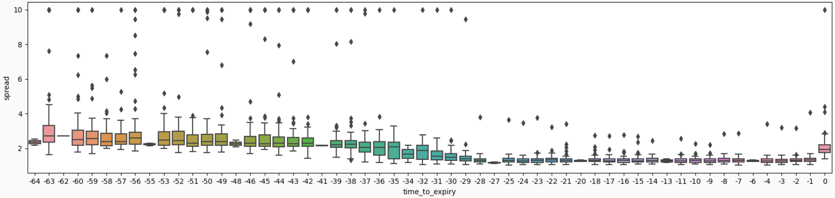
# Data Preprocessing

- We have daily open,high,low,close, open interest, volume directly in the database.
- The return we consider is return of close. The spread is average spread in tick during a day.
- The volatility is estimated by Garman Klass estimator. We have encountered several problems, the negative price, and the open, close is not between low and high.

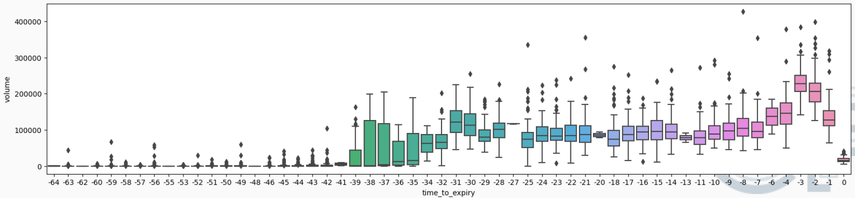
## Data Processing(Stationarization)

We show that several variables have a seasonality w.r.t time to expiry for example, spread, volume and open interest.

- Seasonality of spread for CAC 40 future contract.




- Seasonality of volume for CAC 40 future contract.



## Data Processing(Stationarization)

- We need remove seasonalities to focus on dynamics around this average behaviour. For variables spread, volume and open interest, we do

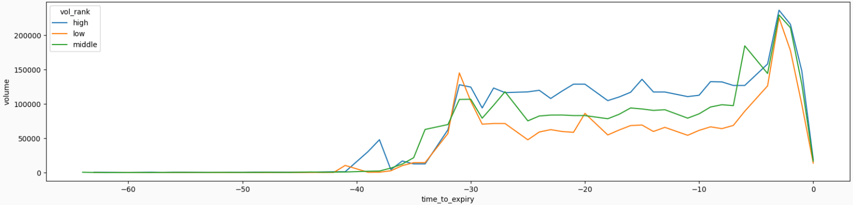
$$\bar{L}^i(\delta_j) =< L^i_\cdot(\delta_k) >_{\delta_k=\delta_j}$$

$$\hat{L}^i_t(\delta_j) = \frac{L^i_t(\delta_j)}{\bar{L}^i(\delta_j)}$$

- The avg point in denominator is 33.58 .

# Data Processing(Stationarization)

Seasonalities we considered are unconditionnal,i.e., they don't depend on other variables. But what if the assumption fails? We consider $< L_{\cdot}^{i}(\delta_k)|\sigma_{\cdot}^{i}(\delta_k) \in I >_{\delta_k=\delta_j}$

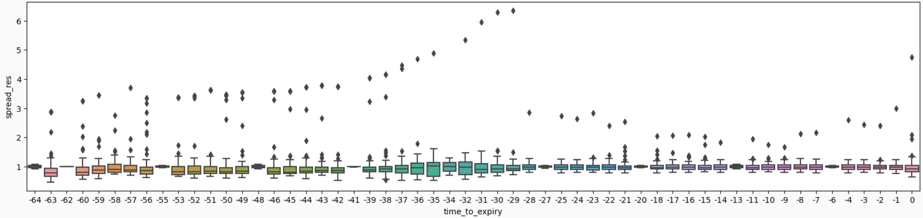



We can model this phenomenon in the future.

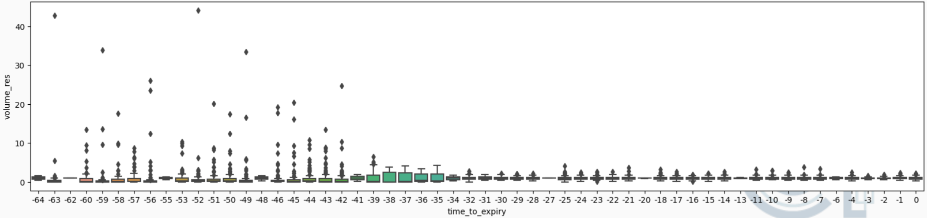
# Data Processing(Stationarization)

The figure below shows the effects of the deseasonality for CAC 40.

- deseasonalized spread for CAC 40 future contract




- deseasonalized volume for CAC 40 future contract



# Decomposition in long term and short term

- In order to find a universal model, we'd better find a characteristic of a future contract to represent itself, so we decompose it into long term and short term, and the long term represent the future contract and short term is the random part we want to explicate.
- Firstly, we roll future contract by liquidity rank(maturity) r to get the prodgen.

$$\hat{L}^i_t(r) = \begin{cases} \hat{L}^i_t(T_{r+1}) & T_0 < t \leq T_1 \\ \hat{L}^i_t(T_{r+2}) & T_1 < t \leq T_2 \\ \dots \end{cases}$$

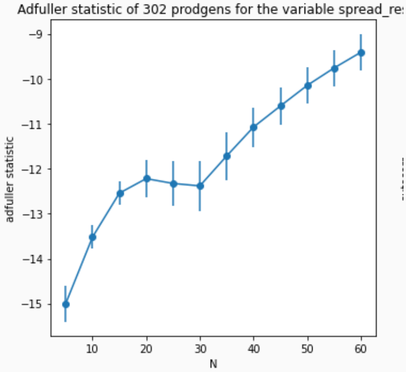
- Secondly, decompose prodgen in short term and long term using moving averge

$$LT(\hat{L}^i_t(r)) = \frac{1}{N} \sum_{k=0}^{N} \hat{L}^i_{t-k}(r)$$

$$ST(\hat{L}^i_t(r)) = \hat{L}^i_t(r) / LT(\hat{L}^i_t(r))$$

# Choice of N in MA

- The key is to chose N to make $ST(\hat{L}_t^i(r))$ more stationnary possible. The N is different for variables.
- Here we have 302 prodgens in total. for a fix variable , we calculate augmented ADF statistic for these 302 prodgens for different Ns, an example for spread is shown below.



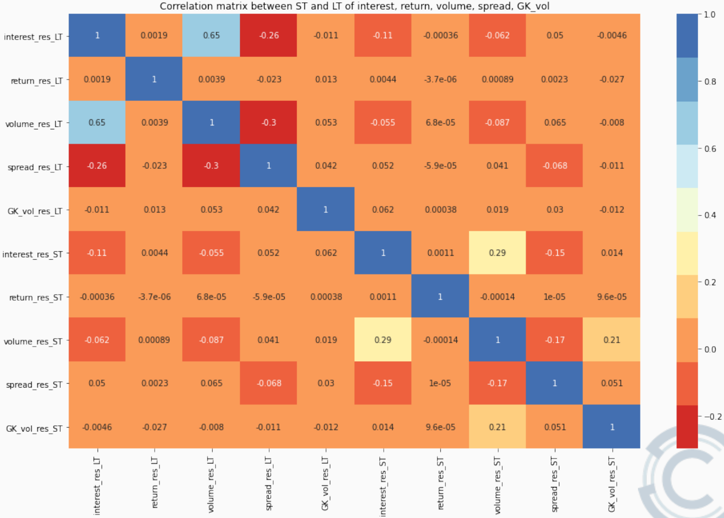

Choice of N for different variables

- spread: 30
- volume: 25
- volatility: 25
- open interest: 20
- return : 15

# Methodology

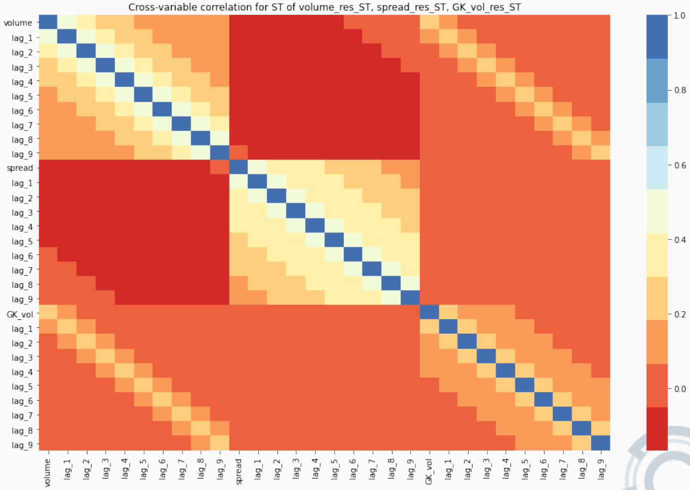
# Correlation analysis

- Before linear model, we will look at correlation matrix between different variables in long term and short term over all prodgens.



# Correlation analysis

- Our focus is to predict short term of liquidity metrics, so we plot correlations between short term of the variables observed at the same time or with a delay.



# Model description

We use a generalized VAR model which is

**VAR Model**

Let $Y_t = (y_{1t}, y_{2t}, \ldots, y_{nt})'$ denote an $(n \times 1)$ vector of time series variables, $D_t$ represents an $(l \times 1)$ matrix of deterministic components, $X_t$ represents an $(m \times 1)$ matrix of exogenous variables, The mean-ajusted $p$-lag vector autoregressive VAR $(p)$ model

$$Y_{t+1} = \sum_{k=0}^{p} \Pi_k Y_{t-k} + \Phi D_{t+1} + GX_t + \varepsilon.$$

where $\Pi_i$, $\Phi$ and $G$ are parameter matrices and $\varepsilon_t$ is an $(n \times 1)$ unobservable zero mean white noise vector process with time invariant covariance matrix $\Sigma$.

In reality, we use

- $Y$ represents volume,spread,vol in short term (n=3).
- $X$ represents volume,spread,vol in long term and open interest,return in long and short term (m=7).
- $D$ represents time to expiry, day of week, week of month, week of year, is vacation, liquidity rank (l=6).

# Estimation and Metrics

- First we divide the dataset(all prodgens) into different intervals of 300 days , we put first 150 consecutive trading days into in sample data and last 150 consecutive trading days as out of sample data.
- We use standard maximum-likelihood method to estimate parameters in the VAR model using in sample data.(During implementation, we used OLS for three variables in short term. We didn't use VAR alogrithm in statsmodel because it can't predict what we expect. )
- To choose lag(model order), we use the lag which minimizes Akaike Information Criteria(AIC)

$$
\begin{aligned}
AIC(p) &= \ln|\hat{\Sigma}(p)| + \frac{2}{N_{obs}}(pn^2 + nl + nm) \\
&= \ln|\hat{\Sigma}(p)| + \frac{2}{N_{obs}}(9p + 39)
\end{aligned}
$$

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AIC | -7.899 | -8.498 | -8.526 | -8.527* | -8.527 | -8.523 | -8.512 | -8.496 | -8.484 | -8.470 | -8.464 |

## Benchmark Result

An example of OLS for predicting spread in short term.

| | | | |
|---|---|---|---|
| **Dep. Variable:** | spread_res_ST | **R-squared:** | 0.306 |
| **Model:** | OLS | **Adj. R-squared:** | 0.306 |
| **Method:** | Least Squares | **F-statistic:** | 5844. |
| **Date:** | Thu, 14 May 2020 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 19:27:31 | **Log-Likelihood:** | 46084. |
| **No. Observations:** | 291889 | **AIC:** | -9.212e+04 |
| **Df Residuals:** | 291866 | **BIC:** | -9.188e+04 |
| **Df Model:** | 22 | | |
| **Covariance Type:** | nonrobust | | |

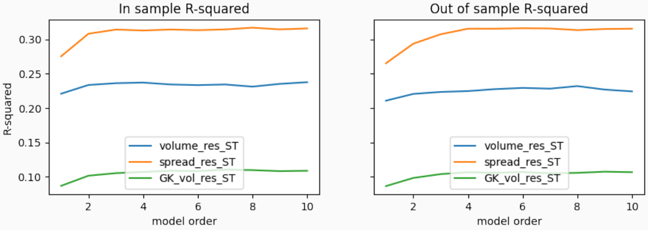
| | **coef** | **std err** | **t** | **P> \|t\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **const** | 0.4123 | 0.003 | 126.664 | 0.000 | 0.406 | 0.419 |
| **volume_res_ST_lag_1** | -0.0042 | 0.000 | -9.010 | 0.000 | -0.005 | -0.003 |
| **volume_res_ST_lag_2** | -0.0006 | 0.000 | -1.263 | 0.207 | -0.002 | 0.000 |
| **volume_res_ST_lag_3** | -0.0004 | 0.000 | -0.964 | 0.335 | -0.001 | 0.000 |
| **spread_res_ST_lag_1** | 0.3558 | 0.002 | 192.806 | 0.000 | 0.352 | 0.359 |
| **spread_res_ST_lag_2** | 0.1581 | 0.002 | 81.643 | 0.000 | 0.154 | 0.162 |
| **spread_res_ST_lag_3** | 0.1143 | 0.002 | 62.149 | 0.000 | 0.111 | 0.118 |

# Benchmark Result

| | coef | std err | t | P> \|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **GK_vol_res_ST_lag_1** | 0.0051 | 0.001 | 7.673 | 0.000 | 0.004 | 0.006 |
| **GK_vol_res_ST_lag_2** | 0.0057 | 0.001 | 8.401 | 0.000 | 0.004 | 0.007 |
| **GK_vol_res_ST_lag_3** | 0.0042 | 0.001 | 6.356 | 0.000 | 0.003 | 0.006 |
| **volume_res_LT** | 0.0077 | 0.001 | 9.569 | 0.000 | 0.006 | 0.009 |
| **spread_res_LT** | -0.0463 | 0.001 | -32.491 | 0.000 | -0.049 | -0.044 |
| **GK_vol_res_LT** | 0.2887 | 0.051 | 5.629 | 0.000 | 0.188 | 0.389 |
| **interest_res_LT** | 0.0014 | 0.001 | 1.981 | 0.048 | 1.48e-05 | 0.003 |
| **return_res_LT** | 0.4400 | 0.101 | 4.347 | 0.000 | 0.242 | 0.638 |
| **interest_res_ST** | -0.0186 | 0.001 | -26.718 | 0.000 | -0.020 | -0.017 |
| **return_res_ST** | 8.914e-08 | 1.46e-07 | 0.613 | 0.540 | -1.96e-07 | 3.74e-07 |
| **time_to_expiry** | 2.141e-06 | 3.61e-06 | 0.592 | 0.554 | -4.94e-06 | 9.23e-06 |
| **day_of_week** | 0.0021 | 0.000 | 7.684 | 0.000 | 0.002 | 0.003 |
| **week_of_month** | 0.0007 | 0.000 | 2.329 | 0.020 | 0.000 | 0.001 |
| **week_of_year** | -8.395e-05 | 2.64e-05 | -3.186 | 0.001 | -0.000 | -3.23e-05 |
| **is_vacation** | 0.0961 | 0.003 | 37.455 | 0.000 | 0.091 | 0.101 |
| **liquidity_rank** | 0.0003 | 0.000 | 0.972 | 0.331 | -0.000 | 0.001 |

# Benchmark Result

- We have got R-squared for different model order.




- We have used AIC to select the model order. The bette order is 3, the correspond R-squared result is shown below.

| R-squared | volume | volatility | spread |
|---|---|---|---|
| in sample | 0.236 | 0.103 | 0.305 |
| out of sample | 0.221 | 0.104 | 0.315 |

# Analysis of source error

- We have predicted the liquidity metrics in short term. To predict the variable itself, we use

$$pred(L^i_{t+1}(r)) = pred(ST(L^i_{t+1}(r))) \times LT(L^i_t(r))) \times \bar{L}^i(\delta_j)$$

The R-square is shown in table below.

| R-squared | volume | volatility | spread |
|---|---|---|---|
| in sample | 0.653 | 0.440 | 0.653 |
| out of sample | 0.647 | 0.335 | 0.565 |

- Here we have two source errors, one part is the prediction of short term, the second is the assumption that long term is constant. In order to analysis the source of error, we do

$$pred(L^i_{t+1}(r)) = pred(ST(L^i_{t+1}(r))) \times LT(L^i_{t+1}(r))) \times \bar{L}^i(\delta_j)$$

The new R-square is shown in table below.

| R-squared | volume | volatility | spread |
|---|---|---|---|
| in sample | 0.679 | 0.494 | 0.677 |
| out of sample | 0.672 | 0.372 | 0.609 |

# Conclusion

- We can see that linear model has a better prediction on short term of spread and volume, but it has a fair performance on volatility.
- This model will be used as a benchmark for the comparison with deep model.

## Apprendice I

An example of OLS for predicting volume in short term.

| | | | |
|---|---|---|---|
| **Dep. Variable:** | volume_res_ST | **R-squared:** | 0.237 |
| **Model:** | OLS | **Adj. R-squared:** | 0.237 |
| **Method:** | Least Squares | **F-statistic:** | 4119. |
| **Date:** | Thu, 14 May 2020 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 19:47:27 | **Log-Likelihood:** | -3.7463e+05 |
| **No. Observations:** | 291889 | **AIC:** | 7.493e+05 |
| **Df Residuals:** | 291866 | **BIC:** | 7.495e+05 |
| **Df Model:** | 22 | | |
| **Covariance Type:** | nonrobust | | |

| | **coef** | **std err** | **t** | **P> \|t\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **const** | 0.6937 | 0.014 | 50.430 | 0.000 | 0.667 | 0.721 |
| **volume_res_ST_lag_1** | 0.3271 | 0.002 | 165.958 | 0.000 | 0.323 | 0.331 |
| **volume_res_ST_lag_2** | 0.0988 | 0.002 | 49.247 | 0.000 | 0.095 | 0.103 |
| **volume_res_ST_lag_3** | 0.0624 | 0.002 | 32.948 | 0.000 | 0.059 | 0.066 |
| **spread_res_ST_lag_1** | -0.1089 | 0.008 | -13.966 | 0.000 | -0.124 | -0.094 |
| **spread_res_ST_lag_2** | -0.0163 | 0.008 | -1.993 | 0.046 | -0.032 | -0.000 |
| **spread_res_ST_lag_3** | -0.0184 | 0.008 | -2.365 | 0.018 | -0.034 | -0.003 |

# Apprendice I

| | coef | std err | t | P> \|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **GK_vol_res_ST_lag_1** | 0.0034 | 0.003 | 1.193 | 0.233 | -0.002 | 0.009 |
| **GK_vol_res_ST_lag_2** | -0.0014 | 0.003 | -0.485 | 0.628 | -0.007 | 0.004 |
| **GK_vol_res_ST_lag_3** | -0.0130 | 0.003 | -4.636 | 0.000 | -0.019 | -0.008 |
| **volume_res_LT** | -0.1862 | 0.003 | -54.746 | 0.000 | -0.193 | -0.180 |
| **spread_res_LT** | -0.0106 | 0.006 | -1.759 | 0.079 | -0.022 | 0.001 |
| **GK_vol_res_LT** | 1.4327 | 0.217 | 6.609 | 0.000 | 1.008 | 1.858 |
| **interest_res_LT** | 0.0684 | 0.003 | 22.667 | 0.000 | 0.062 | 0.074 |
| **return_res_LT** | -0.8127 | 0.428 | -1.900 | 0.057 | -1.651 | 0.026 |
| **interest_res_ST** | 0.1455 | 0.003 | 49.338 | 0.000 | 0.140 | 0.151 |
| **return_res_ST** | 4.754e-08 | 6.15e-07 | 0.077 | 0.938 | -1.16e-06 | 1.25e-06 |
| **time_to_expiry** | -1.051e-05 | 1.53e-05 | -0.688 | 0.492 | -4.04e-05 | 1.94e-05 |
| **day_of_week** | 0.0007 | 0.001 | 0.573 | 0.566 | -0.002 | 0.003 |
| **week_of_month** | 0.0029 | 0.001 | 2.277 | 0.023 | 0.000 | 0.005 |
| **week_of_year** | -0.0013 | 0.000 | -11.647 | 0.000 | -0.002 | -0.001 |
| **is_vacation** | -0.2816 | 0.011 | -25.965 | 0.000 | -0.303 | -0.260 |
| **liquidity_rank** | 0.0003 | 0.001 | 0.279 | 0.780 | -0.002 | 0.003 |

An example of OLS for predicting volatility in short term.

| | | | |
|---|---|---|---|
| **Dep. Variable:** | GK_vol_res_ST | **R-squared:** | 0.104 |
| **Model:** | OLS | **Adj. R-squared:** | 0.104 |
| **Method:** | Least Squares | **F-statistic:** | 1533. |
| **Date:** | Thu, 14 May 2020 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 19:50:48 | **Log-Likelihood:** | -2.5916e+05 |
| **No. Observations:** | 291889 | **AIC:** | 5.184e+05 |
| **Df Residuals:** | 291866 | **BIC:** | 5.186e+05 |
| **Df Model:** | 22 | | |
| **Covariance Type:** | nonrobust | | |

| | **coef** | **std err** | **t** | **P> \|t\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **const** | 0.6252 | 0.009 | 67.506 | 0.000 | 0.607 | 0.643 |
| **volume_res_ST_lag_1** | 0.0112 | 0.001 | 8.447 | 0.000 | 0.009 | 0.014 |
| **volume_res_ST_lag_2** | 0.0035 | 0.001 | 2.567 | 0.010 | 0.001 | 0.006 |
| **volume_res_ST_lag_3** | 0.0029 | 0.001 | 2.245 | 0.025 | 0.000 | 0.005 |
| **spread_res_ST_lag_1** | 0.0326 | 0.005 | 6.210 | 0.000 | 0.022 | 0.043 |
| **spread_res_ST_lag_2** | 0.0148 | 0.006 | 2.687 | 0.007 | 0.004 | 0.026 |
| **spread_res_ST_lag_3** | -0.0113 | 0.005 | -2.167 | 0.030 | -0.022 | -0.001 |

# Apprendice I

| | coef | std err | t | P> \|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **GK_vol_res_ST_lag_1** | 0.2375 | 0.002 | 125.500 | 0.000 | 0.234 | 0.241 |
| **GK_vol_res_ST_lag_2** | 0.1041 | 0.002 | 53.850 | 0.000 | 0.100 | 0.108 |
| **GK_vol_res_ST_lag_3** | 0.0618 | 0.002 | 32.709 | 0.000 | 0.058 | 0.066 |
| **volume_res_LT** | -0.0315 | 0.002 | -13.758 | 0.000 | -0.036 | -0.027 |
| **spread_res_LT** | -0.0268 | 0.004 | -6.595 | 0.000 | -0.035 | -0.019 |
| **GK_vol_res_LT** | -2.2362 | 0.146 | -15.321 | 0.000 | -2.522 | -1.950 |
| **interest_res_LT** | 0.0131 | 0.002 | 6.443 | 0.000 | 0.009 | 0.017 |
| **return_res_LT** | -2.4265 | 0.288 | -8.424 | 0.000 | -2.991 | -1.862 |
| **interest_res_ST** | -0.0025 | 0.002 | -1.281 | 0.200 | -0.006 | 0.001 |
| **return_res_ST** | -3.845e-07 | 4.14e-07 | -0.929 | 0.353 | -1.2e-06 | 4.27e-07 |
| **time_to_expiry** | 1.236e-05 | 1.03e-05 | 1.201 | 0.230 | -7.8e-06 | 3.25e-05 |
| **day_of_week** | 0.0077 | 0.001 | 9.910 | 0.000 | 0.006 | 0.009 |
| **week_of_month** | -0.0028 | 0.001 | -3.267 | 0.001 | -0.005 | -0.001 |
| **week_of_year** | -0.0006 | 7.5e-05 | -8.205 | 0.000 | -0.001 | -0.000 |
| **is_vacation** | -0.1055 | 0.007 | -14.446 | 0.000 | -0.120 | -0.091 |
| **liquidity_rank** | -0.0006 | 0.001 | -0.717 | 0.474 | -0.002 | 0.001 |