Analysis of Return Predictions from Trade Flow

Name: Qiqi Chen

Student Number: 12249648

1.Data

The data include trades and quotes for pairs of 'ETH-BTC', 'ETH-USD' and 'BTC-USD' in 2018. The trades data include the information like timestamp, price, size and who initiated the trade. The book data includes snapshots of the top the book and mid price with respect to each timestamp. Therefore, trade flow can be easily calculated and from the trades data and we can get forward return from the change of mid price in book data.

2.Strategy

The basic logic of the strategy to profit in the market is that if trade flow can be used to estimate the forward trend of the market, we'll use the information of current trade flow to predict future returns. For τ and T, I chose 300 seconds and 600 seconds based on the frequency of data. Split the data into training set and testing set with a training period of 3 hours.

For each timestamp that trade happened, I calculated the flow by formula $F_t^{(\tau)} = V_{(t-\tau,t)}^B - V_{(t-\tau,t)}^S$, and then I calculated the forward return in T seconds by formula $r_t^{(T)} = \frac{P_{t+T}^{Mid}}{P_t^{Mid}} - 1$. In the training dataset, I ran the regression between forward return and trade flow to get β. In the testing dataset, I used $\beta * F_t^{(\tau)}$ as a strategy indicator. If its absolute value is larger than j, I entered the market with a long or short position depending on the sign of indicator. Then my strategy return for each entry actually equals my position*actual forward return.

3.Analysis

(1) Reliability of β

To examine the reliability of β, I firstly checked the value and T-stats of it. From my point of view, when trade flow is positive, which means a lot of buyer-initiated trades happen, the price of the contract should go up in the next few minutes. Therefore, β should have a positive sign. But we can see from the pair 'BTC-USD', the value is negative. Also, the absolute t-values for 'ETH-BTC' and 'ETH-USD' show the β is significant to forward return prediction. Therefore, the model fits well in sample for some pairs.

	ETH-BTC	ETH-USD	BTC-USD
Beta	5.365325E-20	1.141435E-23	-3.900087E-24
T value	13.604891	2.073968	-0.884327
Observations	3339	27573	38374

At the same time, we want to test the model in the testing dataset to check if trade flow can predict forward return accurately. Therefore, I summarized its prediction accuracy by using the following table. Here 'POSITIVE' means positive return, 'NEGATIVE' means negative return. The accuracy is the number of cases of true positive and true negative divided by total case number. For 'ETC-BTC', although t value of β is very large, the model has a bad out of sample performance in the testing dataset. It could be due to sample selection bias. For 'ETC-USD', the accuracy exceeds 0.5, so the model can be used to predict forward, but it still needs to be refined. For 'BTC-USD', the accuracy still is smaller than 0.5. When the prediction accuracy is even lower than 0.5, we don't even need to construct a prediction model to predict the market trend.

ETC-BTC		
	POSITIVE	NEGATIVE
TRUE	4688	1463
FALSE	2592	3612
Accuracy: 0.497855		

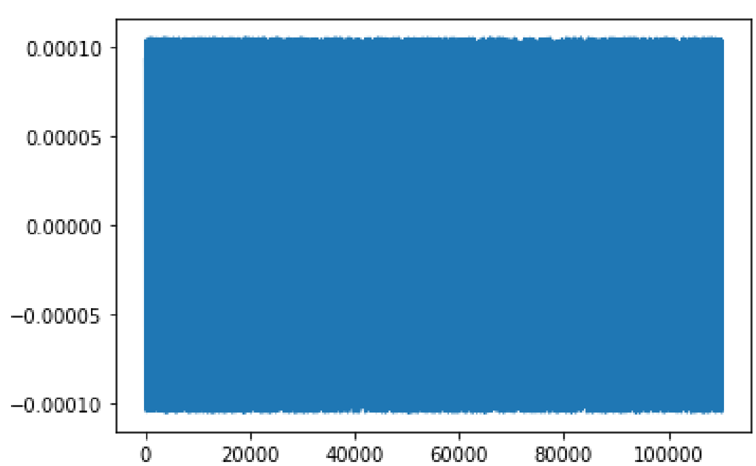
ETC-USD		
	POSITIVE	NEGATIVE
TRUE	27786	23470
FALSE	23588	27375
Accuracy: 0.501433		

BTC-USD		
	POSITIVE	NEGATIVE
TRUE	40211	35992
FALSE	35788	40517
Accuracy: 0.499666		

(2) How to Choose j

Basically, I created a list of potential j with large gaps between each element in the list at first. Then I excluded some j that results in bad performance. I measure the performance by using Sharpe ratio with benchmark return to be 0.

More specifically, take 'ETC-USD' as an example. Below is the plot of estimated forward return.



I created the list [0.00001,0.00002,0.00003,0.00004,0.00005,0.00006,0.00007,0.00008,0.00009] as my initial choices of j and I found the strategy Sharpe ratio were generally decreasing as j goes up.

Therefore, I restricted my j to a range of [0.000015, 0.00003] as my choices of j. Then my best choice of j is 0.000025.

j	Sharpe
0.00001	3.85E-05
0.00002	0.002403
0.00003	0.001949
0.00004	0.001575
0.00005	-0.00153
0.00006	-0.00205
0.00007	-0.00487
0.00008	-0.00596
0.00009	-0.01471

j	Sharpe
0.000015	0.001296
0.000016	0.001566
0.000017	0.001701
0.000018	0.001857
0.000019	0.002409
0.000020	0.002403
0.000021	0.002306
0.000022	0.001975
0.000023	0.002280
0.000024	0.002328
0.000025	0.003048
0.000026	0.002239
0.000027	0.002512
0.000028	0.002749
0.000029	0.002266

(3) Return Analysis

After choosing the best j for each pair, I calculated the return of each entry. The return I defined can also be used to measure the model's performance on the testing dataset.

As we can see from the table below, 'ETH-USD' got the highest mean and Sharpe under this strategy. So, we can still have significant profit by analyzing trade flow. The Sharpe ratio for the other two pairs are neither positive. All the return distribution is quite symmetric. All the return distribution has fat tails which means the strategy has some disadvantages when extreme situation occurs.

	ETH-BTC	ETH-USD	BTC-USD
j	0.000180	0.000025	0.000014
count	2753	77722	93467
mean	-0.000035	0.000012	-0.000024
std	0.002002	0.003785	0.003538
min	-0.007697	-0.024477	-0.036156
25%	-0.001185	-0.001793	-0.001405
50%	0.000000	0.000000	0.000000
75%	0.000513	0.001804	0.001398
max	0.009362	0.025245	0.034722
Sharpe	-0.017335	0.005598	-0.006806
skewness	1.090447	0.089221	-0.062757
kurtosis	2.426433	3.899869	13.251583

(4) Training and Testing Period

From my intuition, I expect a longer training period will lead to a stronger evidence that trade flow matters for forward returns. Also a longer testing period will be helpful to check the validity of the model.

Finally, I think a rolling window of training period can be better to produce good β, since market regime can change in 240 hours, although it can cause computation burden to some extent.