Sunrise Futures Financial Modeling Data Competition

Short Summary

University of Illinois at Urbana-Champaign

Yiwei Shen
Jiahao Zhu
Donghan Liu
Pengyu Chen
Ho Hin Jeremy Wong

1. First Glance of Data

Our goal is to optimize models to increase the prediction power of y_values. At the first glance of data, we observed that y_values are discrete. As shown in the picture, we aim to find the jump points of y_values, trying to explain why an y_value changes suddenly at a specific time point.

V
y_value
0.0625
0.0625
0.03125
0.03125
0.03125
0.03125
0.03125
0.03125
0
0

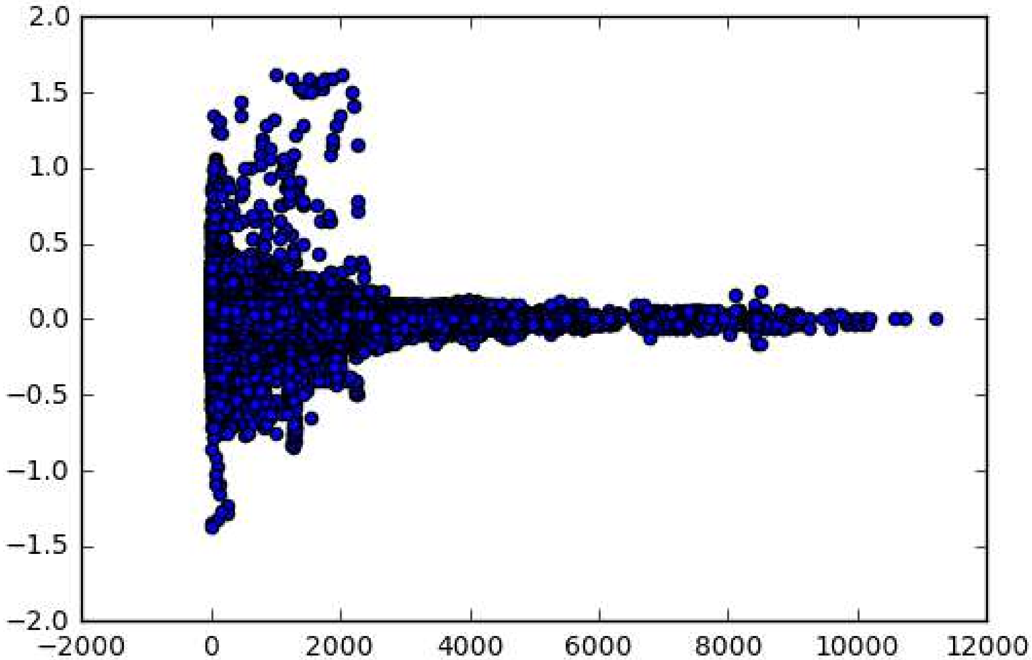
We believe that the status of the current time point depends only on the status of the time point 1 second before but not the past due to the nature of high frequency trading as we assumed. Therefore, we created 1 second time lag in our data set. We shifted the previous second trading price and size down to the current second, and then tried to predict y_value based on the data of the previous second.

```
df['pre_bid_price_inst1'] = df['bid_price_inst1'].shift(1)
df['pre_bid_size_inst1'] = df['bid_size_inst1'].shift(1)
df['pre_ask_price_inst1'] = df['ask_price_inst1'].shift(1)
df['pre_ask_size_inst1'] = df['ask_size_inst1'].shift(1)
df['pre_mid_price_inst1'] = df['mid_price_inst1'].shift(1)

df['pre_bid_price_inst2'] = df['bid_price_inst2'].shift(1)
df['pre_bid_size_inst2'] = df['bid_size_inst2'].shift(1)
df['pre_ask_price_inst2'] = df['ask_price_inst2'].shift(1)
df['pre_ask_size_inst2'] = df['ask_size_inst2'].shift(1)
df['pre_mid_price_inst2'] = df['mid_price_inst2'].shift(1)

df['pre_bid_price_inst3'] = df['bid_price_inst3'].shift(1)
df['pre_bid_size_inst3'] = df['bid_size_inst3'].shift(1)
df['pre_ask_price_inst3'] = df['ask_price_inst3'].shift(1)
df['pre_ask_size_inst3'] = df['ask_size_inst3'].shift(1)
df['pre_mid_price_inst3'] = df['mid_price_inst3'].shift(1)
```

2. Relation Exploration

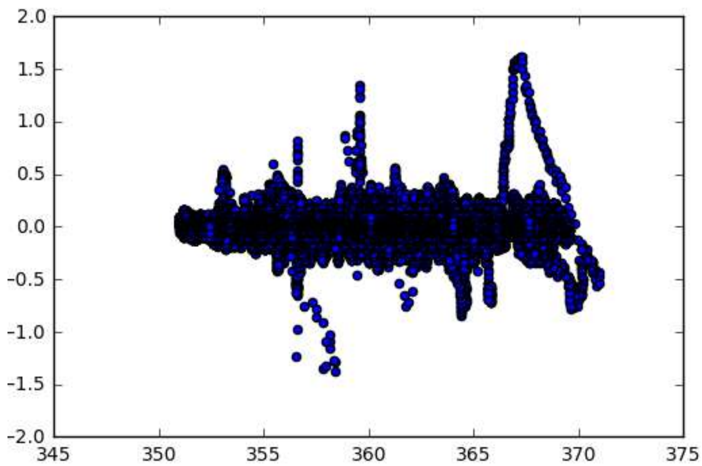
After obtaining our new shifted data set, we began to scatter plot relations between y_value and other independent variables.

1) Size:

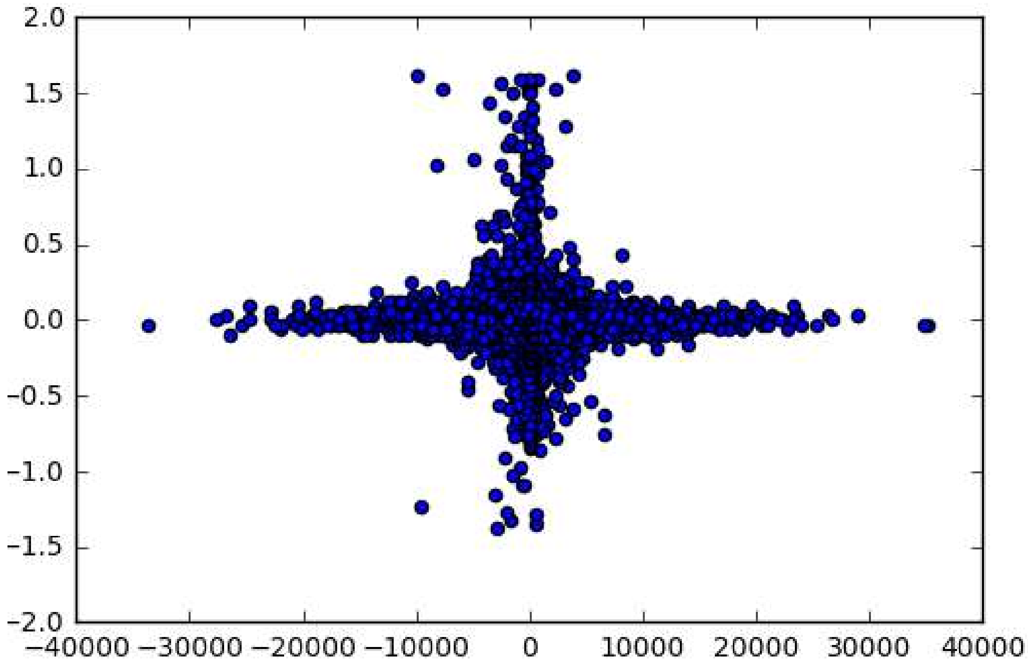


We observed that the y_value v.s. size plots have the same pattern for all three instruments. Small sized observations tend to have higher variances in y_values.

2) Price:
There are no obvious patterns in y_value v.s. prices for all three instruments.



3) Net Trade



The plots of y_values v.s. net trade indicating a pattern that when net traded volumes are close to 0, the variances of y_values become larger. We inspired from these plots that the difference of trading sizes instead of sizes themselves are important to our models. Therefore, we created new columns representing price_diff and size_diff for three instruments.

3. Regression

1) Price

Since we are not sure whether there is any relationship between price and y_value, we tried to regress y_value on prices first.

Dep. Variable:	y_value	R-squared:	0.000
Model:	OLS	Adj. R-squared:	0.000
Method:	Least Squares	F-statistic:	74.19
Date:	Sun, 16 Apr 2017	Prob (F-statistic):	5.61e-48
Time:	10:57:13	Log-Likelihood:	3.1814e+06
No. Observations:	1781941	AIC:	-6.363e+06
Df Residuals:	1781937	BIC:	-6.363e+06
Df Model:	3		
Covariance Type:	nonrobust		

The result turns out that the price is not so much related to y_values.

2) Size
   As we have discovered before, the size differences could be more useful than sizes themselves. Therefore, we tried to regress y_values on size differences.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                y_value   R-squared:                       0.011
Model:                            OLS   Adj. R-squared:                  0.011
Method:                 Least Squares   F-statistic:                     6765.
Date:                Sun, 16 Apr 2017   Prob (F-statistic):               0.00
Time:                        10:59:47   Log-Likelihood:             3.1914e+06
No. Observations:             1781941   AIC:                        -6.383e+06
Df Residuals:                 1781937   BIC:                        -6.383e+06
Df Model:                           3
Covariance Type:            nonrobust
```

It turns out size differences does explain some of the behaviors of y_values.

Further, we tried to include interactions between size_diffs. It turns out that the R-square does not increase much. Therefore, simple model without intersection is more appropriate.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                y_value   R-squared:                       0.011
Model:                            OLS   Adj. R-squared:                  0.011
Method:                 Least Squares   F-statistic:                     2907.
Date:                Sun, 16 Apr 2017   Prob (F-statistic):               0.00
Time:                        11:01:58   Log-Likelihood:             3.1914e+06
No. Observations:             1781941   AIC:                        -6.383e+06
Df Residuals:                 1781933   BIC:                        -6.383e+06
Df Model:                           7
Covariance Type:            nonrobust
```

3) Size*Price
   Now we consider the model include intersections between size_diff and price_diff

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                y_value   R-squared:                       0.011
Model:                            OLS   Adj. R-squared:                  0.011
Method:                 Least Squares   F-statistic:                     2907.
Date:                Sun, 16 Apr 2017   Prob (F-statistic):               0.00
Time:                        11:01:58   Log-Likelihood:             3.1914e+06
No. Observations:             1781941   AIC:                        -6.383e+06
Df Residuals:                 1781933   BIC:                        -6.383e+06
Df Model:                           7
Covariance Type:            nonrobust
------------------------------------------------------------------------------
```

   The model is not improved as well.

4. Final Model

In conclusion, we choose **y_value~ size_diff1 +size_diff2+size_diff3** as our final model due to consideration of R-square and concise of our model.