1.Explore the dataset

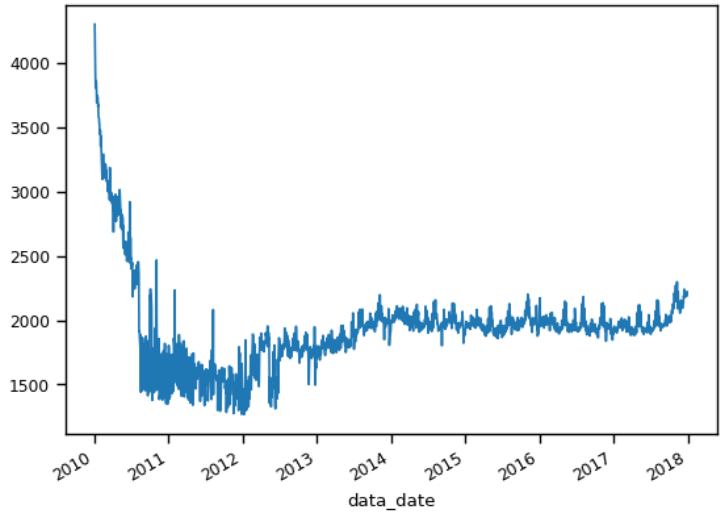
Data_set_n : I recognize them as factors which have predictive effect.

Fillna: for the non-consecutive time series factors, we use forward value to fill NA.
After the merge which all factors and return data, we further fillna by the mean value of the factor.
Then we summarize the different characters of the dataset:

	File	Average Data Frequency (days)	Data Length	Avg IDs/Day	NA Count(%)	Time Span Start	Time Span End	Total Unique IDs	Tradable Unique IDs
0	data_set_1.csv	102	127266	63.888554	3.762199	2010-01-04	2017-12-29	6522	3520
1	data_set_2.csv	102	127266	63.888554	11.500322	2010-01-04	2017-12-29	6522	3520
2	data_set_3.csv	46	254532	63.888554	21.477064	2010-01-04	2017-12-29	6522	3520
3	data_set_4.csv	46	254532	63.888554	3.671837	2010-01-04	2017-12-29	6522	3520
4	data_set_5.csv	102	127266	63.888554	11.835840	2010-01-04	2017-12-29	6522	3520
5	data_set_6.csv	102	127266	63.888554	19.638395	2010-01-04	2017-12-29	6522	3520
6	data_set_7.csv	102	127266	63.888554	18.957145	2010-01-04	2017-12-29	6522	3520
7	data_set_8.zip	2	8345449	4012.235096	0.000000	2010-01-04	2017-12-29	7068	3841
8	data_set_9.zip	2	16176736	4053.405365	0.000000	2010-01-04	2017-12-29	7373	3719
9	data_set_10.csv	1	3281273	1630.041232	0.000000	2010-01-04	2017-12-29	3084	2219
10	data_set_11.csv	1	3281273	1630.041232	0.000000	2010-01-04	2017-12-29	3084	2219

We found that some factors are daily factors, like d10 and d11; And many factors are monthly all quarterly(maybe fundamental information), and span is from 2010-01-04 – 2017-12-29, And the total tradable securities are about 2219(if we want to keep all the factors)

After we merge all the dataset and reference data, the count of sec ids in time series are shown below:



2. Construct technical signals:

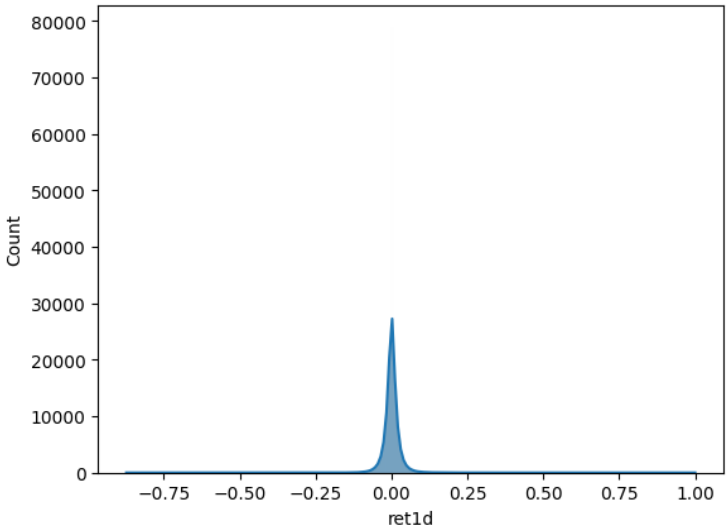
Use the common algo trading signals(by close_price and volume) , we do not have too many choices.

We have:
1. Moving Average
2. EMA - Exponential Moving Average
3. VWMA - Volume Weighted Moving Average
4. BBANDS - Bollinger Bands with different timeperiod
5. MOM – Momentum.
6. Acceleration - Difference in the change of momentum
7. Rate of Change - Rate change of Price.
8. Moving Average Convergence Divergence
9. RSI
10. Price Volume Trend
11. OBV - On Balance Volume
12. Psychological Line Indicator
13. Create volatility by rolling 3, 5, 15 days
14. Create 5, 15 days moving volume volatility
15. Correlation between volume and price.

With different parameters and time span.

The target we want to predict: next day's excess return
The excess return is the demean(split the beta) return of every stock

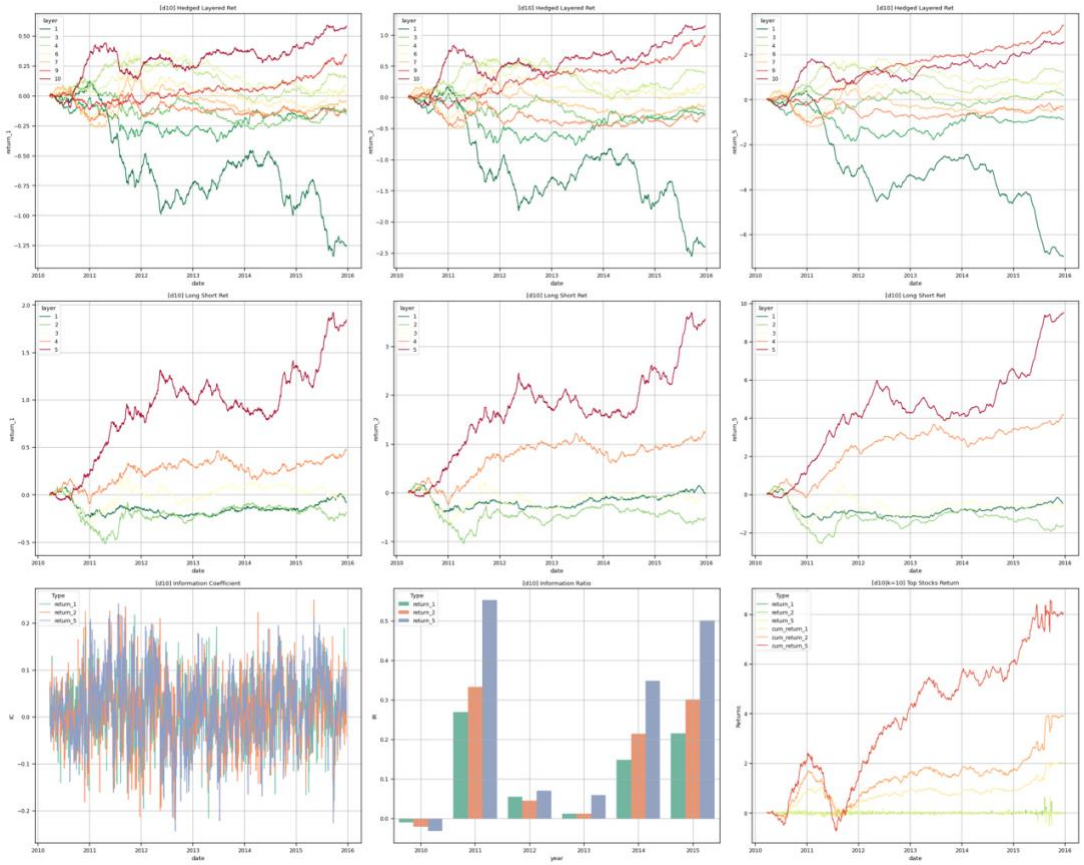
We show the return we want to predict, to see whether they are in a roughly normal distribution(The reason is that we need have balanced labels in samples)



3.Construct signal backtesting system

We split the securities into n layers(can modify) based on the value of the signal, and show the long short return , and ic across the timeseries on cross sections.
And we can change the class parameter to show the different days after return, here we use 1day, 2days, 5days

The class is AlphaSignalAnalyzer()
The example output for a single factor is like this:
We can decide to go long and short according to the layered performance of the factors



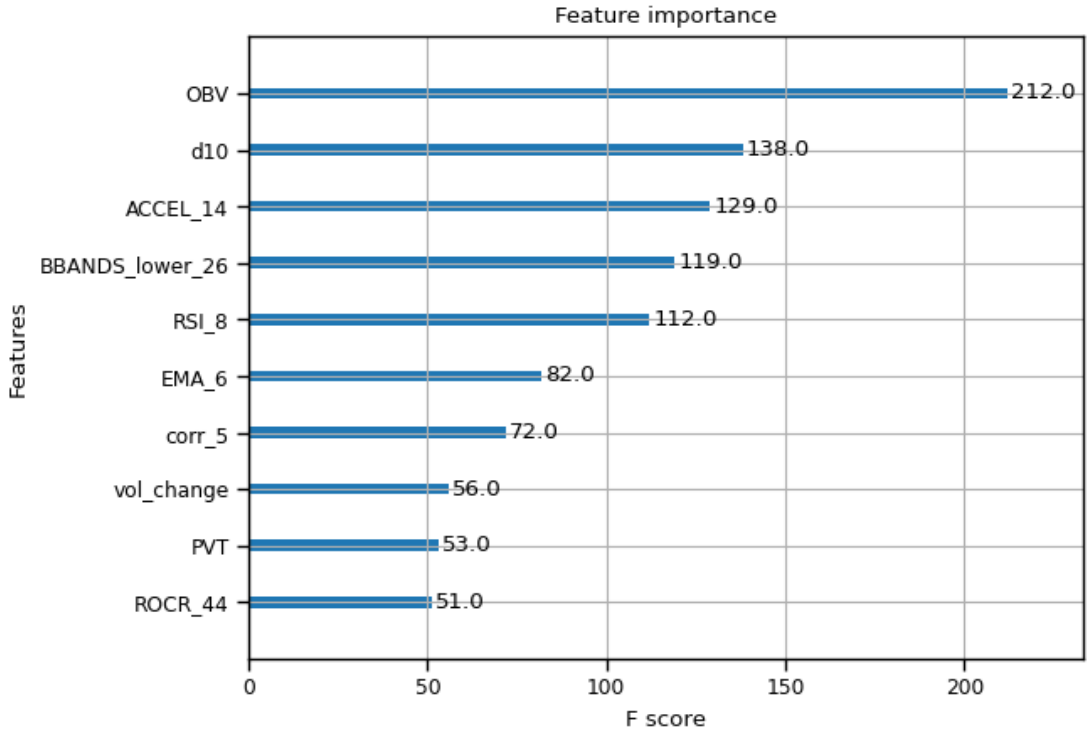
4. Machine Learning Method to ensemble factors

We choose models like linear regression, lasso regression (feature selection), and XGBosst

(non-linear relationship, tree model and boosting method to learn residuals), also use neural network like CNN and LSTM(consider the stock numbers are different in every cross sections and the location maybe vary a lot)

Train_set : start – 20151231
Valid_set: 20160101 – 20161231
Test_set: 20170101 – 20171231

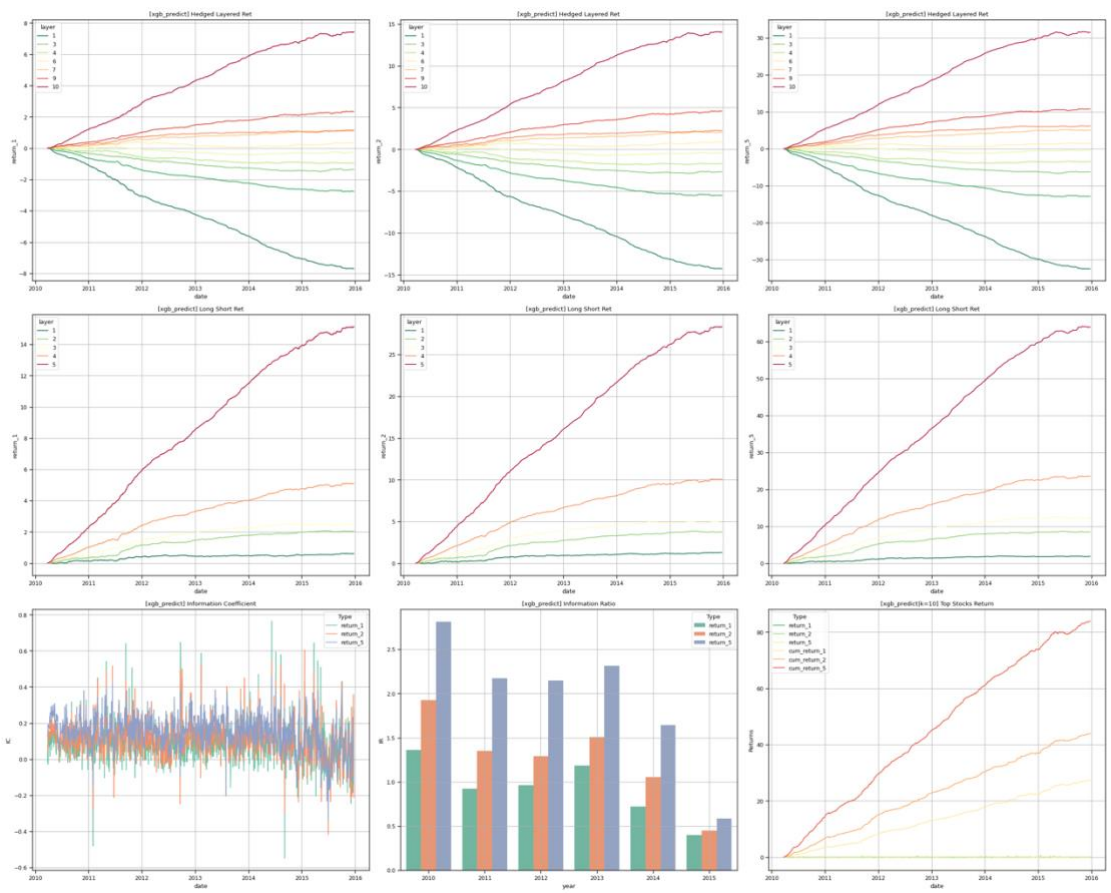
The XGBoost give us the direct understanding of feature importance:



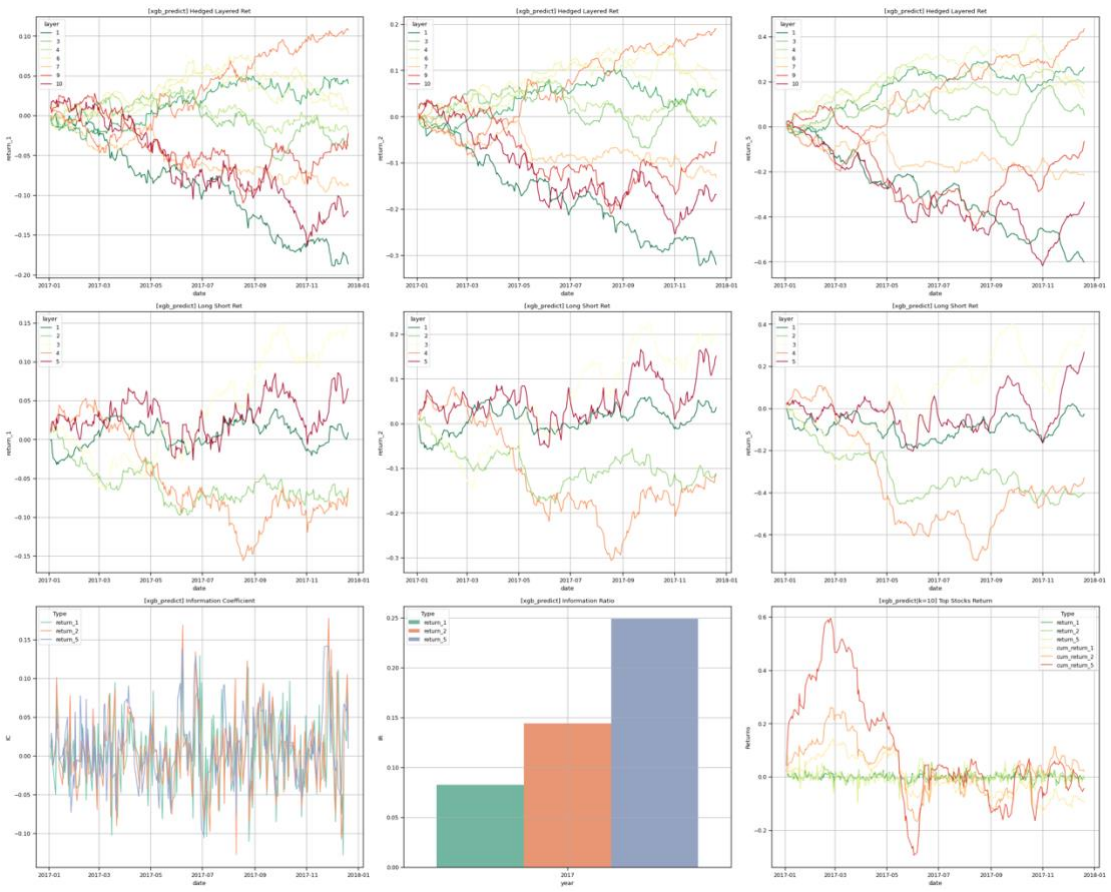
We will combine these top 10 factors with the factors selected by Lasso.

In the Traning Set of XGBoost, We got good performance. But in test set, the performance worse a lot (The factors are not powerful enough and the environment changes in 2017, we maybe can choose rolling traing for the futher research)

Acturally we can see the ic value become lower year by year from 2010-2015



The XGBoost Test set performance:

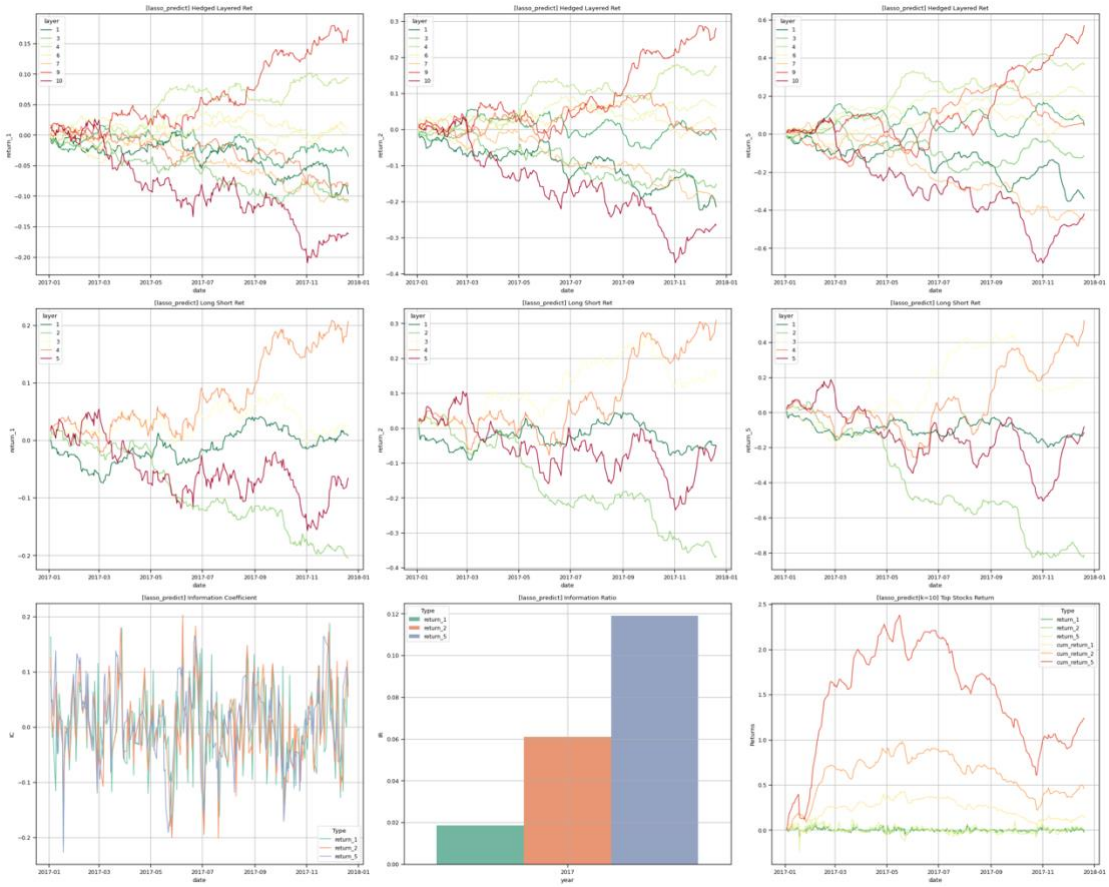


If I was given more time , I will think about the training separately for long end and short end. Maybe Lasso is more suitable for long end and XGBoost is more suitable for short end. And add more details into that.

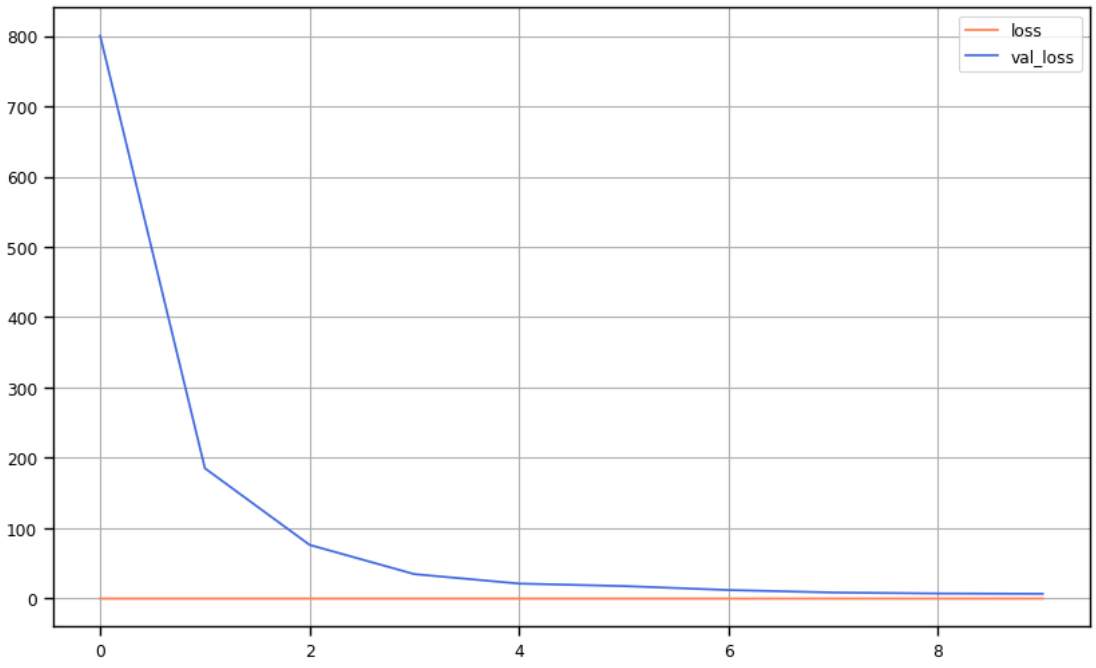
Also we try the Lasso, and got the best test set performance by getting these features:

	coef
vol_change	8.811401e-05
EMA_30	7.633829e-05
sd_15	1.790343e-05
d10	9.230566e-06
BBANDS_lower_14	6.326509e-06
d5	4.064636e-06
d6	3.434214e-06
volsd_15	1.219584e-06
VSMA_46	1.121822e-07
d11	-1.967449e-07
PVT	-4.067648e-07
d9	-1.303094e-06
ROCR_56	-2.452349e-06
BBANDS_upper_50	-4.610758e-06
d4	-4.808972e-06
d7	-6.049434e-06
EMA_50	-6.430125e-06
BBANDS_upper_20	-6.785751e-06
OBV	-7.773814e-06
BBANDS_lower_38	-1.251035e-05
VSMA_18	-2.178488e-05
RSI_8	-4.225711e-05
ROCR_32	-5.644952e-05
ROCR_14	-1.019981e-04
EMA_6	-5.284031e-04

And the performance in test set is :



And we use these selected features(29) on the neural network traing.



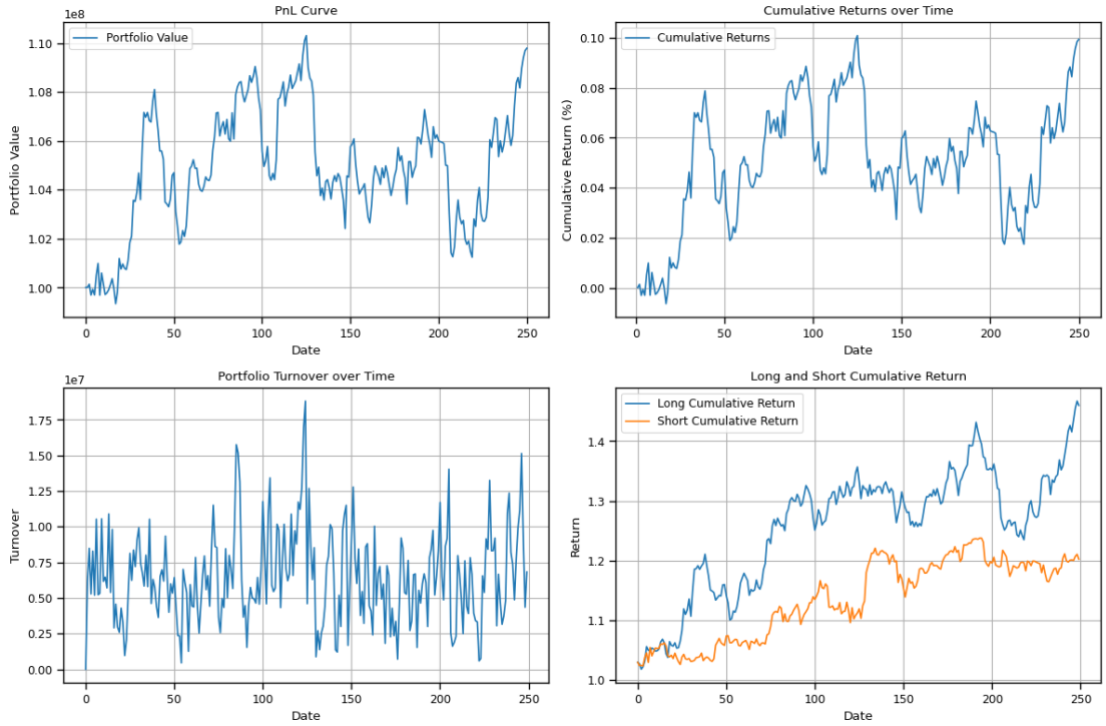
This the train and val loss converge for MLP.

And we also try the LSTM(need more time and computing power to tune the parameter, not fully trained)

5. Portfolio Construction

We long the fist n stocks, and short last n stocks by the rank of ensembled factors.
And the weights are decided by the optimization of min variance,

We design a PortfolioManager class to handle the portfolio construction.
And calculate the return, turnover, maxdrawdown, etc.
The final result may not the best result, we still need more time to futhuer tune the parameter.
And we keep the lomg end and short end are market neutral, we won't suffer from the beta loss.




```
Portfolio Summary
------------------
Sharpe Ratio: 0.057
Annualized Return: 0.098
Maximum Drawdown: 0.082
Long Positions: 0
Short Positions: 0
```

If given more time, we will spend more time to try

1. how to construct more powerful factors
2. how to choose the weights of the long and short end, and add more details into PortfolioManager class
3. Fully tune the parameter of Neural Network and Tree model

I think that must can give us better portfolio.