Alpha Trading Workflow

Analyst: Yuxuan Xia

Date: 2018/06/01

TODO

- Input more effective factors: take advice from people and industry reports
- Quaterly Data and Annually data, how to use them? Decrease the system frequency to quaterly?
- Improve perfomance through deep learning or statistical models?
- Find well-known metrics to express results

Workflow

\checkmark stands for finished and \vartriangle stands for TODO

- Universe definition
- Factors collection and preprocessing
  - △ Factors collection
    - Sources
      - balance sheet
      - cash flow statement
      - income statement
      - earning report
    - Econometric Classifications
      - value
      - growth
      - profitability
      - market size
      - liquidity
      - volatility
      - Momentom
      - Financial leverage (debt-to-equity ratio)
  - Factors preprocessing
    - △daily, quaterly, annually
    - continuous: rescale, outliers
    - ✓discrete: rank
- Factors screening and combination
  - Factors screening
    - ✓Factors' correlation
    - ✓Factors' foreseeablity
    - Fama-Macbeth regression
  - △Factors combination

        - PCA, FA
        - Financial Modeling
        - Linear combination to maximize Sharpe ratio
        - Non-linear learning algorithms
            - ✓AdaBoost
            - Reinforcement learning
- Portfolio allocation

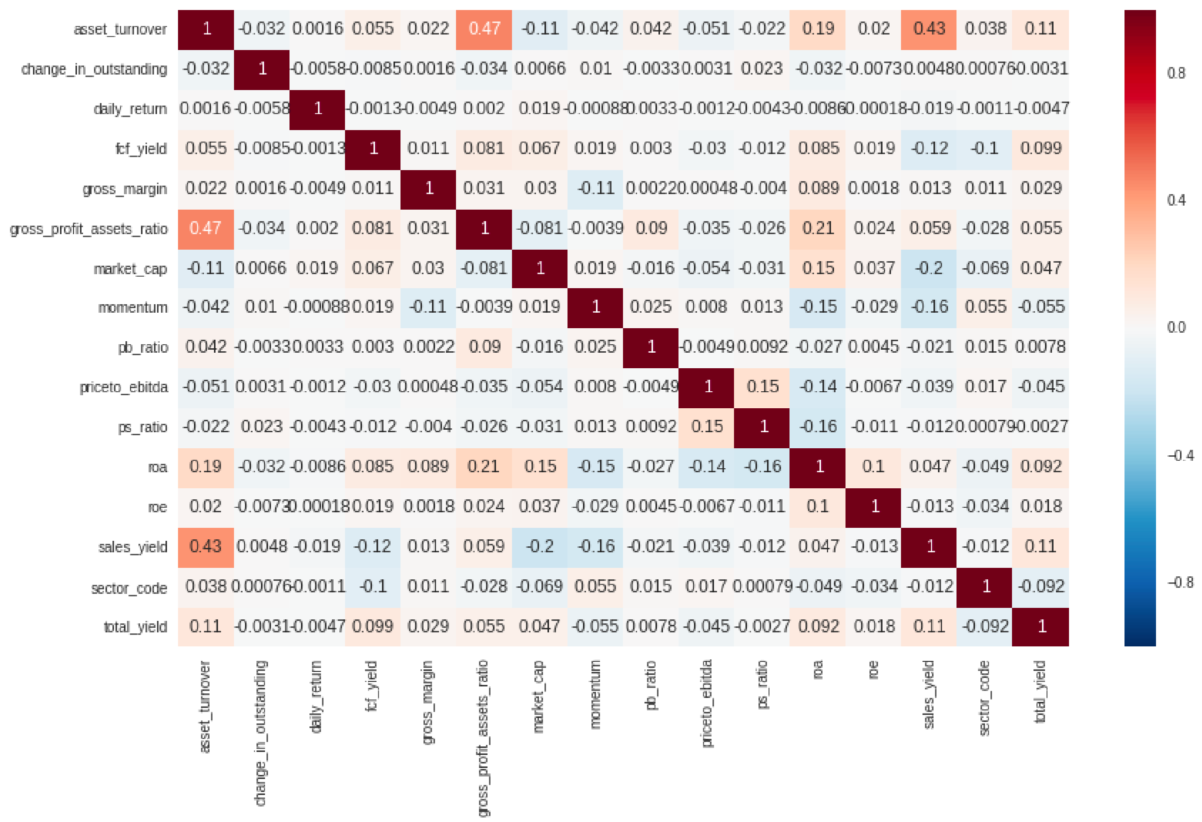
Factors' Correlations

Here, I use correlation matrix as the measure. The difference from the second result is that the correlation matrix is calculated by the rank data rather than the raw data

Two ICs comparison

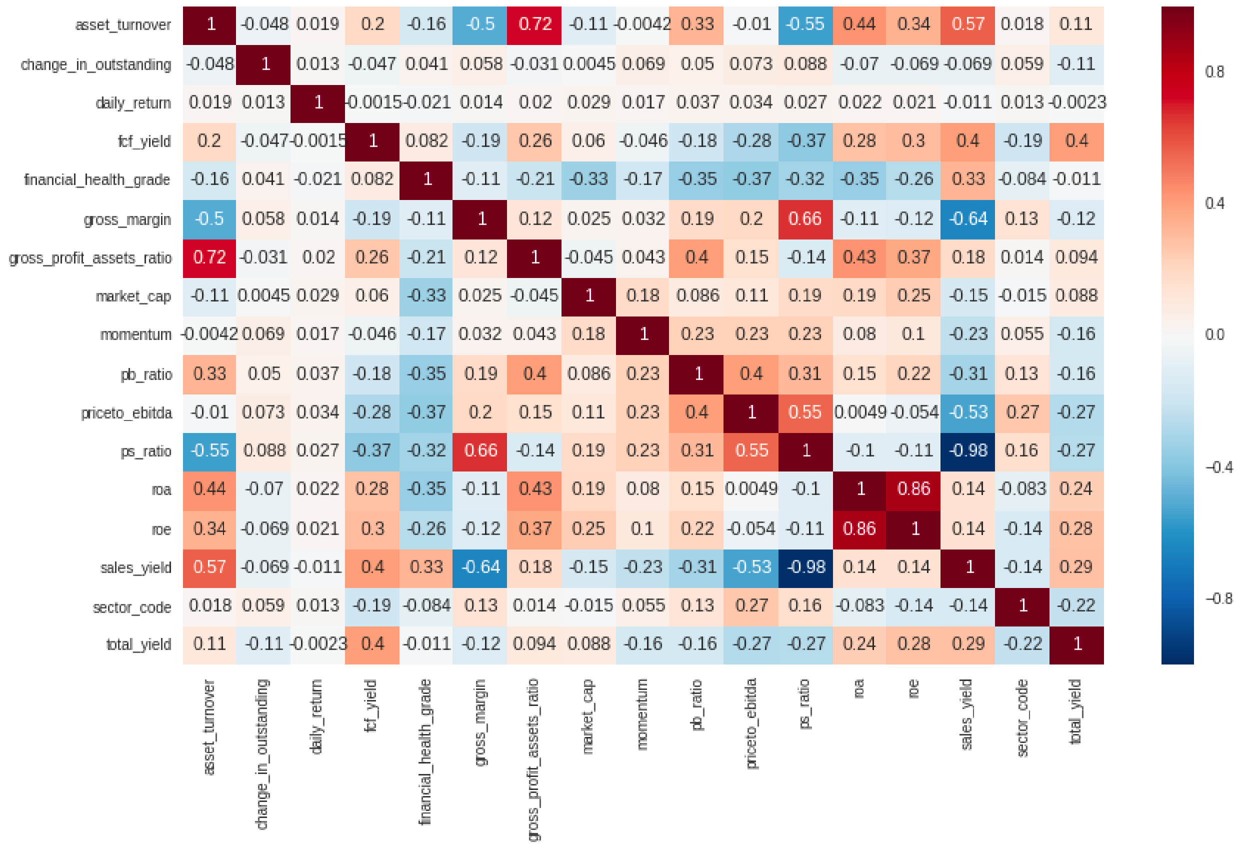
- Pearson's IC: If the sample size is moderate or large and the population is normal, then, in the case of the bivariate normal distribution, the sample correlation coefficient is the maximum likelihood estimate of the population correlation coefficient, and is asymptotically unbiased and efficient, which roughly means that it is impossible to construct a more accurate estimate than the sample correlation coefficient. The number itself has no sense if you don't find a proper way or "common sense" to interpret it. Multi-variate Gaussian distribution give us such a common sense of how it should looks like.
- Spearman's IC: while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). Since We only care about the monotonic relationships. Spearman's IC wins.

Regular IC(Pearson's correlation coefficient) for each factors



	asset_turnover	change_in_outstanding	daily_return	fcf_yield	gross_margin	gross_profit_assets_ratio	market_cap	momentum	pb_ratio	priceto_ebitda	ps_ratio	roa	roe	sales_yield	sector_code	total_yield
asset_turnover	1	-0.032	0.0016	0.055	0.022	0.47	-0.11	-0.042	0.042	-0.051	-0.022	0.19	0.02	0.43	0.038	0.11
change_in_outstanding	-0.032	1	-0.0058	-0.0085	0.0016	-0.034	0.0066	0.01	-0.0033	0.0031	0.023	-0.032	-0.0073	0.0048	0.00076	0.0031
daily_return	0.0016	-0.0058	1	-0.0013	-0.0049	0.002	0.019	-0.00088	0.0033	-0.0012	-0.0043	-0.0086	0.00018	-0.019	-0.0011	-0.0047
fcf_yield	0.055	-0.0085	-0.0013	1	0.011	0.081	0.067	0.019	0.003	-0.03	-0.012	0.085	0.019	-0.12	-0.1	0.099
gross_margin	0.022	0.0016	-0.0049	0.011	1	0.031	0.03	-0.11	0.0022	0.00048	-0.004	0.089	0.0018	0.013	0.011	0.029
gross_profit_assets_ratio	0.47	-0.034	0.002	0.081	0.031	1	-0.081	-0.0039	0.09	-0.035	-0.026	0.21	0.024	0.059	-0.028	0.055
market_cap	-0.11	0.0066	0.019	0.067	0.03	-0.081	1	0.019	-0.016	-0.054	-0.031	0.15	0.037	-0.2	-0.069	0.047
momentum	-0.042	0.01	-0.00088	0.019	-0.11	-0.0039	0.019	1	0.025	0.008	0.013	-0.15	-0.029	-0.16	0.055	-0.055
pb_ratio	0.042	-0.0033	0.0033	0.003	0.0022	0.09	-0.016	0.025	1	-0.0049	0.0092	-0.027	0.0045	-0.021	0.015	0.0078
priceto_ebitda	-0.051	0.0031	-0.0012	-0.03	0.00048	-0.035	-0.054	0.008	-0.0049	1	0.15	-0.14	-0.0067	-0.039	0.017	-0.045
ps_ratio	-0.022	0.023	-0.0043	-0.012	-0.004	-0.026	-0.031	0.013	0.0092	0.15	1	-0.16	-0.011	-0.012	0.00079	0.0027
roa	0.19	-0.032	-0.0086	0.085	0.089	0.21	0.15	-0.15	-0.027	-0.14	-0.16	1	0.1	0.047	-0.049	0.092
roe	0.02	-0.0073	0.00018	0.019	0.0018	0.024	0.037	-0.029	0.0045	-0.0067	-0.011	0.1	1	-0.013	-0.034	0.018
sales_yield	0.43	0.0048	-0.019	-0.12	0.013	0.059	-0.2	-0.16	-0.021	-0.039	-0.012	0.047	-0.013	1	-0.012	0.11
sector_code	0.038	0.00076	0.0011	-0.1	0.011	-0.028	-0.069	0.055	0.015	0.017	0.00079	-0.049	-0.034	-0.012	1	-0.092
total_yield	0.11	-0.0031	-0.0047	0.099	0.029	0.055	0.047	-0.055	0.0078	-0.045	-0.0027	0.092	0.018	0.11	-0.092	1

Spearman's Rank correlation coefficient for each factors



	asset_turnover	change_in_outstanding	daily_return	fcf_yield	financial_health_grade	gross_margin	gross_profit_assets_ratio	market_cap	momentum	pb_ratio	priceto_ebitda	ps_ratio	roa	roe	sales_yield	sector_code	total_yield
asset_turnover	1	-0.048	0.019	0.2	-0.16	-0.5	0.72	-0.11	-0.0042	0.33	-0.01	-0.55	0.44	0.34	0.57	0.018	0.11
change_in_outstanding	-0.048	1	0.013	-0.047	0.041	0.058	-0.031	0.0045	0.069	0.05	0.073	0.088	-0.07	-0.069	-0.069	0.059	-0.11
daily_return	0.019	0.013	1	-0.0015	-0.021	0.014	0.02	0.029	0.017	0.037	0.034	0.027	0.022	0.021	-0.011	0.013	-0.0023
fcf_yield	0.2	-0.047	-0.0015	1	0.082	-0.19	0.26	0.06	-0.046	-0.18	-0.28	-0.37	0.28	0.3	0.4	-0.19	0.4
financial_health_grade	-0.16	0.041	-0.021	0.082	1	-0.11	-0.21	-0.33	-0.17	-0.35	-0.37	-0.32	-0.35	-0.26	0.33	-0.084	-0.011
gross_margin	-0.5	0.058	0.014	-0.19	-0.11	1	0.12	0.025	0.032	0.19	0.2	0.66	-0.11	-0.12	-0.64	0.13	-0.12
gross_profit_assets_ratio	0.72	-0.031	0.02	0.26	-0.21	0.12	1	-0.045	0.043	0.4	0.15	-0.14	0.43	0.37	0.18	0.014	0.094
market_cap	-0.11	0.0045	0.029	0.06	-0.33	0.025	-0.045	1	0.18	0.086	0.11	0.19	0.19	0.25	-0.15	-0.015	0.088
momentum	-0.0042	0.069	0.017	-0.046	-0.17	0.032	0.043	0.18	1	0.23	0.23	0.23	0.08	0.1	-0.23	0.055	-0.16
pb_ratio	0.33	0.05	0.037	-0.18	-0.35	0.19	0.4	0.086	0.23	1	0.4	0.31	0.15	0.22	-0.31	0.13	-0.16
priceto_ebitda	-0.01	0.073	0.034	-0.28	-0.37	0.2	0.15	0.11	0.23	0.4	1	0.55	0.0049	-0.054	-0.53	0.27	-0.27
ps_ratio	-0.55	0.088	0.027	-0.37	-0.32	0.66	-0.14	0.19	0.23	0.31	0.55	1	-0.1	-0.11	-0.98	0.16	-0.27
roa	0.44	-0.07	0.022	0.28	-0.35	-0.11	0.43	0.19	0.08	0.15	0.0049	-0.1	1	0.86	0.14	-0.083	0.24
roe	0.34	-0.069	0.021	0.3	-0.26	-0.12	0.37	0.25	0.1	0.22	-0.054	-0.11	0.86	1	0.14	-0.14	0.28
sales_yield	0.57	-0.069	-0.011	0.4	0.33	-0.64	0.18	-0.15	-0.23	-0.31	-0.53	-0.98	0.14	0.14	1	-0.14	0.29
sector_code	0.018	0.059	0.013	-0.19	-0.084	0.13	0.014	-0.015	0.055	0.13	0.27	0.16	-0.083	-0.14	-0.14	1	-0.22
total_yield	0.11	-0.11	-0.0023	0.4	-0.011	-0.12	0.094	0.088	-0.16	-0.16	-0.27	-0.27	0.24	0.28	0.29	-0.22	1

How to rule out redundant factors and why Spearman's rank correlation coefficients?

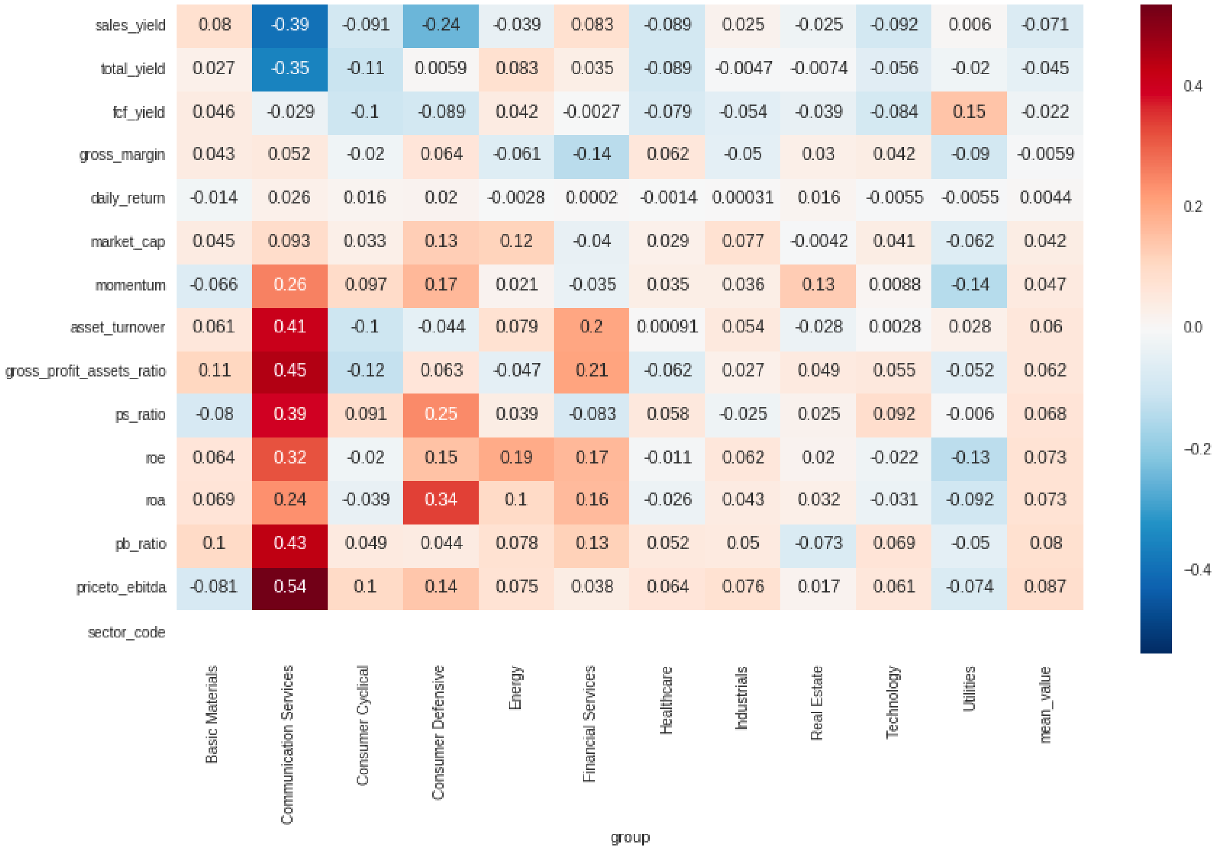
From the correlation coefficients below, we can again conclude that Spearman's rank IC is far more robust. Take ps_ratio and sales_yield as a example. $ps\_ratio = \frac{\text{adjusted close price}}{\text{sales per share}}$ whereas $sales\_yield = \frac{\text{sales per share}}{\text{price}}$ Ahthogh the price in sales_yield formula is vague in our data source we can see roughly speaking, these two variable should be inverse of each other. The Spearman's rank correlation coefficient is -0.98 which verifies this statement, and we should avoid using both of these factors, which would exeggarate the impact of this peticular factor. However, we can not see such identity in the Pearson's regular correlation coefficients. It's quite misleading actually and that's why we choose Spearman's rank IC.

Factors' Foreseeability

Mehods

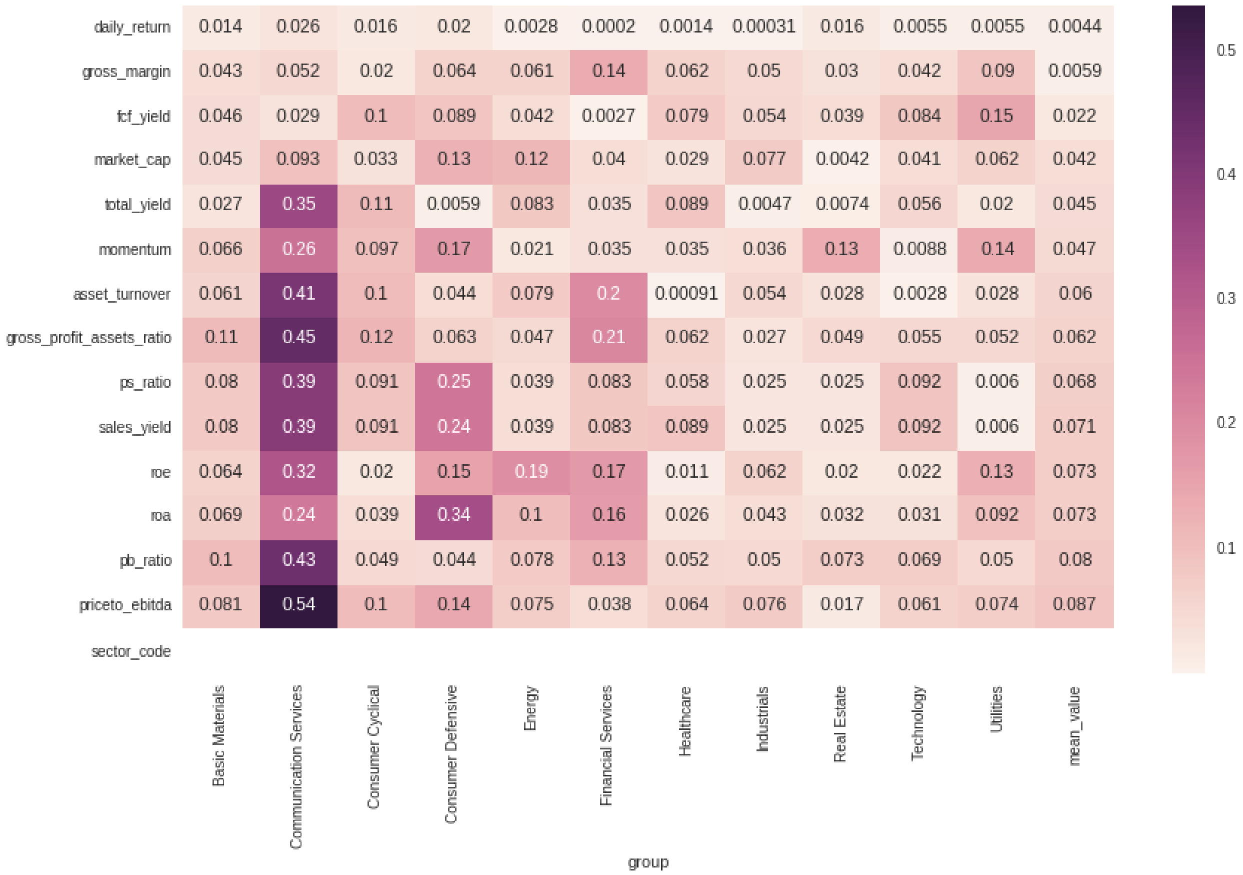
- Spearman's rank correlation coefficients
- Fama-Macbeth regression: Not only consider the foreseeability of factors itself but also consider the co-vary of different factors, which means rule out factors if the returns can be explained by the recent factors.

Spearman's rank IC for factors vs. forward returns



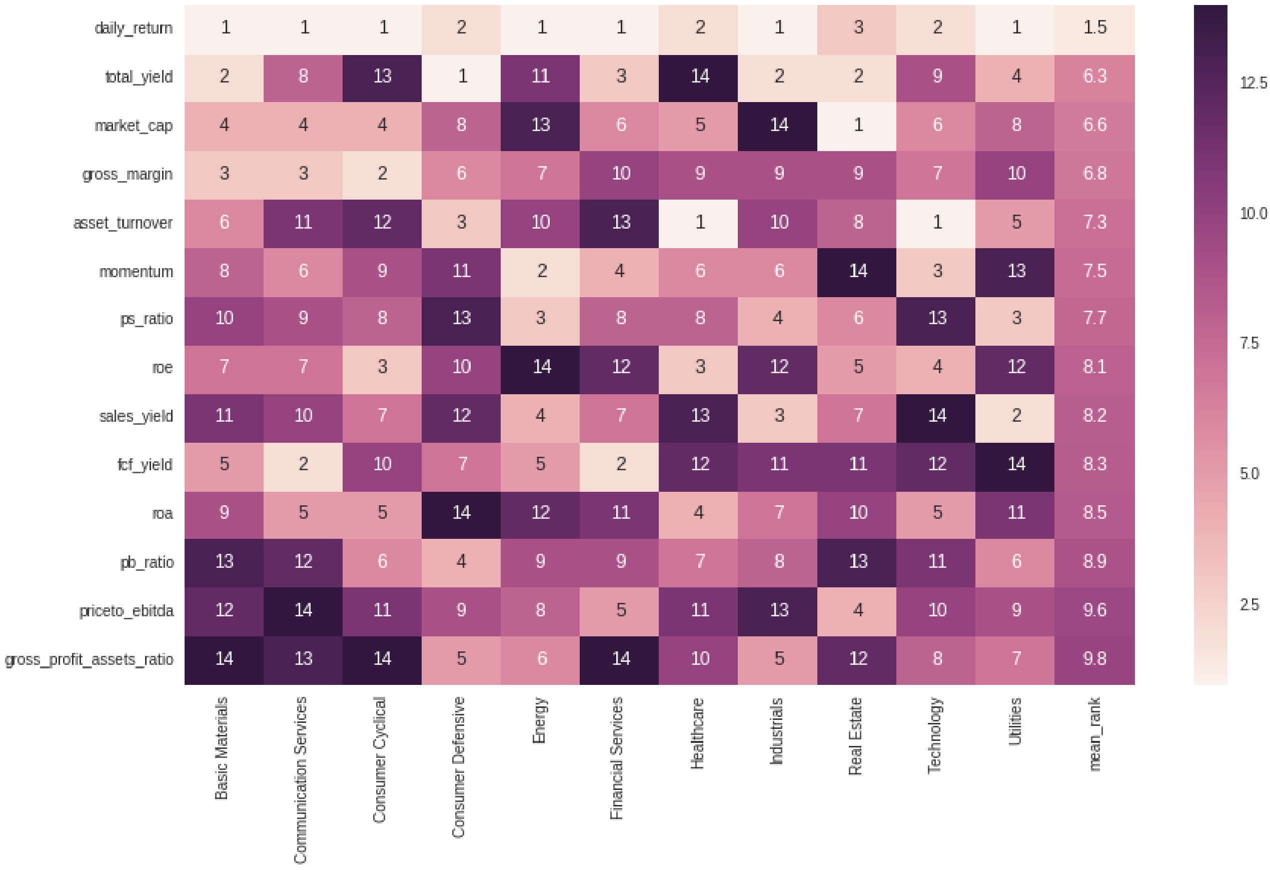
	Basic Materials	Communication Services	Consumer Cyclical	Consumer Defensive	Energy	Financial Services	Healthcare	Industrials	Real Estate	Technology	Utilities	mean_value
sales_yield	0.08	-0.39	-0.091	-0.24	-0.039	0.083	-0.089	0.025	-0.025	-0.092	0.006	-0.071
total_yield	0.027	-0.35	-0.11	0.0059	0.083	0.035	-0.089	-0.0047	-0.0074	-0.056	-0.02	-0.045
fcf_yield	0.046	-0.029	-0.1	-0.089	0.042	-0.0027	-0.079	-0.054	-0.039	-0.084	0.15	-0.022
gross_margin	0.043	0.052	-0.02	0.064	-0.061	-0.14	0.062	-0.05	0.03	0.042	-0.09	-0.0059
daily_return	-0.014	0.026	0.016	0.02	-0.0028	0.0002	-0.0014	0.00031	0.016	-0.0055	-0.0055	0.0044
market_cap	0.045	0.093	0.033	0.13	0.12	-0.04	0.029	0.077	-0.0042	0.041	-0.062	0.042
momentum	-0.066	0.26	0.097	0.17	0.021	-0.035	0.035	0.036	0.13	0.0088	-0.14	0.047
asset_turnover	0.061	0.41	-0.1	-0.044	0.079	0.2	0.00091	0.054	-0.028	0.0028	0.028	0.06
gross_profit_assets_ratio	0.11	0.45	-0.12	0.063	-0.047	0.21	-0.062	0.027	0.049	0.055	-0.052	0.062
ps_ratio	-0.08	0.39	0.091	0.25	0.039	-0.083	0.058	-0.025	0.025	0.092	-0.006	0.068
roe	0.064	0.32	-0.02	0.15	0.19	0.17	-0.011	0.062	0.02	-0.022	-0.13	0.073
roa	0.069	0.24	-0.039	0.34	0.1	0.16	-0.026	0.043	0.032	-0.031	-0.092	0.073
pb_ratio	0.1	0.43	0.049	0.044	0.078	0.13	0.052	0.05	-0.073	0.069	-0.05	0.08
priceto_ebitda	-0.081	0.54	0.1	0.14	0.075	0.038	0.064	0.076	0.017	0.061	-0.074	0.087
sector_code												

Spearman's rank IC (absolute value) for factors vs. forward returns



	Basic Materials	Communication Services	Consumer Cyclical	Consumer Defensive	Energy	Financial Services	Healthcare	Industrials	Real Estate	Technology	Utilities	mean_value
daily_return	0.014	0.026	0.016	0.02	0.0028	0.0002	0.0014	0.00031	0.016	0.0055	0.0055	0.0044
gross_margin	0.043	0.052	0.02	0.064	0.061	0.14	0.062	0.05	0.03	0.042	0.09	0.0059
fcf_yield	0.046	0.029	0.1	0.089	0.042	0.0027	0.079	0.054	0.039	0.084	0.15	0.022
market_cap	0.045	0.093	0.033	0.13	0.12	0.04	0.029	0.077	0.0042	0.041	0.062	0.042
total_yield	0.027	0.35	0.11	0.0059	0.083	0.035	0.089	0.0047	0.0074	0.056	0.02	0.045
momentum	0.066	0.26	0.097	0.17	0.021	0.035	0.035	0.036	0.13	0.0088	0.14	0.047
asset_turnover	0.061	0.41	0.1	0.044	0.079	0.2	0.00091	0.054	0.028	0.0028	0.028	0.06
gross_profit_assets_ratio	0.11	0.45	0.12	0.063	0.047	0.21	0.062	0.027	0.049	0.055	0.052	0.062
ps_ratio	0.08	0.39	0.091	0.25	0.039	0.083	0.058	0.025	0.025	0.092	0.006	0.068
sales_yield	0.08	0.39	0.091	0.24	0.039	0.083	0.089	0.025	0.025	0.092	0.006	0.071
roe	0.064	0.32	0.02	0.15	0.19	0.17	0.011	0.062	0.02	0.022	0.13	0.073
roa	0.069	0.24	0.039	0.34	0.1	0.16	0.026	0.043	0.032	0.031	0.092	0.073
pb_ratio	0.1	0.43	0.049	0.044	0.078	0.13	0.052	0.05	0.073	0.069	0.05	0.08
priceto_ebitda	0.081	0.54	0.1	0.14	0.075	0.038	0.064	0.076	0.017	0.061	0.074	0.087
sector_code												

Rank of the Spearman's rank IC (absolute value) for factors vs. forward returns



	Basic Materials	Communication Services	Consumer Cyclical	Consumer Defensive	Energy	Financial Services	Healthcare	Industrials	Real Estate	Technology	Utilities	mean_rank
daily_return	1	1	1	2	1	1	2	1	3	2	1	1.5
total_yield	2	8	13	1	11	3	14	2	2	9	4	6.3
market_cap	4	4	4	8	13	6	5	14	1	6	8	6.6
gross_margin	3	3	2	6	7	10	9	9	9	7	10	6.8
asset_turnover	6	11	12	3	10	13	1	10	8	1	5	7.3
momentum	8	6	9	11	2	4	6	6	14	3	13	7.5
ps_ratio	10	9	8	13	3	8	8	4	6	13	3	7.7
roe	7	7	3	10	14	12	3	12	5	4	12	8.1
sales_yield	11	10	7	12	4	7	13	3	7	14	2	8.2
fcf_yield	5	2	10	7	5	2	12	11	11	12	14	8.3
roa	9	5	5	14	12	11	4	7	10	5	11	8.5
pb_ratio	13	12	6	4	9	9	7	8	13	11	6	8.9
priceto_ebitda	12	14	11	9	8	5	11	13	4	10	9	9.6
gross_profit_assets_ratio	14	13	14	5	6	14	10	5	12	8	7	9.8

Alpha Factor Combination

construct an aggregate alpha factor which has its return distribution profitable. The term "profitable" here means condense, little turnover, significant in the positive return.

Methods

linear methods

- normalize factors and try a linear combination
- rank each factor and then sum up
- Financial modeling
- linear combination to maximize Sharpe ratio

Non-linear methods

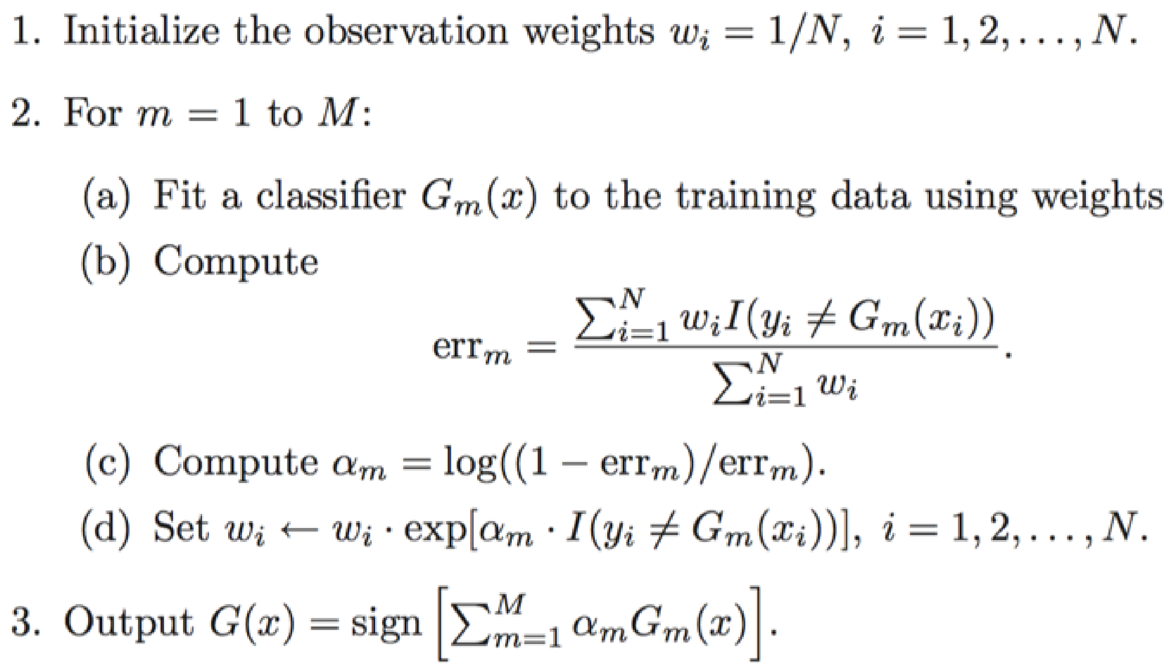
- AdaBoost
- Reinforement Learning

AdaBoost

Description

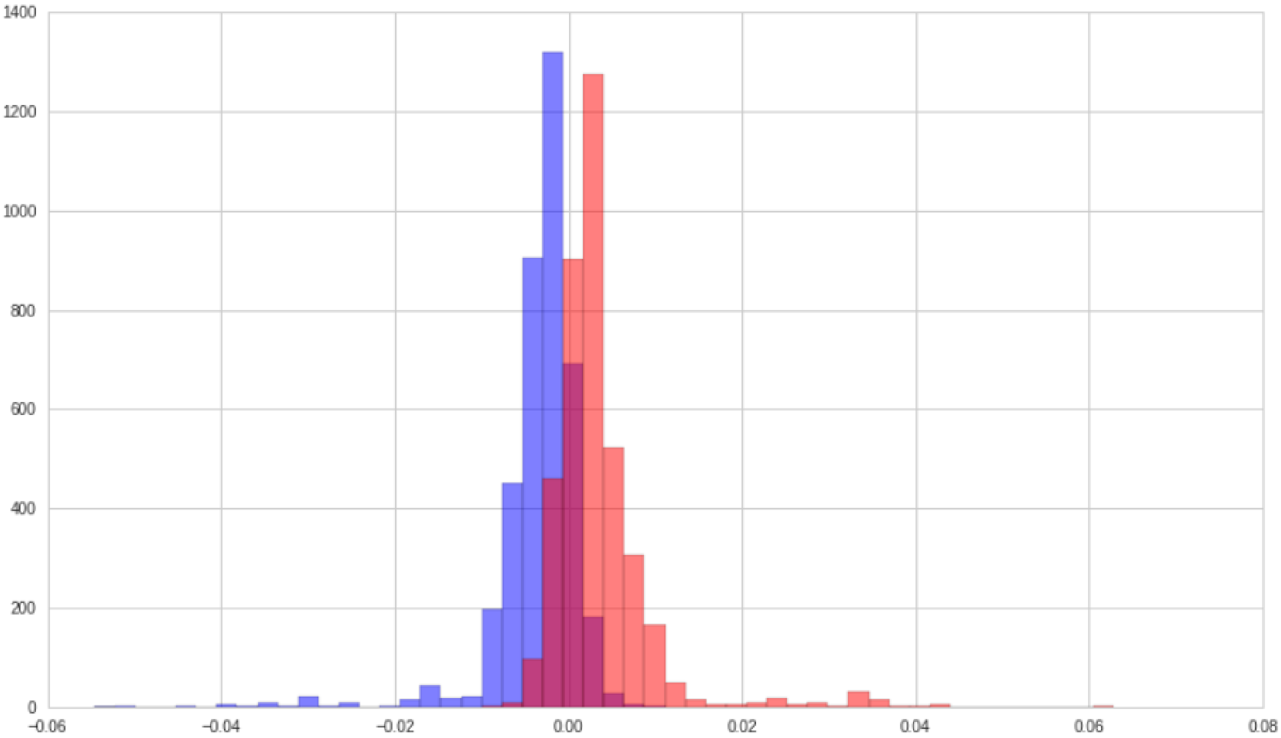
The algorithm sequentially applies a weak classification to modified versions of the data. By increasing the weights of the missclassified observations, each weak learner focuses on the error of the previous one. The predictions are aggregated through a weighted majority vote.

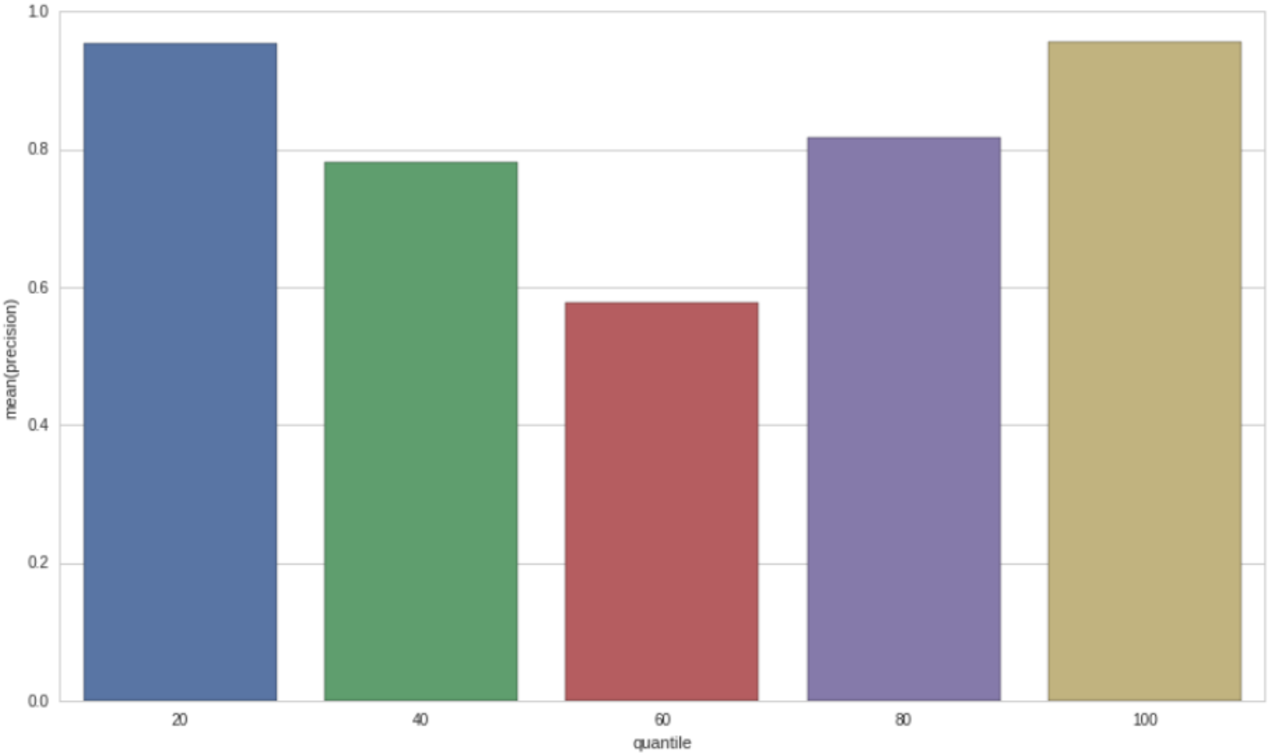
Algorithm



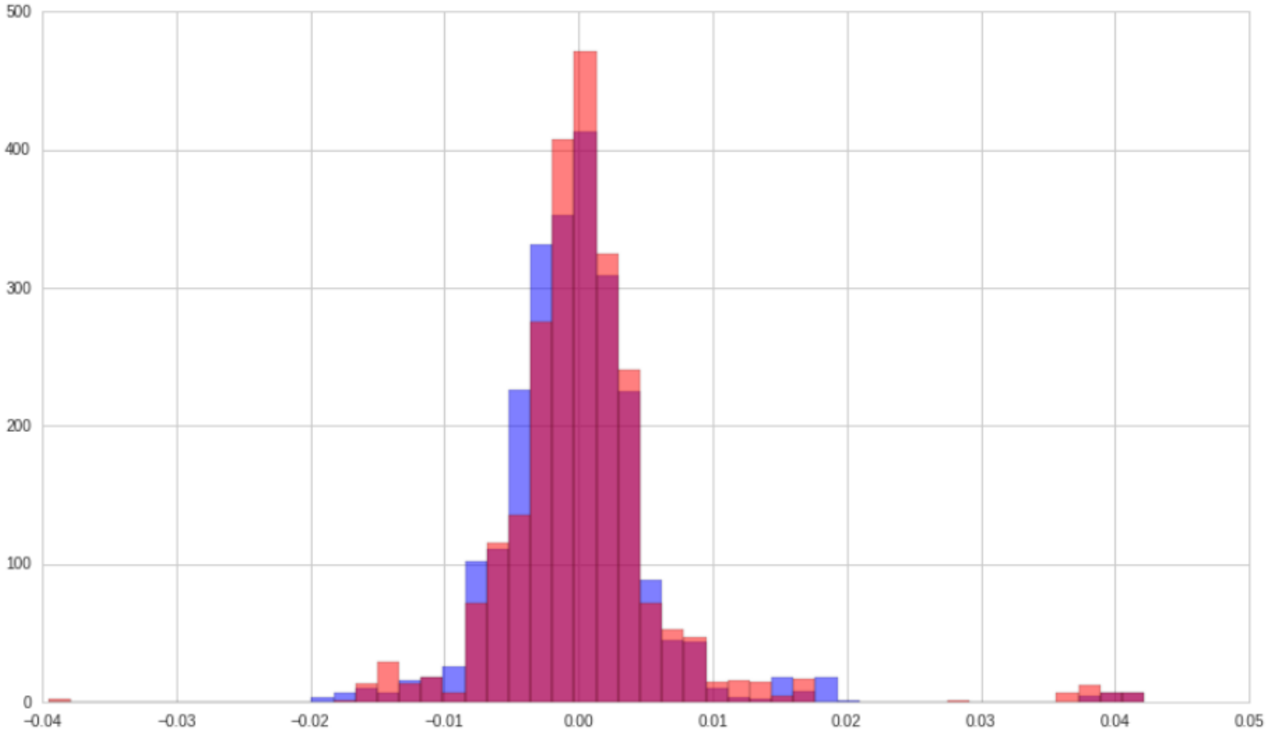
```
1. Initialize the observation weights $w_i = 1/N,\ i = 1, 2, \ldots, N$.
2. For $m = 1$ to $M$:
    (a) Fit a classifier $G_m(x)$ to the training data using weights $w_i$.
    (b) Compute
        $$\mathrm{err}_m = \frac{\sum_{i=1}^N w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^N w_i}.$$
    (c) Compute $\alpha_m = \log((1 - \mathrm{err}_m)/\mathrm{err}_m)$.
    (d) Set $w_i \leftarrow w_i \cdot \exp[\alpha_m \cdot I(y_i \neq G_m(x_i))],\ i = 1, 2, \ldots, N$.
3. Output $G(x) = \mathrm{sign}\left[\sum_{m=1}^M \alpha_m G_m(x)\right]$.
```

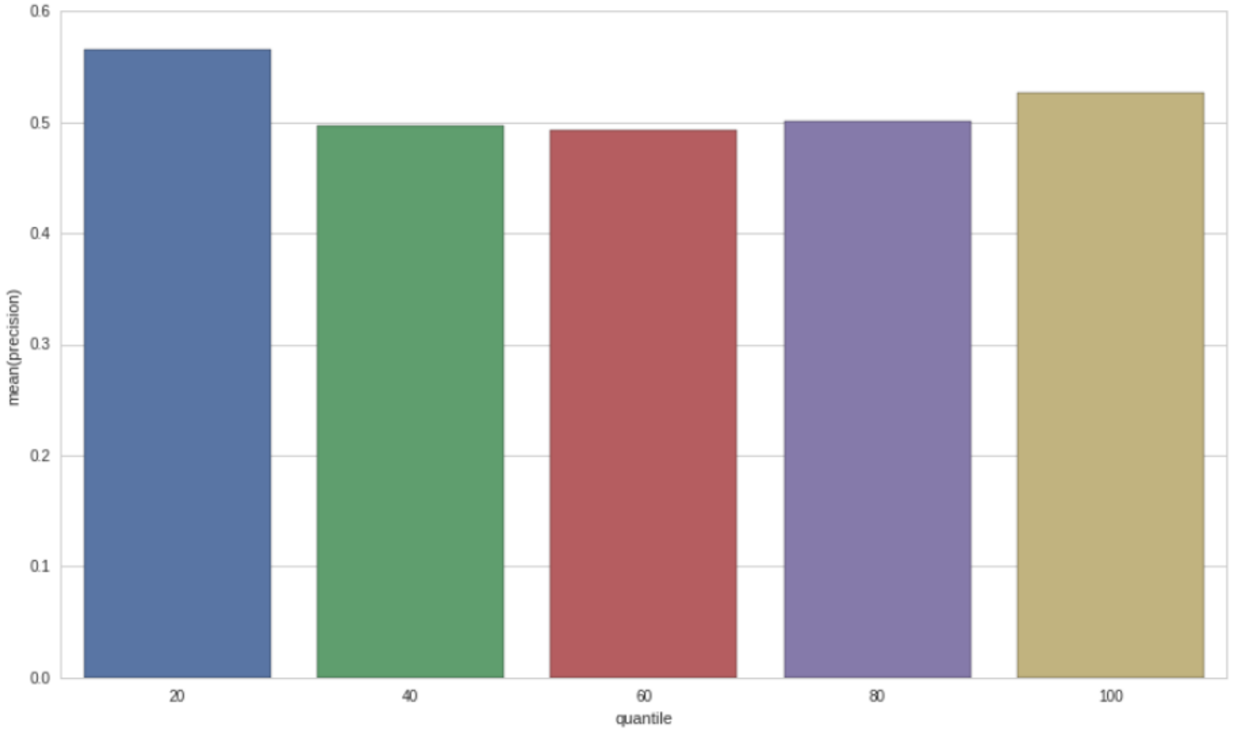
Train set


1400
1200
1000
800
600
400
200
0
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08


1.0
0.8
0.6
0.4
0.2
0.0
mean(precision)
20
40
60
80
100
quantile

Test set





References

- Jonathan Larkin, *A Professional Quant Equity Workflow*. August 31, 2016
- *A Practitioner's Guide to Factor Models*. The Research Foundation of The Institute of Chartered Financial Analysts
- Thomas Wiecki, Machine Learning on Quantopian
- Inigo Fraser Jenkins, *Using factors with different alpha decay times: The case for non-linear combination*
- PNC, *Factor Analysis: What Drives Performance?*
- O'Shaughnessy, *Alpha or Assets? — Factor Alpha vs. Smart Beta*. April 2016
- *O'Shaughnessy Quarterly Investor Letter Q1 2018*

- Jiantao Zhu, Orient Securities, *Alpha Forecasting - Factor-Based Strategy Research Series 13*
- Yang Song, Bohai Securities, *Multi-Factor Models Research: Single Factor Testing*, 2017/10/11