Which anomaly portfolio will make money in the next month?

20201181 Jihwan Oh

I. Introduction

I explain about out project motivation and purpose with literature review. In finance, anomaly refers to when the actual result differs from the expected result predicted by the model. Examples of anomaly include “January effect”. The January effect refers to a pattern in which stock prices, which performed poorly in the fourth quarter of last year in January, rise much more than in other months. According to the “THE HISTORY OF THE CROSS SECTION OF STOCK RETURNS” (2016):

『Asset pricing research continues to uncover new anomalies at an impressive rate. Harvey, Liu, and Zhu (2015) document 314 factors identified by the literature, with the majority being identified during the last 15 years. Cochrane (2011) summarizes the state of the literature by noting: “We thought 100% of the cross-sectional variation in expected returns came from the CAPM, now we think that’s about zero and a zoo of new factors describes the cross-section.”』

Therefore, anomaly portfolio is worth researching. Our team showed interest in this special feature, Anomaly, and we decided to analyze it through machine learning and deep learning.

The reason why we use ML/DL is as follows. The “Empirical Asset Pricing via Machine Learning” (2020) is a research paper that analyses popular machine learning techniques and how machine learning can be implemented into finance to create more efficient models with better results. Our goal is to predict and analyze the anomaly portfolio return in these two methods as recurrent neural network (RNN) and logistic regression (LR).

II. Data

DATE	beta_1	dtv_12	isff_1	ivff_1	me	srev	tv_1	eprd	\
1967-01	11.8004	-4.4299	3.5865	12.1847	-13.0562	-0.1126	11.4709	8.6986	
1967-02	2.1382	-2.7018	-1.2576	4.7215	-5.1638	2.8017	4.7486	-5.0862	
1967-03	0.2358	-1.6275	4.8673	0.6764	-3.3238	-1.5156	0.3803	-0.3200	
1967-04	3.0167	0.5737	-3.6645	-3.0035	-0.7099	-1.9171	-0.8516	-2.8200	
1967-05	1.3046	-6.4407	-1.2705	3.4395	-4.8868	-4.0637	0.6068	1.9194	
...	...	...	...	...	...	...	...	...	
2021-08	0.7348	1.3942	-1.4633	-1.3650	0.6421	2.9832	-0.4001	-1.1960	
2021-09	9.5032	-2.3366	1.9337	1.0624	-1.4418	-1.7750	1.1031	2.1055	
2021-10	1.7722	5.5664	0.9665	-2.3577	8.4341	2.6571	0.5204	-2.6409	
2021-11	-4.0661	4.6815	5.1696	-5.6675	7.9488	8.6912	-3.1802	-10.0804	
2021-12	-10.1847	1.9730	-2.0947	-15.6574	7.3379	5.0166	-19.7398	-2.3719	

[660 rows x 118 columns]

Fig.1: Monthly anomaly portfolio return data during 1967/01 to 2021/12.

We are using "Monthly Anomaly Data" which was handled in class. Since we have six anomaly files: frictions, intangibles, investment, momentum, profitability, value-growth. By merging all these data together, we get almost 200 anomalies portfolio return during 1967/01 to 2021/12. By dropping the anomalies which have at least one NA values. Finally, we have 118 anomalies during 1967/01 to 2021/12. *(Fig.1)*

Using python package *statsmodels.tsa.stattools – adfuller*, we can check the p-value of stationary. The maximum p-value in 118 anomalies is 0.00089. It doesn't over alpha = 0.05. Therefore, all anomaly portfolio return (time series) data is stationary. It means that we don't need to use scaling like log or standard scaling. Therefore, we use data without any scaling.

III. Models

Now, we get a preprocessed data. The flowchart of our model is as follows. *(Fig.2)*

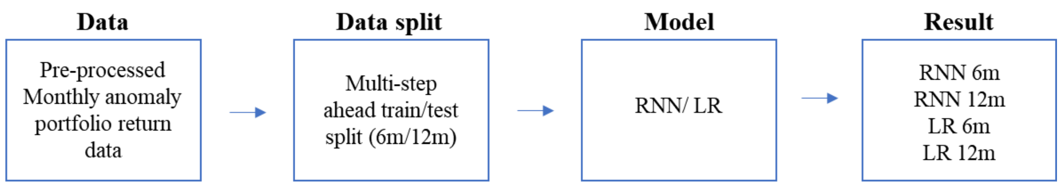


Fig.2: Flowchart of model

First, in the time series data, we use the train set as n number of data and then a multi-step ahead prediction method to predict the $(n+1)^{th}$ value. n is called 'window size'. For example, if window size n=12, predict the 2019/01 value using 2018/01 to 2018/12 data. In our project, we use window size as 6months and 12months. Because the portfolio returns are dependent with seasonality. 6 months and 12 months mean half a year and a year. These two are typical periods of seasonality. It shows in detail as follows. *(Fig.3)*

Then split the data (660months: 1967/01 to 2021/12) into train set (80%, 528 months: 1967/01 to 2010/12) and test set (20%, 132 months: 2011/01 to 2021/12).

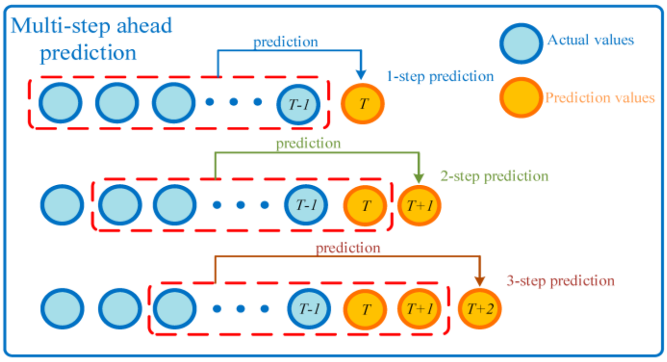
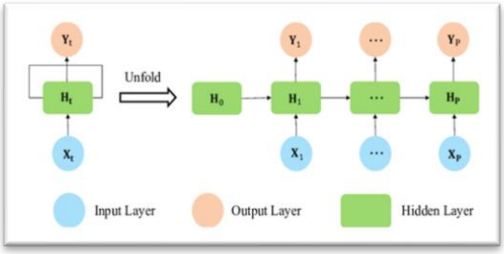


Fig.3: Multi-step ahead prediction method

We use two machine learning methods recurrent neural network (RNN) and logistic regression (LR) to predict anomaly portfolio returns.

RNN is an artificial neural network which uses to predict time series data. It has two powerful properties the hidden state stores a lot of information about the past efficiently and the non-linear dynamic allows then to update the hidden state in complicated ways. *(Fig.4)*

Logistic regression is a binary classification method. We labeled the data by setting the anomaly portfolio return to 0 if it is negative and 1 if it is positive. Now, we can find pattern by train set and then, we can classify which anomaly portfolio return is positive or negative in test set. *(Fig.5)*



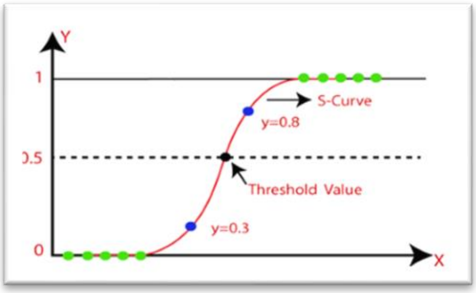


Fig.4: Recurrent neural network (RNN)

Fig.5: Logistic regression (LR)

IV. Result & Conclusion

Fig.6 shows that predict the value of anomaly portfolio returns using RNN in the test set period. *Fig.7* shows that predict the value of probability (close to 1: positive return, close to 0: negative return) using LR in the test set period. Both window sizes are 6 months. Now, we use the winner minus loser method that chooses top 10%/25% and bottom 10%/25% of anomalies in predicted data. And then, apply it to the real data. We also did window size = 12 months case.

→ 118 anomalies

		0	1	2	3	4	5	6	7	8	9	...
2011/01	0	1.215188	-1.048021	0.892622	0.272997	-1.129442	0.772301	-2.121192	-0.205771	1.177544	1.143759	...
2011/02	1	-0.720311	0.774981	0.856529	-1.674666	0.049734	0.323547	-1.897976	-2.145447	1.299806	-0.125764	...
2011/03	2	-0.260738	-1.135028	-0.306466	0.562326	0.399118	0.452634	-2.488044	-0.547396	-0.779630	0.987896	...
2011/04	3	-1.271842	-1.671114	0.622064	-1.173645	-2.110644	0.002207	-2.668441	-1.346239	1.233120	0.857585	...
.	4	-2.232063	-0.005992	0.374860	-1.629015	0.586636	-1.196059	-3.122563	-0.840958	0.460028	-0.785726	...
.	...	...	...	...	...	...	...	...	...	...	...	...
.	127	-1.291644	0.313031	-0.202873	-1.625271	1.976310	-0.090402	-1.919806	-1.623312	1.383723	0.172523	...
2021/09	128	-1.618121	1.270279	0.665616	-0.470819	0.589398	-1.544467	0.320897	-0.798086	0.046815	0.640145	...
2021/10	129	1.089907	-0.615519	0.183236	0.297655	-1.204493	-1.390528	-0.761146	0.016006	0.569612	-0.249674	...
2021/11	130	0.232263	-0.253987	-0.314123	-1.649233	2.023220	-1.064724	-0.216776	-3.161397	0.186290	1.163643	...
2021/12	131	-2.259713	0.700570	0.218412	-1.675514	1.856935	-2.048966	-1.308275	-1.842998	-0.454938	-0.067477	...

→ 118 anomalies

		0	1	2	3	4	5	6	7	8	9	...
2011/01	0	0.541110	0.439242	0.584586	0.465029	0.447901	0.437301	0.490426	0.575134	0.578125	0.566980	...
2011/02	1	0.490536	0.529014	0.568984	0.398067	0.600040	0.472884	0.447461	0.420607	0.567549	0.498904	...
2011/03	2	0.525153	0.452620	0.559949	0.475480	0.497582	0.431074	0.484504	0.488085	0.525991	0.670137	...
2011/04	3	0.484308	0.368452	0.581047	0.415660	0.460352	0.454762	0.457849	0.475854	0.579283	0.565918	...
.	4	0.493842	0.468521	0.588475	0.448383	0.562910	0.447169	0.455304	0.431290	0.548379	0.509892	...
.	...	...	...	...	...	...	...	...	...	...	...	...
.	127	0.441730	0.636653	0.485842	0.300041	0.647502	0.441208	0.355856	0.381852	0.663611	0.574371	...
2021/09	128	0.522594	0.639726	0.545996	0.394110	0.596393	0.419048	0.400860	0.440775	0.676017	0.445018	...
2021/10	129	0.520350	0.468515	0.556339	0.402729	0.477810	0.448537	0.463928	0.459421	0.598236	0.560289	...
2021/11	130	0.473300	0.577982	0.606014	0.443947	0.687052	0.469897	0.468404	0.357930	0.568919	0.628152	...
2021/12	131	0.497062	0.702268	0.621803	0.431372	0.709202	0.438084	0.451744	0.279858	0.560776	0.581201	...

Fig.6: RNN predict (window size = 6 months)

Fig.7: LR predict (window size = 6 months)

Fig.8 shows that RNN winner minus loser portfolio with 25% and window size = 12.

	High(12d,25%)	Low(12d,25%)	High-Low(12d,25%) \	High_ANOMALIES(12d,25%) \	LOW-ANOMALIES(12d,25%)
0	9.6668	-28.7714	38.4382	ol, cpq_12, epq_1, srev, resid6_6, eg_1, oca, …	dpia, tv_1, rev_1, noa, dnca, rev_6, roe_1, db…
1	21.1869	-37.7893	58.9762	epq_6, tbiq_12, sim_1, gpa, dtv_12, ebp, ilr_6…	eprd, tv_1, ivff_1, oca, dnca, cto, ivc, cei, …
2	30.8945	-1.7007	32.5952	sgq_1, ilr_6, r1n, resid6_6, dfin, eg_6, im_1,…	tv_1, dwc, ig2, rev_1, dnca, dtv_12, ep, ta, i…
3	13.3187	13.4789	-0.1602	cop, r5n, ilr_6, rev_6, im_12, resid6_6, dp, p…	tv_1, me, dtv_12, dpia, eprd, beta_1, ivff_1, …
4	17.5938	-10.8363	28.4301	r5n, dp, resid6_12, aci, ia, r5a, vhp, cpq_12,…	tv_1, beta_1, ivff_1, pta, r10n, srev, rev_12,…
..	…	…	…	…	…
127	41.3997	-37.0624	78.4621	droe_6, ta, rev_12, cpq_6, ope, oca, cim_1, r5…	tv_1, spq_6, ivff_1, eprd, bmq_12, nsi, beta_1…
128	-77.2391	8.8815	-86.1206	vhp, roe_1, spq_6, roe_6, r1n, eg_6, cim_1, r6…	beta_1, srev, ivc, dwc, ivg, dp, nsi, pda, noa…
129	-66.6138	84.7912	-151.4050	r1n, tbiq_6, ilr_1, r6_12, ebp, cpq_12, tbiq_1…	rev_6, aci, p52w_12, em, srev, ta, ir, resid11…
130	136.4989	-40.1194	176.6183	eg_1, ile_1, resid6_12, ato, r5a, droe_1, etr,…	eprd, ivff_1, spq_1, dur, ta, bmq_12, nsi, spq…
131	55.5248	-40.2059	95.7307	cim_12, r6_12, cpq_12, cop, r15a, r5a, eg_6, a…	beta_1, srev, eprd, nsi, ivff_1, cei, dp, tv_1…

Fig.8: Winner minus loser portfolio result. (RNN, window size = 12 months, High/Low 25%)

Fig.9 and *Fig.10* show that top 10 most frequently anomalies in each model. The 'momentum' factors are most frequently selected in winner portfolio. However, 'intangibles' and 'frictions' factors are most frequently selected in loser portfolio. It means that 'momentum' factors have high probability that it will make a positive return and 'intangibles' and 'frictions' factors have high probability that it will make a negative return.

High_ANOMALIES (6d,10%)	Low_ANOMALIES (6d,10%)	High_ANOMALIES (6d,25%)	Low_ANOMALIES (6d,25%)	High_ANOMALIES (12d,10%)	Low_ANOMALIES (12d,10%)	Low_ANOMALIES (12d,25%)	Low_ANOMALIES (12d,25%)
cim_1	eprd	cim_1	eprd	r11_1	eprd	cim_1	eprd
r11_1	ivff_1	eg_1	cei	cim_1	tv_1	eg_1	cei
r11_6	tv_1	r11_1	pda	sim_1	ivff_1	r11_1	noa
p52w_12	beta_1	r11_6	ivff_1	r6_6	beta_1	r5a	dwc
eg_1	em	r5a	oa	eg_1	nsi	sim_1	poa
sim_1	nsi	sim_1	poa	r11_6	srev	r11_6	nsi
r1n	dnoa	r15a	nsi	p52w_12	dwc	p52w_12	dfnl
im_1	me	droe_1	pta	p52w_6	em	resid6_6	tv_1
r6_6	noa	r6_6	dwc	roe_1	dbe	r15a	pta
spq_1	cei	epq_1	ivc	cpq_1	me	resid6_12	dcoa

Fig.9: Top 10 frequently anomalies in RNN

High_ANOMALIES (6d,10%)	Low_ANOMALIES (6d,10%)	High_ANOMALIES (6d,25%)	Low_ANOMALIES (6d,25%)	High_ANOMALIES (12d,10%)	Low_ANOMALIES (12d,10%)	Low_ANOMALIES (12d,25%)	Low_ANOMALIES (12d,25%)
r11_1	dwc	r11_1	dii	r11_1	dwc	r11_1	dii
r6_6	pda	r6_6	dwc	r6_6	dii	resid6_12	dwc
eg_1	dii	resid6_12	dnoa	resid6_12	poa	r6_6	poa
r11_6	nsi	eg_1	nsi	eg_1	nsi	eg_1	nsi
resid6_12	cei	r6_12	ig	r11_6	pda	r6_12	pda
r6_1	ig	r11_6	cei	r6_12	ig	r11_6	dcoa
droe_1	eprd	resid11_6	pda	eg_6	noa	r6_1	dnoa
r6_12	noa	r6_1	pta	droe_1	cei	resid11_6	ivff_1
resid11_6	dcoa	resid6_6	poa	r6_1	pta	droe_1	cei
p52w_6	ivff_1	droe_1	dac	r5a	eprd	resid6_6	pta

Fig.10: Top 10 frequently anomalies in LR

Fig.11 shows that the final monthly anomaly portfolio return. (Unit: %) RNN,6m,10% is best model as 0.63% monthly return. However, in 25% model, LR is better than RNN.

Also, we calculate the Sharpe ratio. (*Fig.12)*

Type	LR	**RNN**
High-Low(6m,10%)	0.52	0.63
High-Low(6m,25%)	0.55	0.40
High-Low(12m,10%)	0.52	0.72
High-Low(12m,25%)	0.50	0.44

Fig.11: Final monthly anomaly portfolio return.

Type	LR	**RNN**
High-Low(6m,10%)	0.12	0.13
High-Low(6m,25%)	0.10	0.11
High-Low(12m,10%)	0.10	0.16
High-Low(12m,25%)	0.09	0.14

Fig.12: Sharpe ratio of each model.

Whole Sharpe ratios are small. Return of our model is quite good. But standard deviation of returns is quite large. The RNN is better than LR in the Sharpe ratio.

In conclusion, RNN model has better performance than LR in the winner minus loser portfolio return and sharp ratio. The reason for that is RNN is store the sequential information and use nonlinear dynamics, but LR is not. *The conclusion details (formula, etc.) in code file.

V. Literature reference

[1]: Juhani T. Linnainmaa Michael R. Roberts, "THE HISTORY OF THE CROSS SECTION OF STOCK RETURNS" - *NATIONAL BUREAU OF ECONOMIC RESEARCH*, 2016

[2]: Shihao Gu, Bryan Kelly, Dacheng Xiu, "Empirical Asset Pricing via Machine Learning", *The Review of Financial Studies*, 2020

[3]: Anomaly information: https://global-q.org/testingportfolios.html