Which anomaly portfolio will make money in the next month?

20205097 Kim Ji Hyun
20192033 Aida Kurmangali
20201181 Oh Ji Hwan

TABLE OF CONTENTS

01 PROJECT MOTIVATION

02 DATA ANALYSIS

03 MACHINE LEARNING MODEL

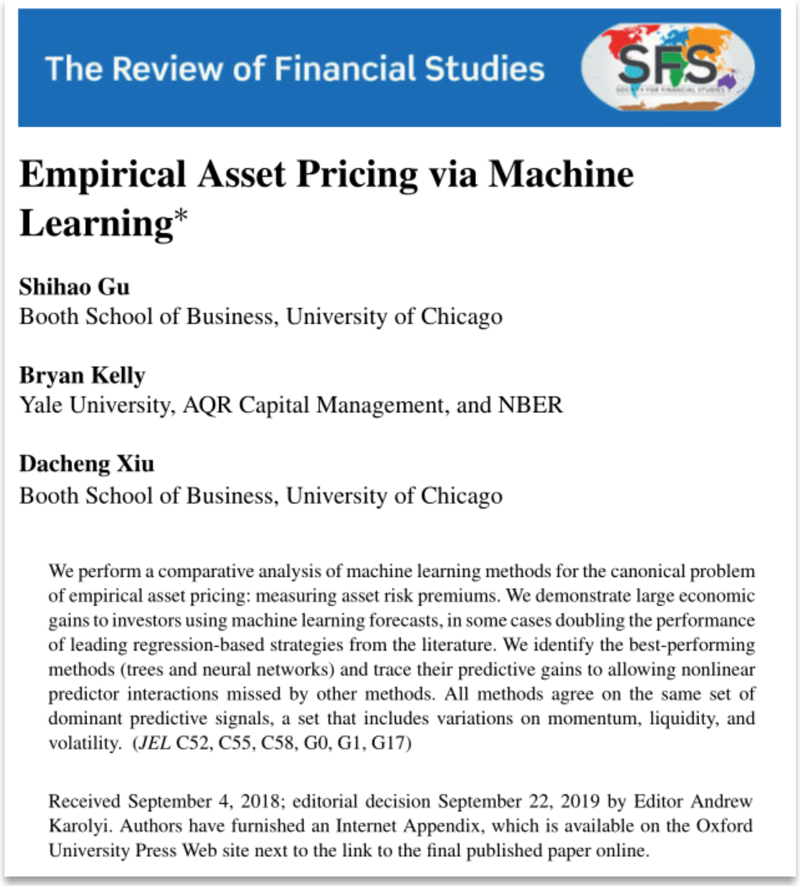
04 PORTFOLIO CREATION

01 PROJECT MOTIVATION

Why anomaly portfolio data ?

- Financial data release date is not fixed.
- Anomalies can be provided daily or yearly.
- In our anomaly portfolio return data, data processing about releasing problem has already been processed.
- Using anomalies result is more realistic.

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



The Review of Financial Studies

Empirical Asset Pricing via Machine Learning*

Shihao Gu
Booth School of Business, University of Chicago

Bryan Kelly
Yale University, AQR Capital Management, and NBER

Dacheng Xiu
Booth School of Business, University of Chicago

We perform a comparative analysis of machine learning methods for the canonical problem of empirical asset pricing: measuring asset risk premiums. We demonstrate large economic gains to investors using machine learning forecasts, in some cases doubling the performance of leading regression-based strategies from the literature. We identify the best-performing methods (trees and neural networks) and trace their predictive gains to allowing nonlinear predictor interactions missed by other methods. All methods agree on the same set of dominant predictive signals, a set that includes variations on momentum, liquidity, and volatility. (*JEL* C52, C55, C58, G0, G1, G17)

Received September 4, 2018; editorial decision September 22, 2019 by Editor Andrew Karolyi. Authors have furnished an Internet Appendix, which is available on the Oxford University Press Web site next to the link to the final published paper online.

"

The bottom-up S&P 500 forecast from the generalized linear model, in contrast, delivers an ***R2 of 0.71%.***

Trees and neural networks improve upon this further, generating monthly out of sample ***R2's between 1.08% to 1.80%*** per month.

"

02

DATA ANALYSIS

Data preprocessing

“Monthly Anomaly Portfolio Return Data”

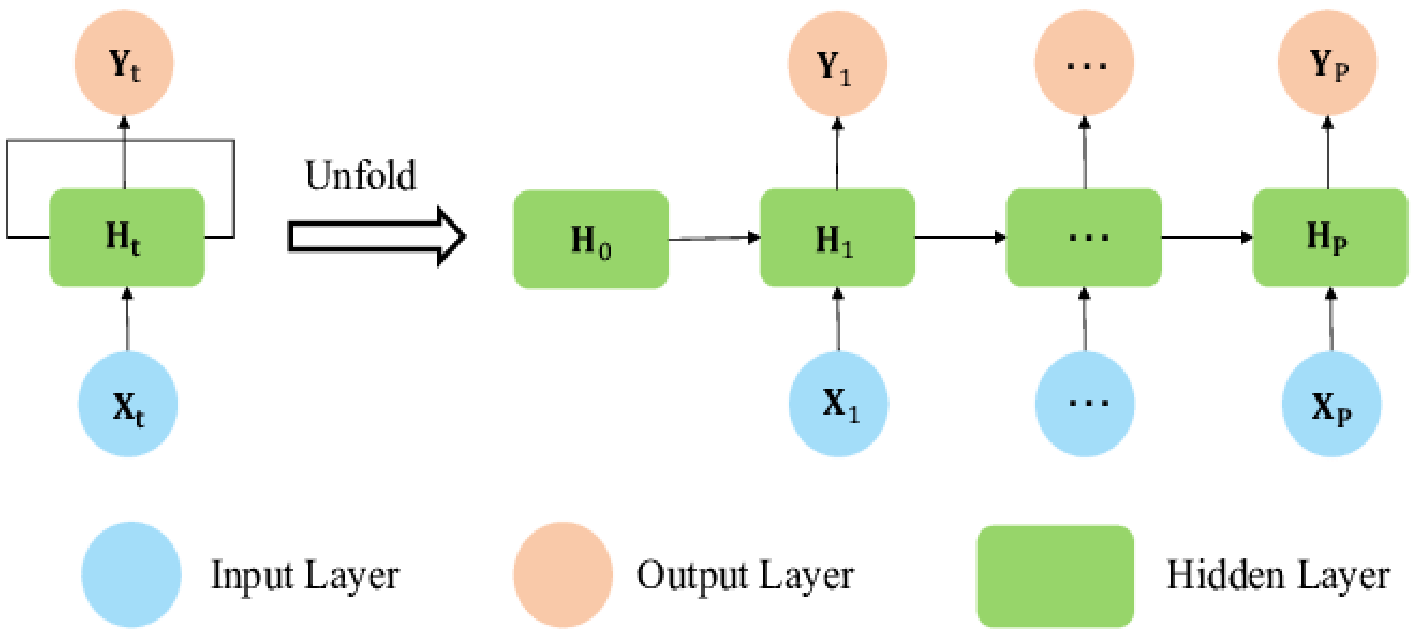
DATE	beta_1	dtv_12	isff_1	ivff_1	me	srev	tv_1	eprd	etl	etr	...	epq_12
1967-01	11.8004	-4.4299	3.5865	12.1847	-13.0562	-0.1126	11.4709	8.6986	3.0873	-0.3883	...	9.4271
1967-02	2.1382	-2.7018	-1.2576	4.7215	-5.1638	2.8017	4.7486	-5.0862	-3.8491	0.6776	...	-3.6533
1967-03	0.2358	-1.6275	4.8673	0.6764	-3.3238	-1.5156	0.3803	-0.3200	-0.2502	-3.9915	...	1.8400
1967-04	3.0167	0.5737	-3.6645	-3.0035	-0.7099	-1.9171	-0.8516	-2.8200	1.0904	-1.7424	...	-2.6575
1967-05	1.3046	-6.4407	-1.2705	3.4395	-4.8868	-4.0637	0.6068	1.9194	1.5000	2.3227	...	3.3314
...	...	...	...	...	...	...	...	...	...	...	...	...
2021-08	0.7348	1.3942	-1.4633	-1.3650	0.6421	2.9832	-0.4001	-1.1960	-0.6989	-0.7664	...	-2.2072
2021-09	9.5032	-2.3366	1.9337	1.0624	-1.4418	-1.7750	1.1031	2.1055	2.4773	-1.1027	...	0.3463
2021-10	1.7722	5.5664	0.9665	-2.3577	8.4341	2.6571	0.5204	-2.6409	-0.7454	-2.7059	...	-6.8053
2021-11	-4.0661	4.6815	5.1696	-5.6675	7.9488	8.6912	-3.1802	-10.0804	-1.2931	-5.3710	...	0.7586
2021-12	-10.1847	1.9730	-2.0947	-15.6574	7.3379	5.0166	-19.7398	-2.3719	1.8116	5.4579	...	7.8395

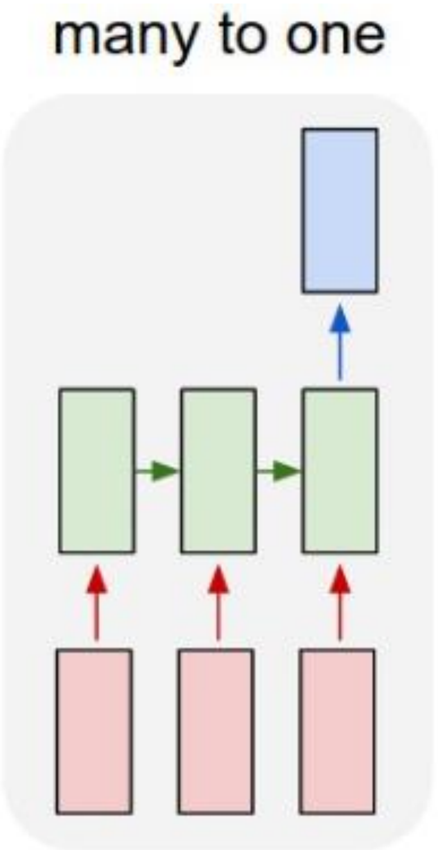
- Drop columns(=anomalies) which contain NaN value -> 118 anomalies
- Period : 1967/01 – 2021/12

03

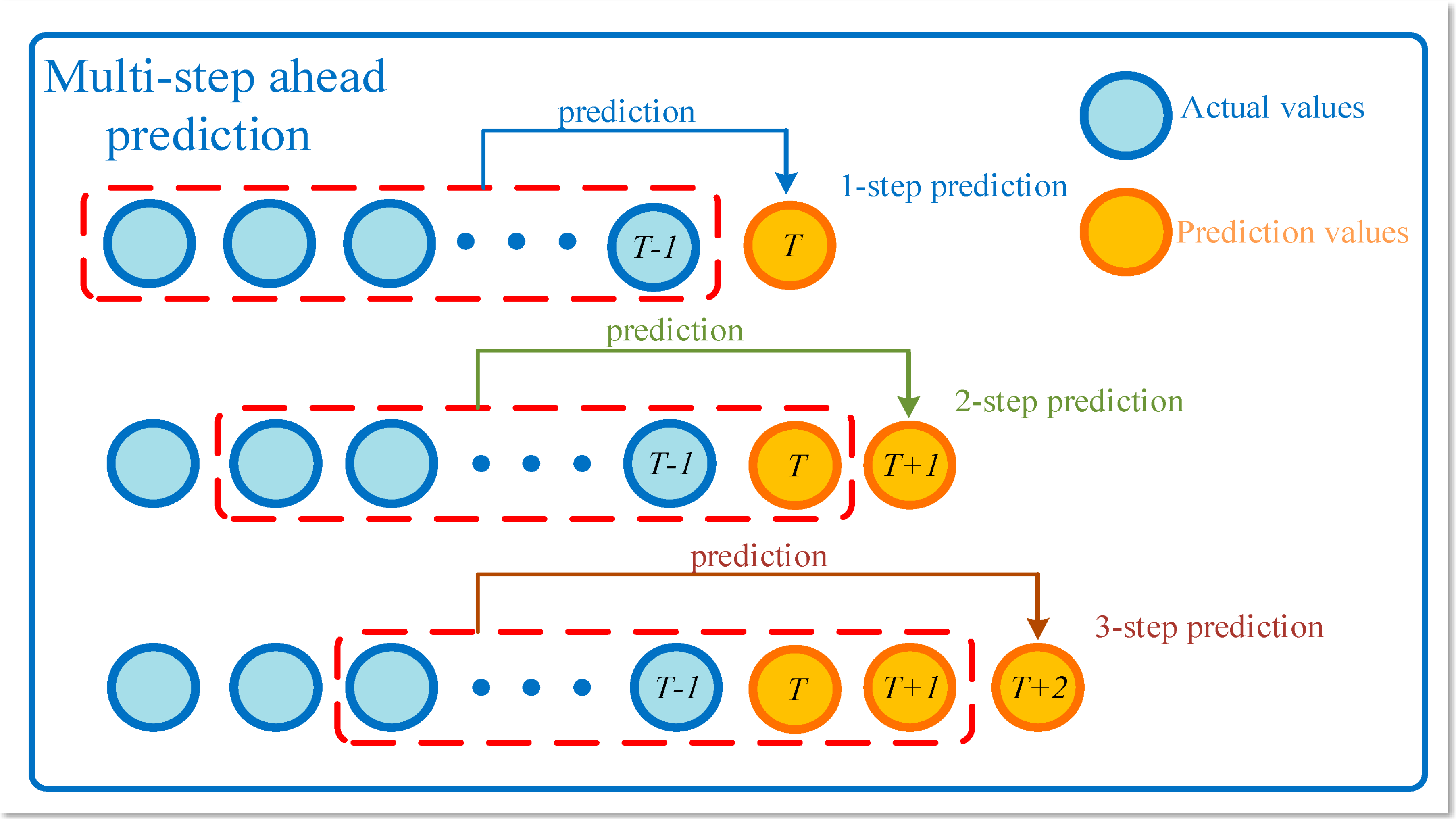
MACHINE LEARNING MODEL

RNN : Recurrent Neural Networks





RNN : Recurrent Neural Networks



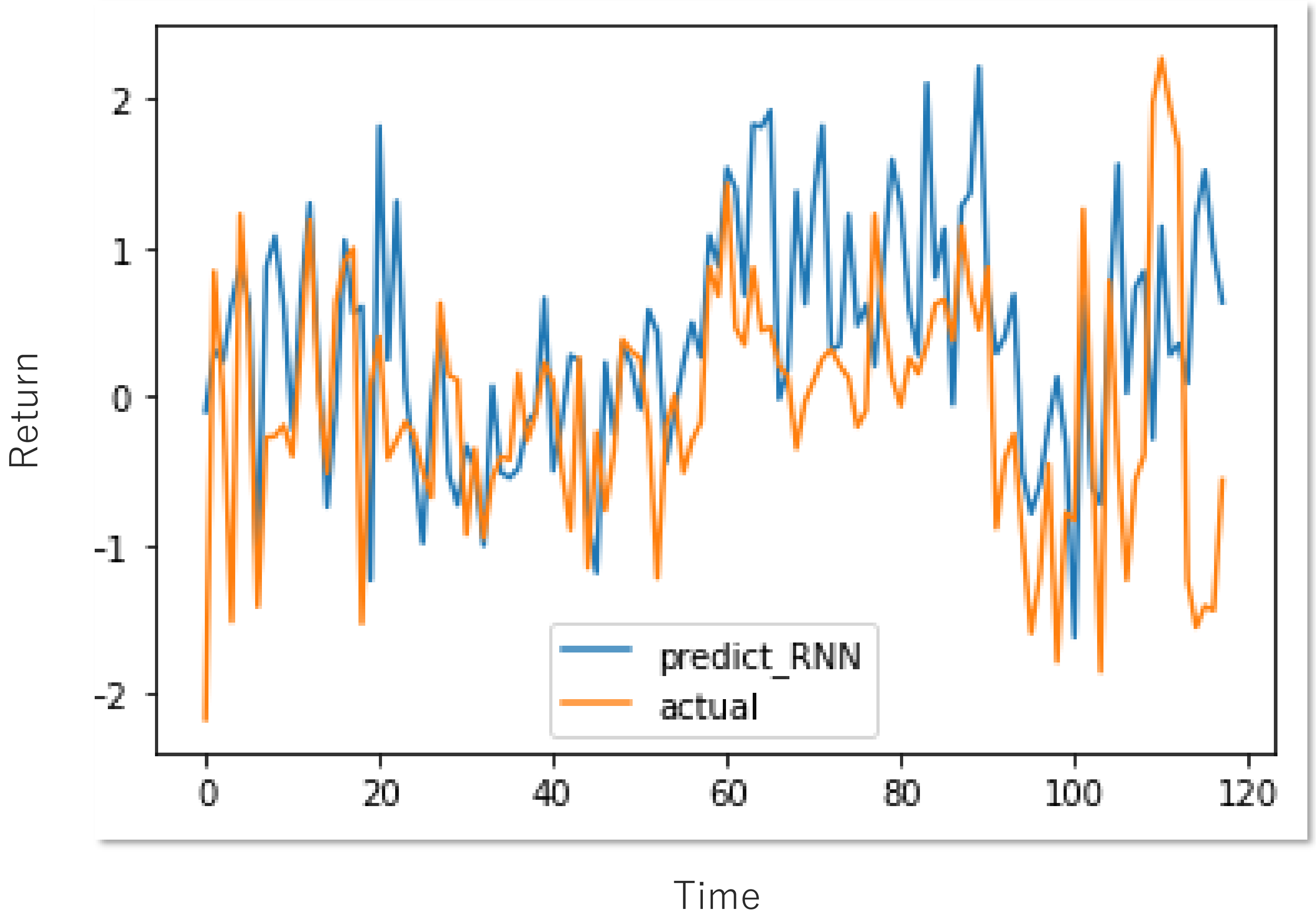
Predicted return : result of RNN with window size 6-month

→ 118 anomalies

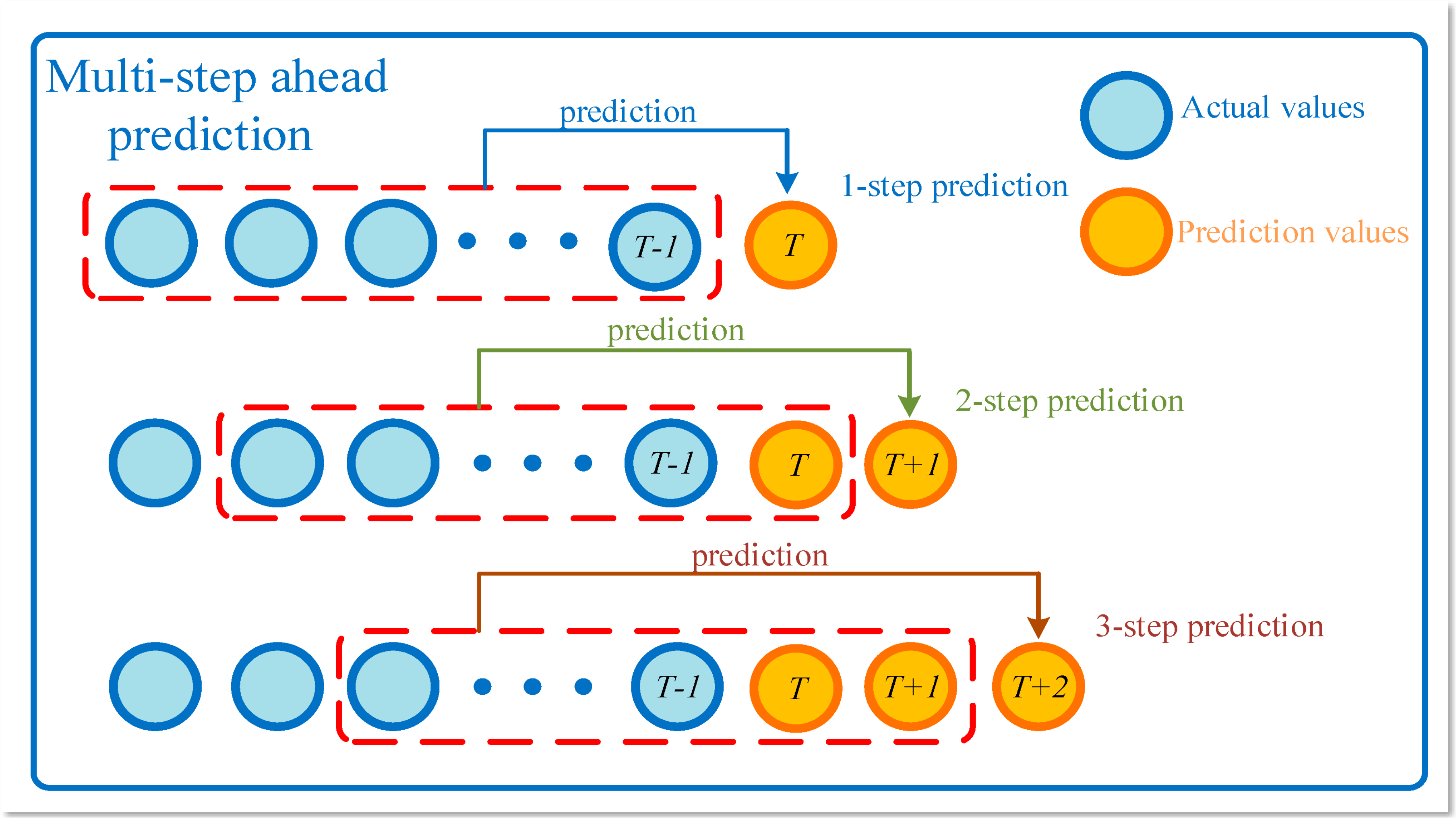
		0	1	2	3	4	5	6	7	8	9	...
2011/01	**0**	0.921818	-1.555905	-0.216784	0.507426	-1.715958	0.582059	-0.006161	1.060925	0.231481	0.815064	...
2011/02	**1**	-0.250684	0.022891	0.995024	-1.657771	0.891085	0.412480	-0.531858	-0.723415	0.759400	-0.467554	...
2011/03	**2**	-0.791201	-0.662408	-0.222583	0.709380	-0.135113	1.072130	-0.710599	-0.004273	-0.075016	0.334186	...
2011/04	**3**	-1.643287	-1.641460	0.188515	-2.331415	-2.341688	-0.458757	-0.377641	-0.404866	0.893355	0.859169	...
.	**4**	-2.105942	-0.349189	0.514675	-1.685540	0.383830	-0.700145	-0.506371	-1.252297	0.667818	0.385201	...
.	**...**	...	...	...	...	...	...	...	...	...	...	...
.	**127**	-1.371055	0.261436	-0.084163	-1.039234	0.891655	0.729431	-1.828175	-1.033825	0.888507	0.737908	...
2021/09	**128**	-1.629615	0.487816	0.331246	0.283571	0.188426	-1.279548	-0.271830	-0.385659	0.319497	0.553288	...
2021/10	**129**	1.106156	-0.065281	0.673837	0.846263	-1.616605	-0.453245	-0.207154	0.163164	0.232559	-0.248487	...
2021/11	**130**	-0.347369	0.158365	0.088639	-1.096070	0.956352	-1.108885	0.259125	-2.059382	0.343916	0.585333	...
2021/12	**131**	-2.680436	0.692534	-0.027551	-1.484989	0.894401	-1.364572	-1.067600	-2.184001	0.254836	0.588070	...

132 rows × 118 columns

Comparison of Predicted and Actual data



RNN : Recurrent Neural Networks



Predicted return : result of RNN with window size 12-month

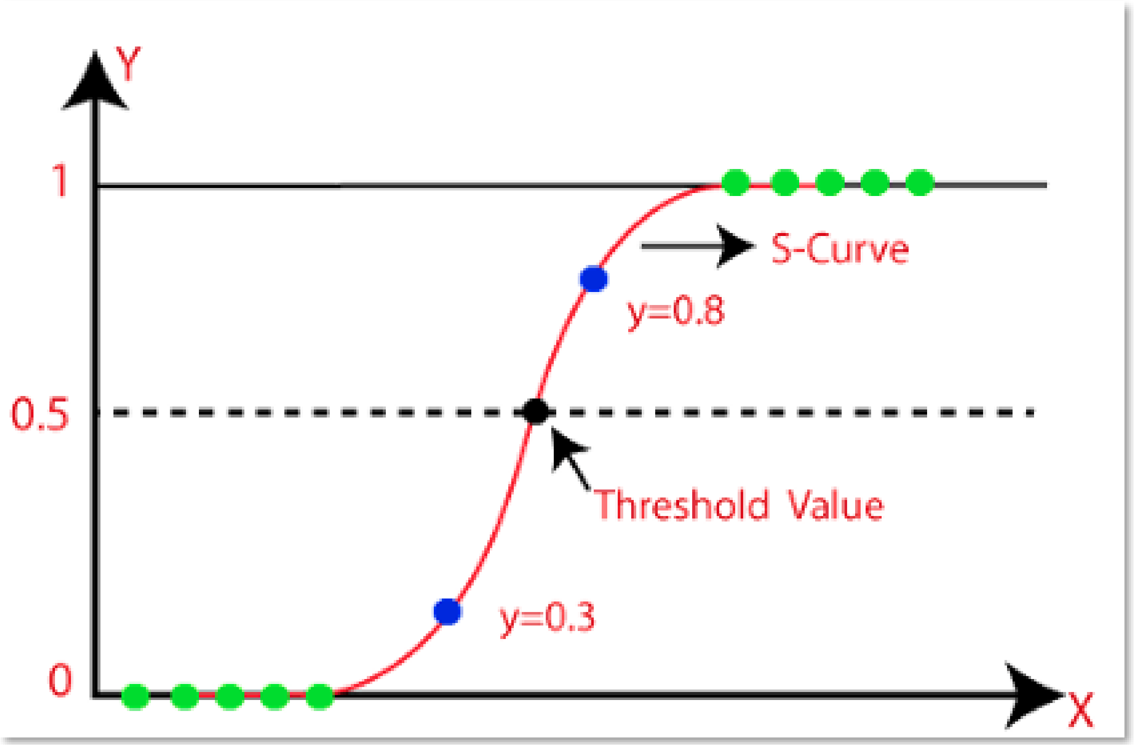
→ 118 anomalies

		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	...

Logistic Regression

$$Where\ f(*) = Logistic\ Regression$$

$$\mathrm{f}(x) = X_T\ , x = X_{T-1}, X_{T-2}, X_{T-3}, \dots X_{T-6}\ or\ X_{T-12}$$



Probability from RL classification

- High probability (close to 1) means that price will increase more.
- If probability close to 0.5 , then maintain the price.
- Low probability(close to 0) means that price will decrease more.

Predicted return : result of LR with window size 6-month

→ 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	...

Predicted return : result of LR with window size 12-month

→ 118 anomalies

		0	1	2	3	4	5	6	7	8	9	...
2011/01	0	0.593016	0.410367	0.647445	0.462422	0.525160	0.451730	0.529627	0.586910	0.594515	0.495245	...
2011/02	1	0.553616	0.497347	0.586675	0.426943	0.641920	0.446779	0.501142	0.451371	0.572068	0.482946	...
2011/03	2	0.582781	0.415391	0.531446	0.510588	0.513813	0.439949	0.493140	0.519602	0.538403	0.698491	...
2011/04	3	0.397974	0.301741	0.596467	0.412630	0.416274	0.418672	0.460874	0.480691	0.593048	0.652117	...
.	4	0.466454	0.486399	0.555478	0.417640	0.526894	0.438318	0.395907	0.356765	0.524760	0.462064	...
.	...	...	...	...	...	...	...	...	...	...	...	...
.	127	0.411250	0.706996	0.427734	0.393667	0.559281	0.436651	0.421064	0.348586	0.719910	0.483272	...
2021/09	128	0.576371	0.437294	0.552878	0.406885	0.274237	0.380133	0.380126	0.383780	0.672458	0.463518	...
2021/10	129	0.679825	0.387306	0.452063	0.480228	0.385244	0.442396	0.586010	0.440676	0.595335	0.499405	...
2021/11	130	0.442842	0.358929	0.639024	0.506307	0.616988	0.440669	0.502177	0.358616	0.568646	0.644830	...
2021/12	131	0.583398	0.666747	0.695378	0.519382	0.747673	0.421789	0.540006	0.284813	0.551877	0.663811	...

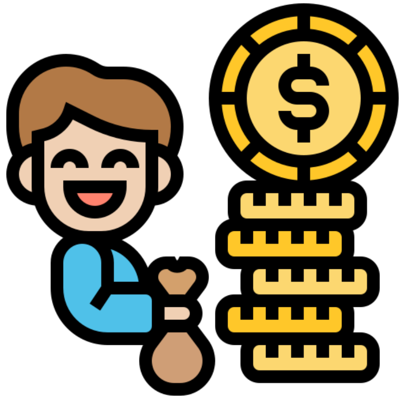
04

PORTFOLIO CREATION

Generating Portfolio

- Using the predicted values, select the high return anomalies list and the low return anomalies list.
  - We have two options. Select the **10% or 25% anomalies for each of group**
- Apply the High and Low group anomalies list to the real anomaly portfolio return data
- Add all the actual returns of each group, and calculate High group return – Low group return

> **"Our portfolio return is**
> **High anomaly group – Low anomaly group"**



Portfolio result on High – Low (Winner – Loser) strategy with RNN

	High(6d,10%)	Low(6d,10%)	High-Low (6d,10%)	High_ANOMALIES(6d,10%)	LOW-ANOMALIES(6d,10%)
2011/01	-3.1463	-5.8132	2.6669	droe_1, eg_6, im_6, bmq_12, cpq_1, sp, sim_1, oca, spq_1, spq_6, spq_12	noa, me, dbe, rev_12, dtv_12, rev_6, opa, ope, dnoa, pta, ivc
2011/02	7.8925	-14.1867	**22.0792**	vhp, p52w_12, r11_12, r6_1, r15a, ioca, im_1, r6_6, eg_1, r1n, sim_1	ivff_1, cei, ope, r10n, dnco, ig, dpia, oca, eprd, pda, oa
2011/03	22.3391	3.6138	18.7253	oca, resid6_6, cpq_1, r1n, r11_6, r6_6, cim_1, sim_1, p52w_12, r11_12, r11_1	bm, rev_1, pta, ia, ig2, vhp, ig, rev_12, cla, ivc, dnoa
2011/04	4.0625	4.2094	-0.1469	im_12, im_1, r11_12, cla, r6_1, cpq_12, r5a, spq_12, dp, p52w_12, cim_1	me, ivff_1, beta_1, dtv_12, dnoa, r5n, dpia, dlno, dnca, ir, dii
2011/05	-6.0234	-5.7864	-0.237	r11_6, cpq_1, r10a, aci, r1n, epq_1, r15a, r5a, r11_1, p52w_12, ol	pta, beta_1, r10n, ivff_1, eprd, bm, srev, dnoa, pda, ta, ig2

We can get return **22%** at 2011/02 through selected anomalies by RNN

Portfolio result on High – Low (Winner – Loser) strategy with LR

	High(6d,10%)	Low(6d,10%)	High-Low (6d,10%)	High_ANOMALIES(6d,10%)	LOW-ANOMALIES(6d,10%)
2011/01	-18.0374	-10.521	-7.5164	r1a, im_12, r11_1, resid6_12, droe_6, resid6_6, droe_12, r6_12, droe_1, r11_6, ile_1	dpia, noa, rev_12, rev_6, ia, pta, poa, dnoa, dnco, dcoa, dbe
2011/02	17.0352	-9.0544	26.0896	resid11_6, r6_1, eg_1, im_6, resid6_6, ilr_12, r6_6, droe_1, r11_1, resid6_12, r15a	pda, nsi, cei, ivff_1, ig, dii, eprd, dfnl, ir, poa, cto
2011/03	16.8504	7.4983	9.3521	sue_6, resid6_12, eg_1, ilr_12, r6_12, r6_6, im_6, r11_6, etr, im_12, r11_1	dwc, ig2, dii, ia, dnco, dac, poa, ir, ig, nsi, dnca
2011/04	2.502	0.434	2.068	dfin, r10a, r6_12, r11_6, droe_1, ilr_6, eg_1, resid6_12, r6_6, resid11_6, r11_1	dtv_12, dnco, noa, dlno, dnca, ir, ivff_1, ep, cei, dac, dpia
2011/05	1.0387	6.2066	-5.1679	im_6, r11_1, cto, r6_12, ope, resid6_6, r15a, resid6_12, eg_1, r6_1, resid11_6	nsi, dwc, pda, dnca, dnoa, dii, cei, ia, ig2, pta, noa

We can get return **-5%** at 2011/05 through selected anomalies by Logistic Regression

The number of positive return on our portfolio

Type	Success of RNN	Success of LR
High-Low(6d,10%)	55%	55%
High-Low(6d,25%)	54%	58%
High-Low(12d,10%)	57%	54%
High-Low(12d,25%)	54%	58%

Average monthly return on Winner minus Loser portfolio

Type	RNN	Return on LR
High-Low(6m,10%)	5.72	6.93
High-Low(6m,25%)	**15.78**	11.47
High-Low(12m,10%)	5.69	7.89
High-Low(12m,25%)	14.63	12.79

Most frequently selected anomalies on Logistic Regression

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

Most frequently selected anomalies on 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%)
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