Quartlery compustat firm characteristics (QUARTERLY)

* Backward fill to trade at a gigher frequency

Jensen,Kelly and Pedersen JKP factors:

* Monthly frequency

GENERAL IDEA:

TWO STAGE APPROACH:

1. Monthly backbone

* Compustat
* Jensen Kelly and Pedersen
* Chen zimmer

1. Daily analysis to tilt from my original monthly position

* Daily futures
* 8K 10K
* Earning call

**10k ideas:**

Une image contenant texte, capture d’écran, Police

Le contenu généré par l’IA peut être incorrect.

Elements en vrac

1. Changer les csv en parquets?

Existing studies :

Zhang (2022) – Deep Architecture Benchmarking

In particular, **LSTM networks with memory and attention mechanisms achieved the highest predictive accuracy** and investment performance,

WHAT’S LSTM:

**The Expected Returns on Machine-Learning Strategies Vitor Azevedoa,∗ , Christopher Hoegnerb , Mihail Velikovc:**

<https://afajof.org/management/viewp.php?n=75544#:~:text=up%20to%201.63,returns%20are%20also%20accompanied%20by>

“We follow common practice and include only common equity stocks (CRSP share code 10 or 11)” :

 **Share code 10** = U.S.‐listed common shares

 **Share code 11** = American Depositary Receipts (ADRs), which CRSP treats as common‐equity equivalents

The anomaly signals have varying ranges of values over which they are defined, making it more difficult for neural networks to estimate suitable parameters during training (Singh and Singh, 2020). Consequently, we follow the current literature by percent-ranking all anomaly features into the same range [-1;1] (Kelly et al., 2019; Freyberger et al., 2020; Gu et al., 2020). Missing values are replaced with 0.

one-hot encoding to ensure that our studied models do not suffer multi-collinearity issues

**Enhancing Stock Market Anomalies with Machine Learning**

//https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=3604626

Thereby, we follow the assumptions of Chen and Zimmermann (2020), namely applying a lag of six and three months for annual, respectively, quarterly accounting data.8

While we do not oppose any strict filters for prices or market cap during the data gathering process, we follow Griffin et al. (2010) by including only common equity (i.e., stocks with a WRDS share code of 10, 11, or 12) and excluding any stock that is not listed at the U.S. exchanges NYSE, NASDAQ or AMEX.

We calculate the anomaly set for every firm-month observation available, ranging from 1945 to 2019. However, our main analysis focuses only on the period from 1979 to 2019 (492 months of observations) to reduce the number of missing values in the training set while simultaneously ensuring a large enough and diverse dataset to find profitable patterns. Particularly, analyst recommendations and quarterly-based fundamental data often do not match our quality and quantity requirements before 1979.

The stock-characteristic portfolio-sort approach is among the most common and dominant instruments to measure an anomalies’ potential profitability and determine its statistical significance

V.B. Reducing the high-dimensionality of the factor zoo with unsupervised learning and feature reduction algorithms

A sophisticated reduction or combination of features into a lowerdimensional dataset could filter out unnecessary noise, further improving our algorithms’ performance.

Autoencoders are a special case of Convolutional Neural Networks. -> Complex method for dimensionality reduction.

The lasso regression and elastic net selection follow common practice. The theory-derived selection of anomalies uses only past anomalies with t-statistics above 1.96 and 3, aiming to reduce the noise of non-important and insignificant signa

In total, our approaches indicate that feature reduction might be less potent in the context of anomalies than suggested. While reducing the noise and dimensionality of the dataset, feature reduction might also weaken or eliminate significant signals, decreasing the model’s overall performance. Since many machine learning models have in-built capabilities to handle high-dimensional datasets, feature reduction might be less critical than in classical regression approaches of former literature.

While more training data is generally positively correlated with a model’s performance and ability to generalize, the relevance of older observations decreases due to the non-stationary character of the financial time series. Thus, a shorter time frame with less but more relevant data might increase performance.

It seems that the algorithms can handle the high dimensionality directly by themselves, and any pre-processing reduction methods weaken essential signals.

Une image contenant texte, capture d’écran, Police, conception

Le contenu généré par l’IA peut être incorrect.

Find the right sequence length:

(determines *how many past time‐steps* the model sees in a single forward pass)

Enough history to learn dynaics that matter but not so much ta we force the model to remember old information useless

//Empirical validation: grid or random search over a small set of candidate / times series cross validation

Strict rolling windows:

How to define the good size?

Expanfing windows but decay?

Function of the decay

Retraining every month?

On everything or smaller epoch

--------------------------------------------------------------------------------------------------------------

!!!!

LAGING RULE: CHEN AND ZIMMEMRAN:

 **“Annual accounting data becomes available only 6 months after fiscal‐year‐end.”**

 **“Quarterly accounting data becomes available only 3 months after quarter‐end.”**

**To handle the CZ data:**

Une image contenant texte, capture d’écran, Police, nombre

Le contenu généré par l’IA peut être incorrect.

signaldoc.csv

-> cat.data column

SEEMS THAT THIS IS ALREADY LAGGED TO TAKE INTO ACCOUNT THIS EFFECT

OTHERWISE THE CATEGORIE OF THE ANOMALY IS AVAILABLE ON THE CHE GITHUB ( SEE FAVORIS GOOGLE)

https://github.com/OpenSourceAP/CrossSection/blob/master/SignalDoc.csv

<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4491887&download=yes>

Furthermore, we include eight macroeconomic predictors from Welch and Goyal (2008): dividend-price ratio (DP), earnings-price ratio (EP), book-to-market ratio (BM), net equity expansion (NTIS), Treasury-bill rate (TBL), term spread (TMS), default spread (DFY), and stock variance (SVAR).6

<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4281769#:~:text=Portfolios%20formed%20on%20a%20time,We%20also%20show%20that>

parle de CZ ET JKP

Une image contenant texte, capture d’écran, Police

Le contenu généré par l’IA peut être incorrect.

En 2 etapes?

1. Etapes selection de features avec un lasso
2. Implementation du reseau de neurone

Rf**=**conn**.**raw\_sql("""select mcaldt,tmytm

from crsp.tfz\_mth\_rf

where kytreasnox = 2000001

and mcaldt>='2002-01-01'

and mcaldt<='2024-12-31'""", date\_cols**=**['mcaldt'])

Rf['tmytm']**=**Rf['tmytm']**/**12**/**100

Rf**=**Rf**.**rename(columns**=**{ "mcaldt": "date","tmytm": "1m T-Bill"})