Portfolio Risk Forecast

for TD Asset Management

11th Montreal
Industrial Problem Solving
Workshop

August 27,
2021

Organized by

CENTRE DE RECHERCHES MATHÉMATIQUES

IVADO

Fin ML

The Team

Professors & coordinators

René Garcia
Full Professor,
Université de Montréal

Jean Masson
Managing Director,
TD Asset Management

Ruslan Goyenko
Associate Professor,
McGill University

Jean-François Fortin
Vice-President,
TD Asset Management

Manuel Morales
Associate Professor
& Fin-ML Program Director,
Université de Montréal

Rheia Khalaf
Director, Collaborative
Research & Partnerships,
Fin-ML / IVADO

Avinash Srikanta Prasad
Postdoctoral Fellow,
University of Waterloo

Foulemata Tirera
M. Sc.,
Polytechnique de Montréal

Participants

Yaroslav Babich
Undergraduate,
Mathematics
Université Laval

Kiran Deol
Undergraduate,
University of Alberta

Mohamed Gueye
Master's Student,
Mathematics and Statistics,
Université de Montréal

Qi Guo
Ph.D. Student,
University of Calgary

Thierry Jean
Master's Student,
Computational Medecine,
Université de Montréal

Ehsan Rezaei
Ph.D. Student,
Polytechnique de Montréal

Javad Roustaei
Master's Student,
Financial Mathematics,
Concordia University

Myles Sjogren
Master's Student,
Mathematics and Statistics
University of Calgary

Ernest Tafolong
Master's Student,
Decision Science,
HEC Montréal

Shiva Zokaee
Master's Student,
Industrial Engineering,
GERAD, Université de Montréal

participants are sorted alphabetically

Problem Breakdown

Thierry Jean, Myles Sjogren

1

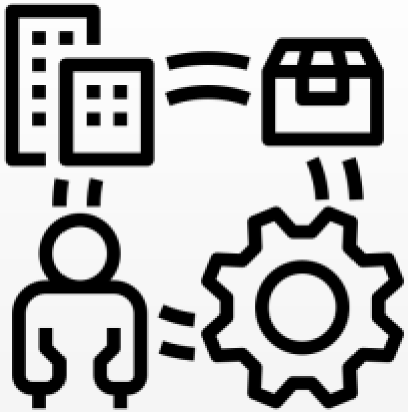
participants are sorted alphabetically

Problem statement

“ **How can we better predict risk and future return ?**

Fundamental factors

describe the underlying financials, such as earnings, market capitalization, and debt levels.



source: https://www.investopedia.com/terms/m/multifactor-model.asp

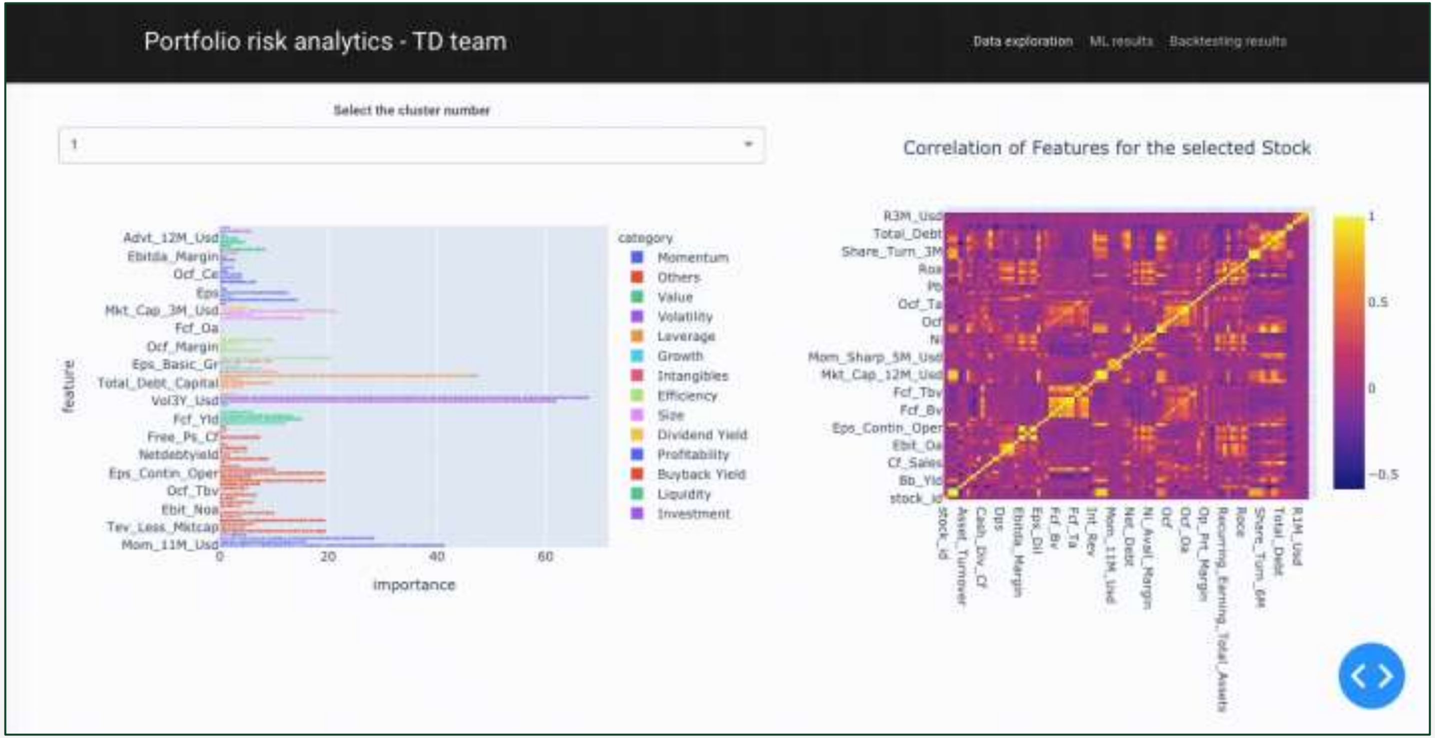
Historical Stock Data

- Tabular data indexed by *stock_id* and *date*
- 1207 individual stocks
- 245 months (from 11-1998 to 03-2019)
- 93 features about fundamental factors
- 4 labels about future/forward total return over 1, 3, 6, and 12 months

stock_id	date	Advt_3M_Usd	Advt_6M_Usd	…	R12M_Usd
13	2006-12-31	0.33	0.27	…	-0.041
13	2007-01-31	0.32	0.28	…	-0.253
13	2007-02-28	0.3	0.3	…	-0.366
17	2015-03-31	0.64	0.7	…	-0.376
17	2015-04-30	0.62	0.66	…	-0.113
17	2015-05-31	0.63	0.64	…	-0.194
17	2015-06-30	0.62	0.63	…	0.309
17	2015-07-31	0.56	0.6	…	2.139
17	2015-08-31	0.47	0.57	…	0.436
17	2015-09-30	0.41	0.54	…	1.398
17	2015-10-31	0.36	0.48	…	1
17	2015-11-30	0.37	0.43	…	2.933
17	2015-12-31	0.32	0.37	…	4.323
17	2016-01-31	0.27	0.32	…	4.447
17	2016-02-29	0.21	0.3	…	4.857
17	2016-03-31	0.31	0.32	…	1.737
17	2016-04-30	0.45	0.38	…	0.275
17	2016-05-31	0.55	0.42	…	0.376
17	2016-06-30	0.61	0.5	…	0.22
17	2016-07-31	0.65	0.58	…	-0.024
17	2016-08-31	0.7	0.64	…	0.467
17	2016-09-30	0.7	0.66	…	0.222
17	2016-10-31	0.66	0.66	…	0.08
17	2016-11-30	0.67	0.69	…	-0.244
17	2016-12-31	0.73	0.71	…	-0.143
17	2017-01-31	0.76	0.72	…	-0.219
17	2017-02-28	0.86	0.8	…	-0.341
17	2017-03-31	0.86	0.81	…	-0.153
17	2017-04-30	0.87	0.82	…	0.104

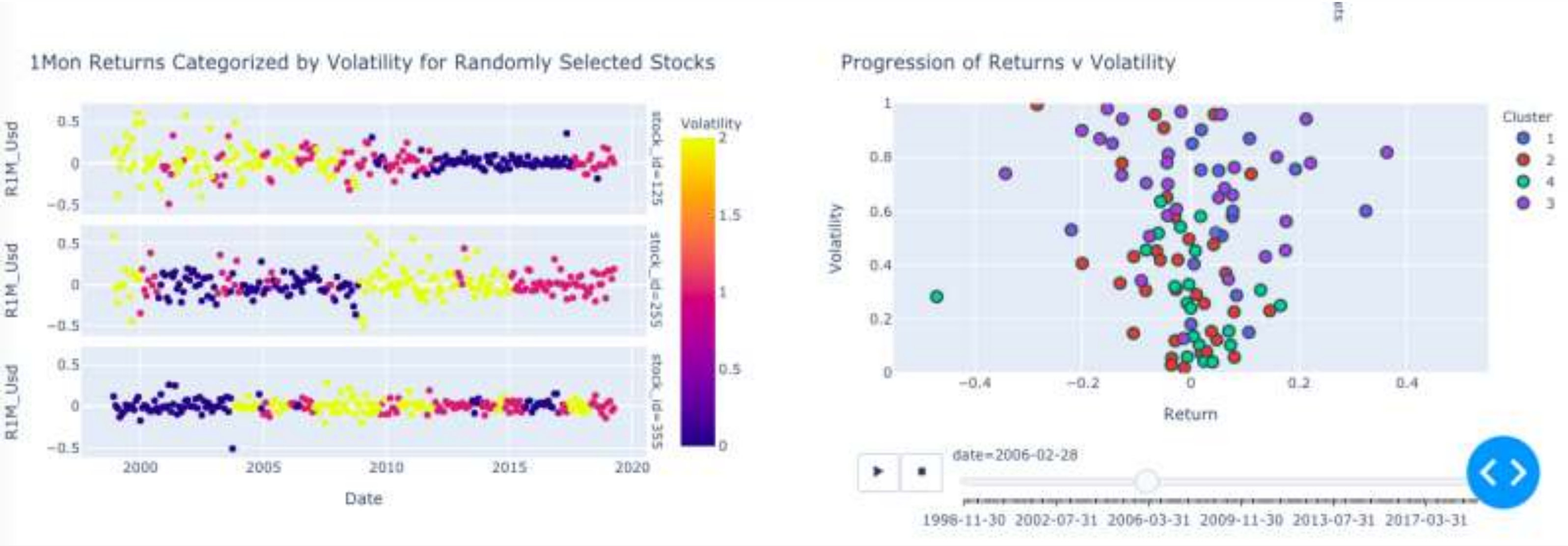
source: https://github.com/shokru/mlfactor.github.io/tree/master/material

Data Exploration Dashboard - Gif



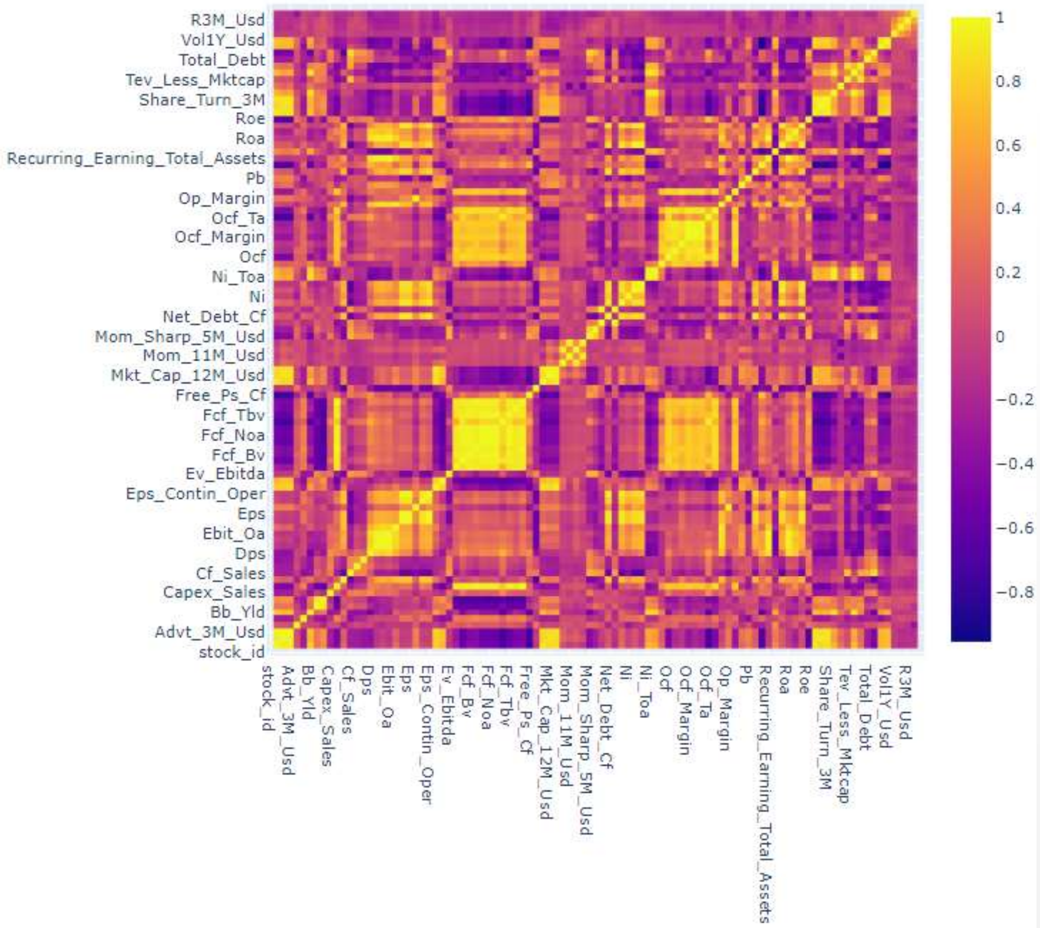
Data Exploration Dashboard – Screenshot 1

Data Exploration Dashboard – Screenshot 2



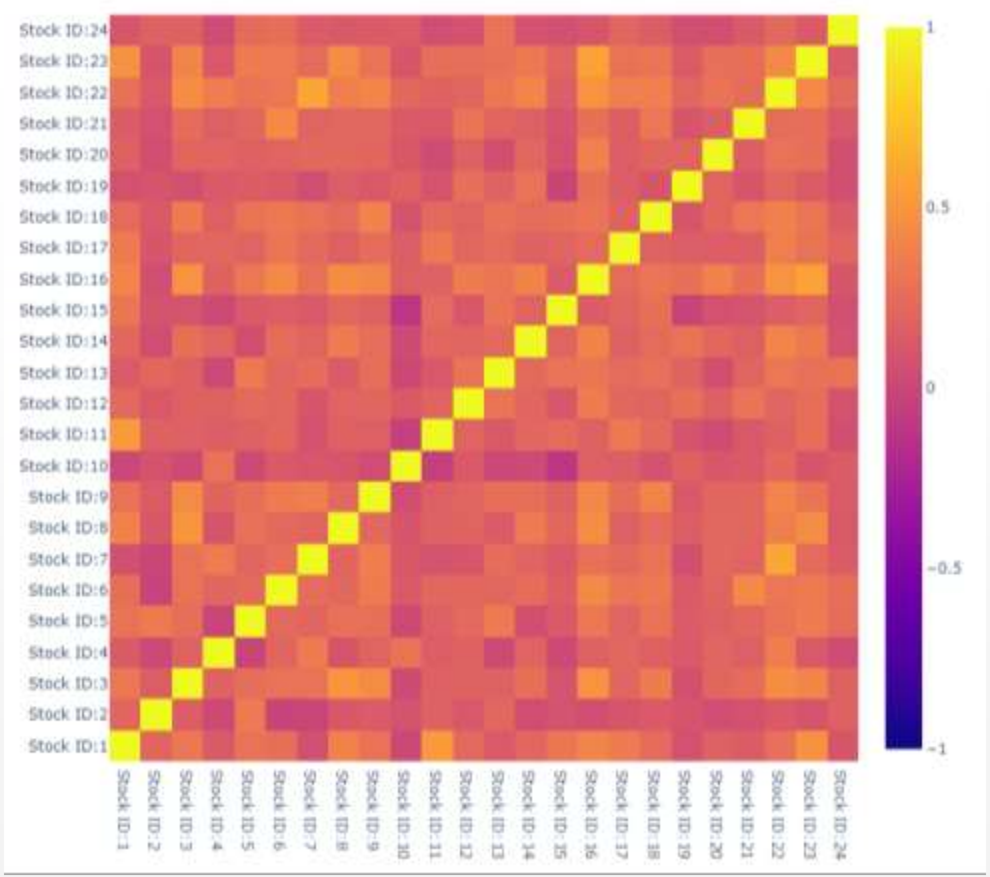
Data Exploration

Correlation Matrix of Features of a single stock

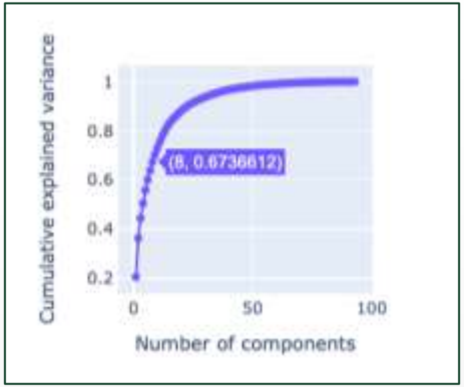


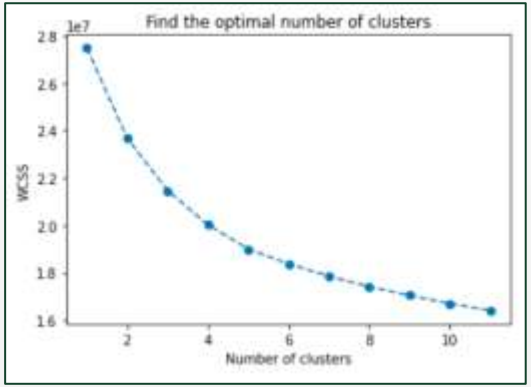
Data Exploration

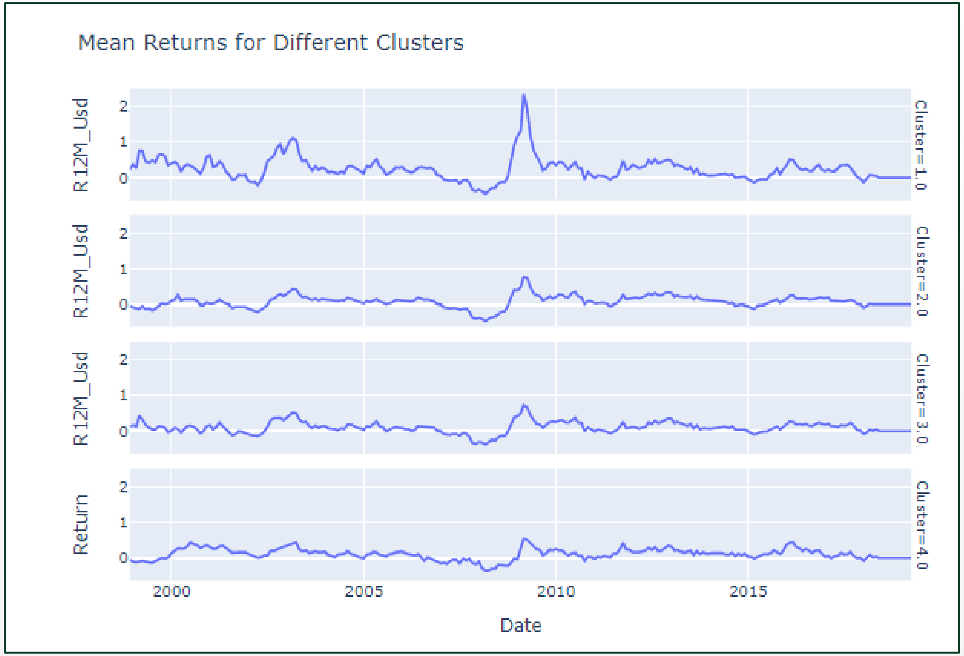
Correlation Matrix of Stocks' 1 month returns



Data Exploration – PCA and Clustering of Stocks







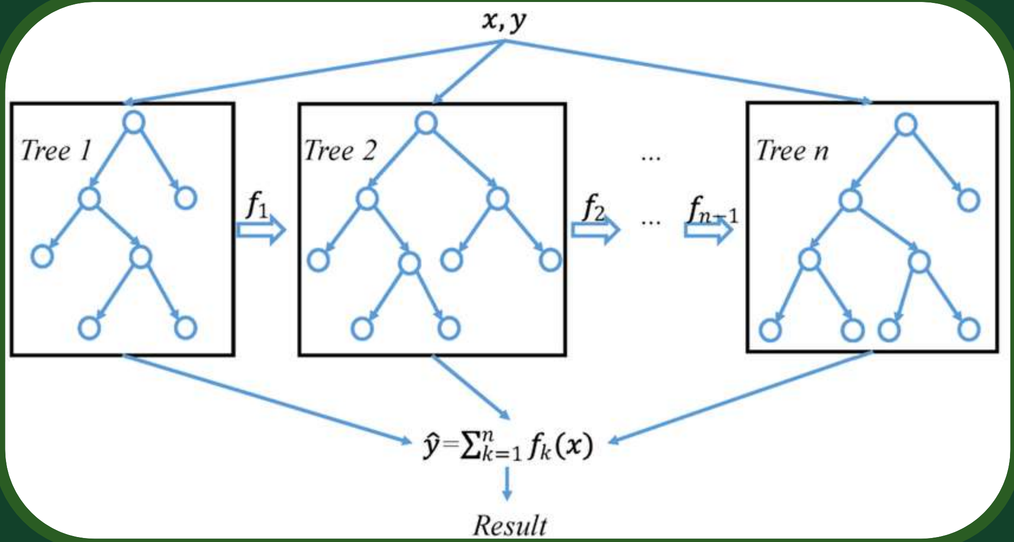
Problem Solving Strategy

3

Strategy

We will try 2 distinct methods to forecast returns

1. **Tree-based models:**
   - Leverages larges datasets with many features
   - Discover subsets of important features
2. **Autoencoders**:
   - Allow for a non-linear dimension reduction of features
   - Can detect stocks with "anomalous" factors



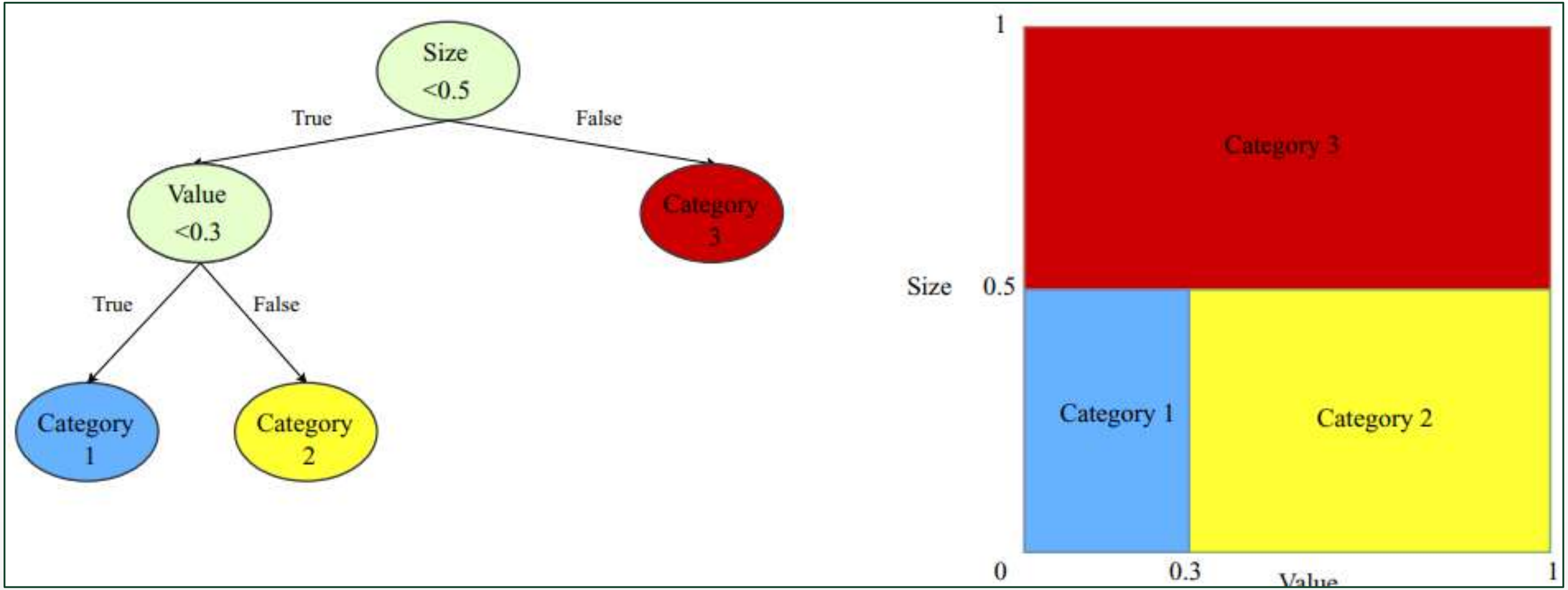
Tree-based models

Yaroslav Babich, Kiran Deol, Myles Sjogren, Ernest Tafolong

participants are sorted alphabetically

Overview of Tree Models

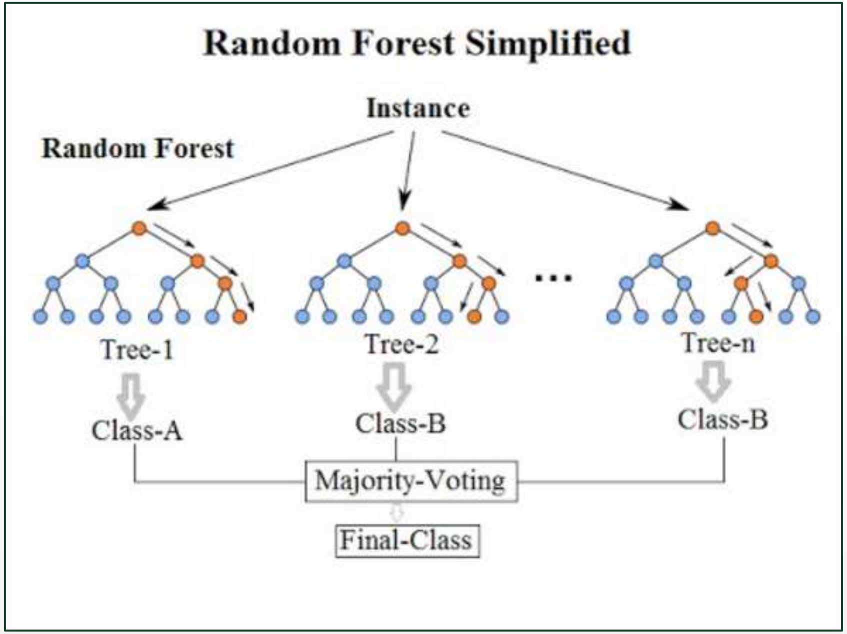
- Aimed at finding groups of observations that behave similarly
- Trees grow through a branching process which splits data according to certain thresholds/categories of a given predictor.
- At each step "impurity" or other error metrics are minimized



Representation of a tree

One tree isn’t Enough, we need a Forest

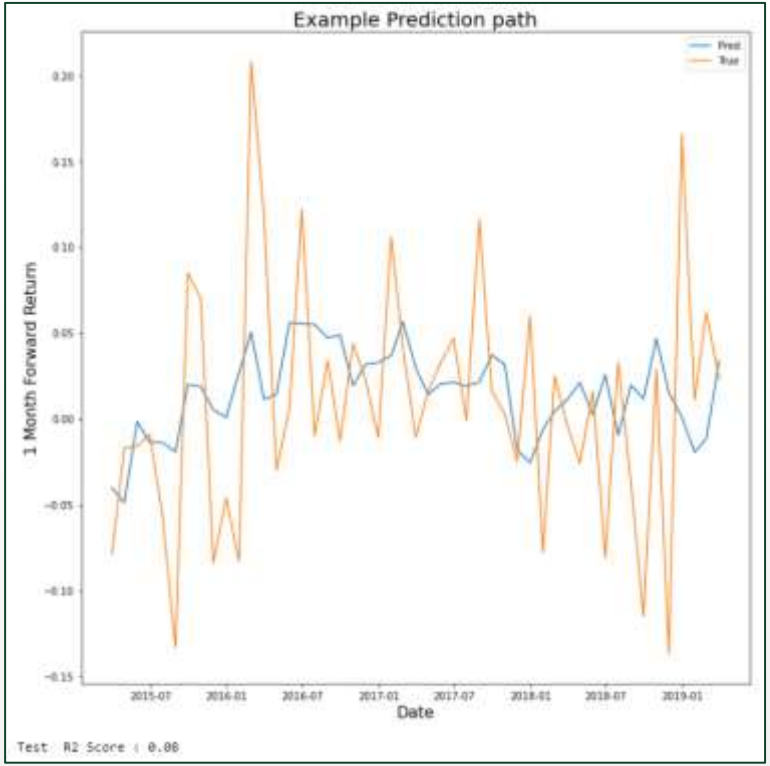
- Boosting: aggregate forecasts from many simpler trees
- Reduces correlation among different trees in the forest
- XGBoost, LightGBM



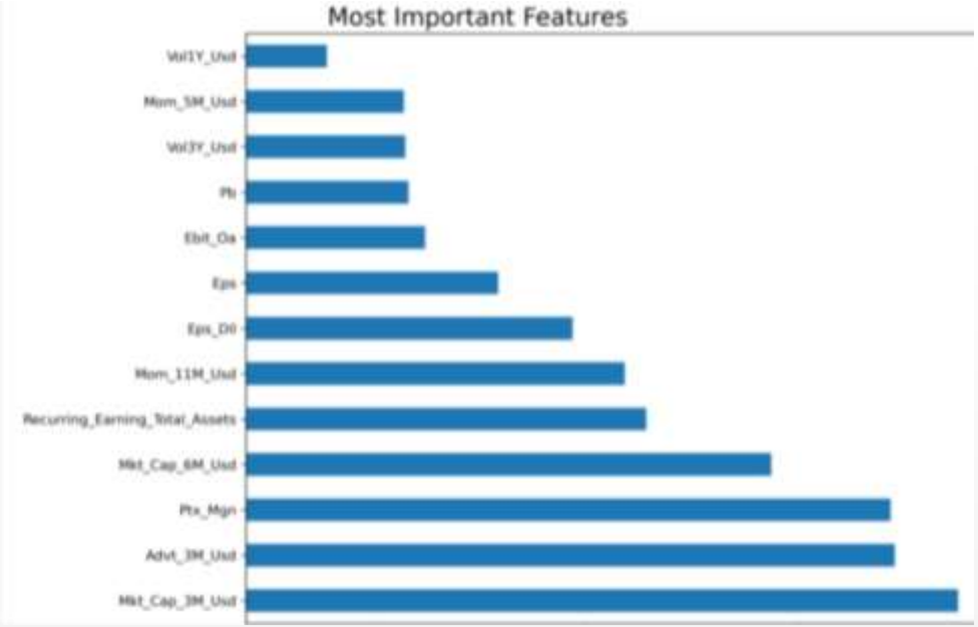
Results

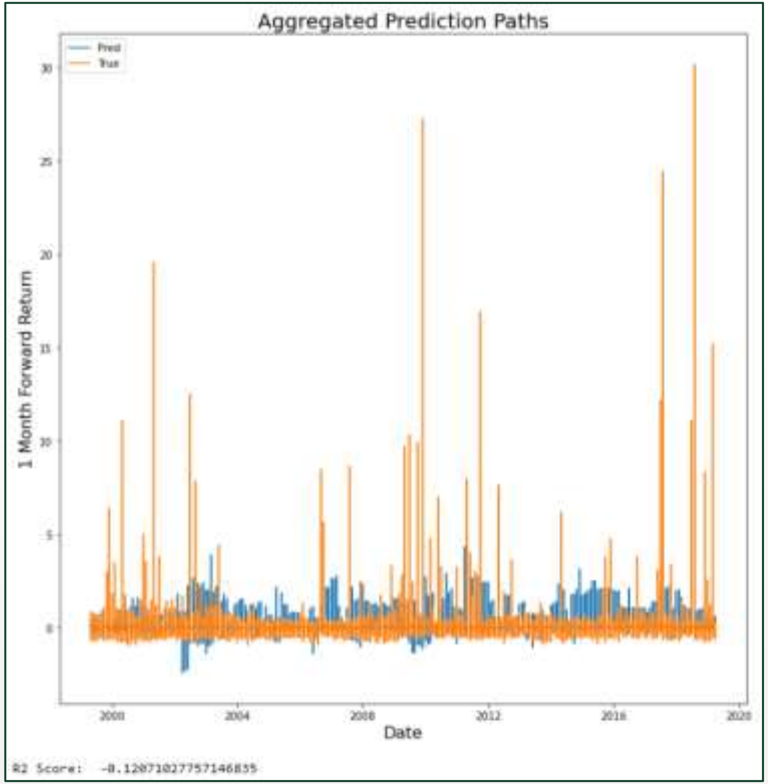
Important considerations:

- Train-Validation-Test split
- Model Hyperparameters
  - no. of trees, depth of trees, regularization parameters, etc.
- How long to Train
- Sparsity of Data
- Model generalization

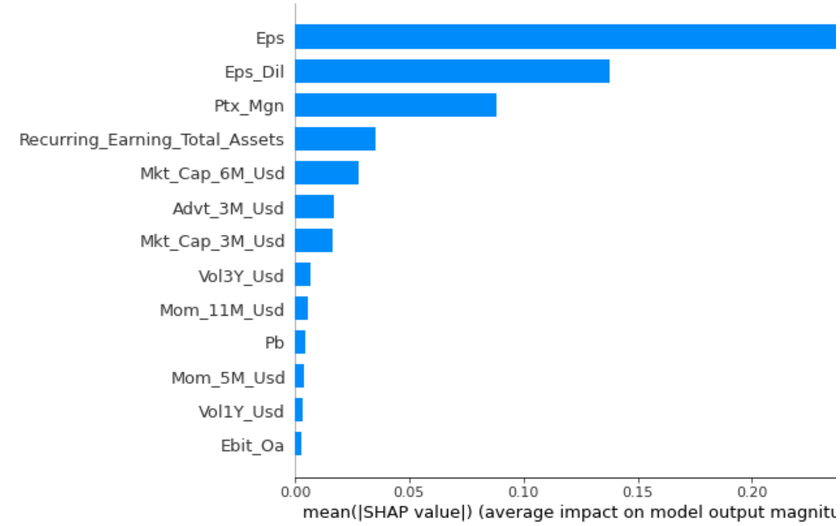


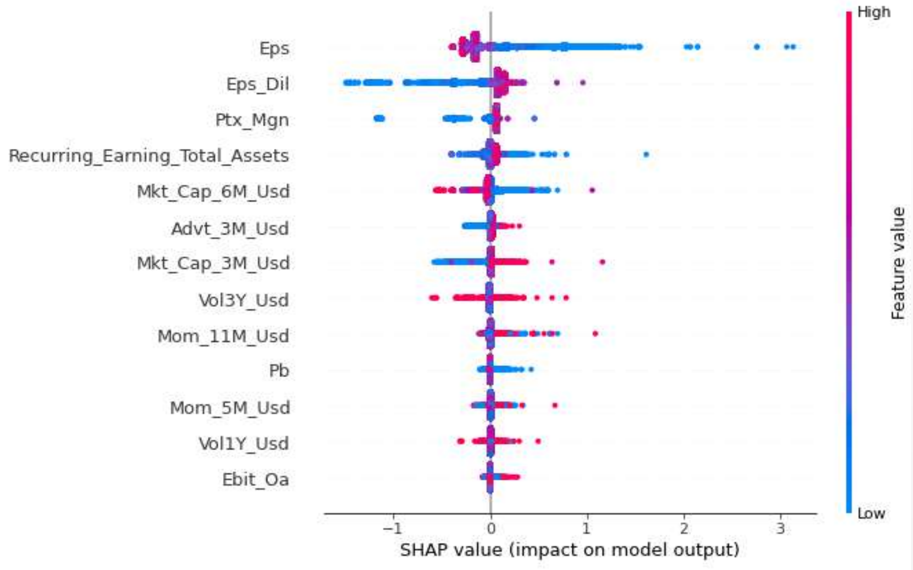
Results


Most Important Features
Vol1Y_Usd
Mom_5M_Usd
Vol3Y_Usd
Pb
Ebit_Oa
Eps
Eps_Dil
Mom_11M_Usd
Recurring_Earning_Total_Assets
Mkt_Cap_6M_Usd
Ptx_Mgn
Advt_3M_Usd
Mkt_Cap_3M_Usd

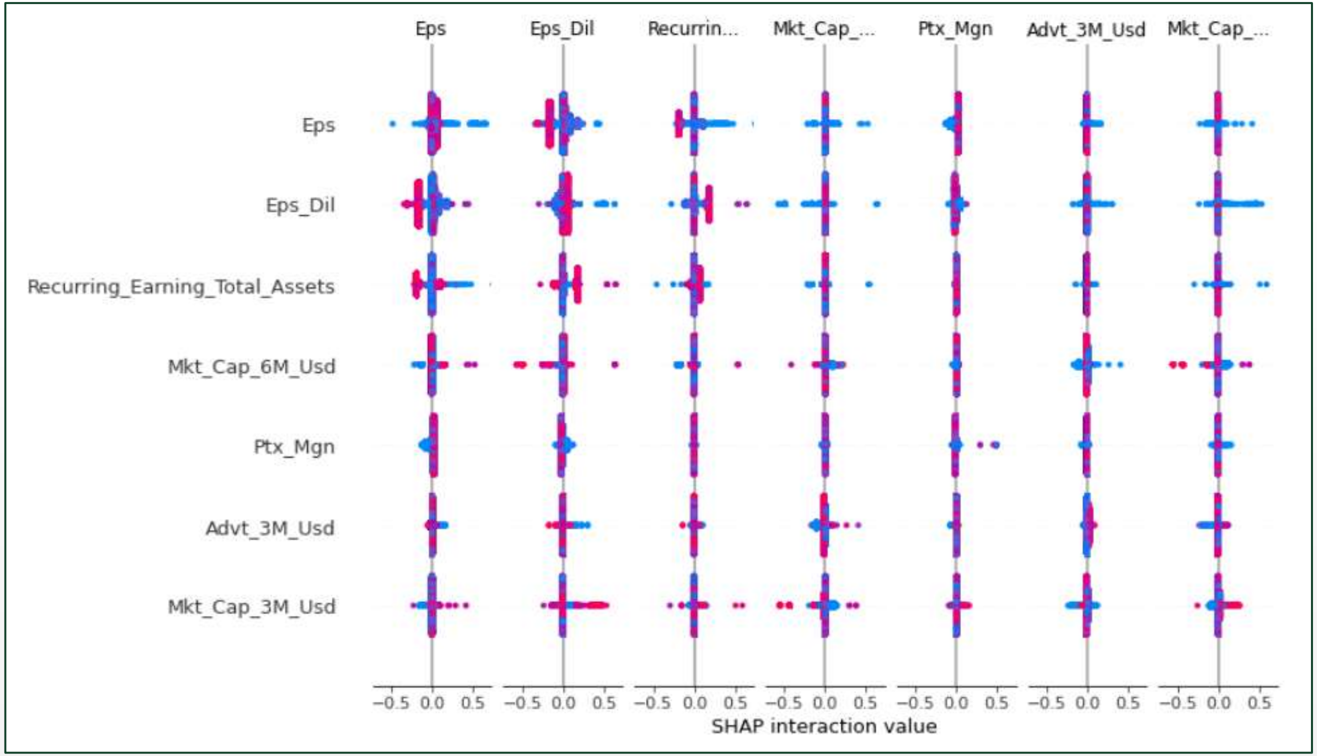

Aggregated Prediction Paths
Pred
True
1 Month Forward Return
0
5
10
15
20
25
30
2000
2004
2008
2012
2016
2020
Date
R2 Score: -0.1207102775714683S

SHAP – Feature importance



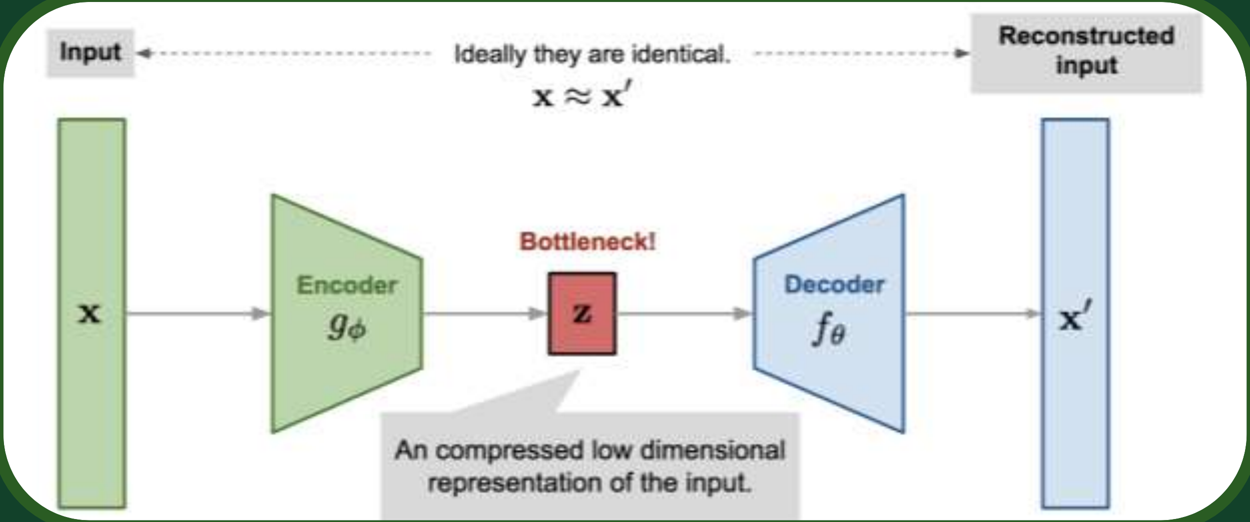


SHAP – Feature importance interactions



Takeaways

- Stock returns are inherently hard to predict at a certain timescale
- Trees takes some fine tuning
- Parameters and features that perform well for a given stock and time window may not generalize to others



Autoencoder models

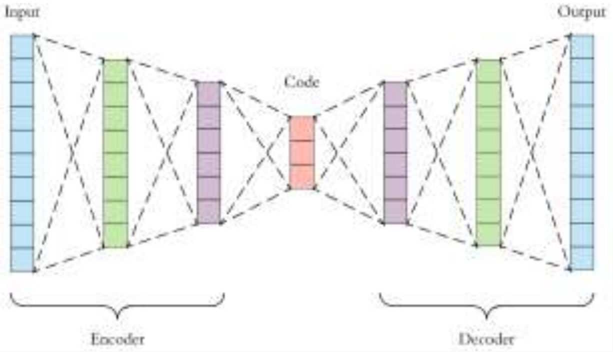
Qi Guo, Ehsan Rezaei, Javad Roustaei, Shiva Zokaee

participants are sorted alphabetically

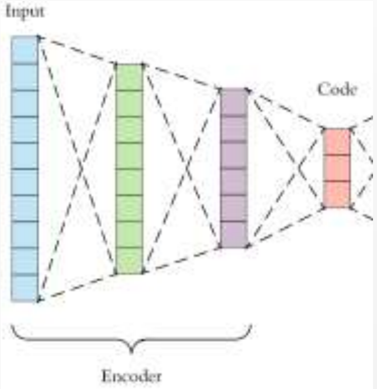
Motivation for using Autoencoders

- The fundamental goal of asset pricing is to understand the behavior of risk premiums.
- The high-dimensional nature of machine learning methods enhances their flexibility relative to more traditional econometric prediction techniques.
- This flexibility brings hope of better approximating the unknown and likely complex data generating process underlying equity risk premiums

A Brief Introduction to Autoencoders



$$\max P(\hat{x}|x)$$



$$x \in \mathbb{R}^n \Rightarrow y \in \mathbb{R}^m \ s.t. \ m \ll n$$

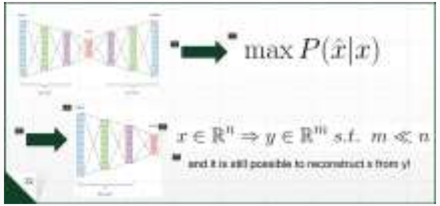
and it is still possible to reconstruct x from y!

Conditional Autoencoders

Cons of AEs:

- ✓ What if we have <u>more information</u> about x? Can we make even a better $\hat{x}$ using fewer dimensions and these extra information z?

$$\max P(\hat{x}|x) \longrightarrow \max P(\hat{x}|x, z)$$

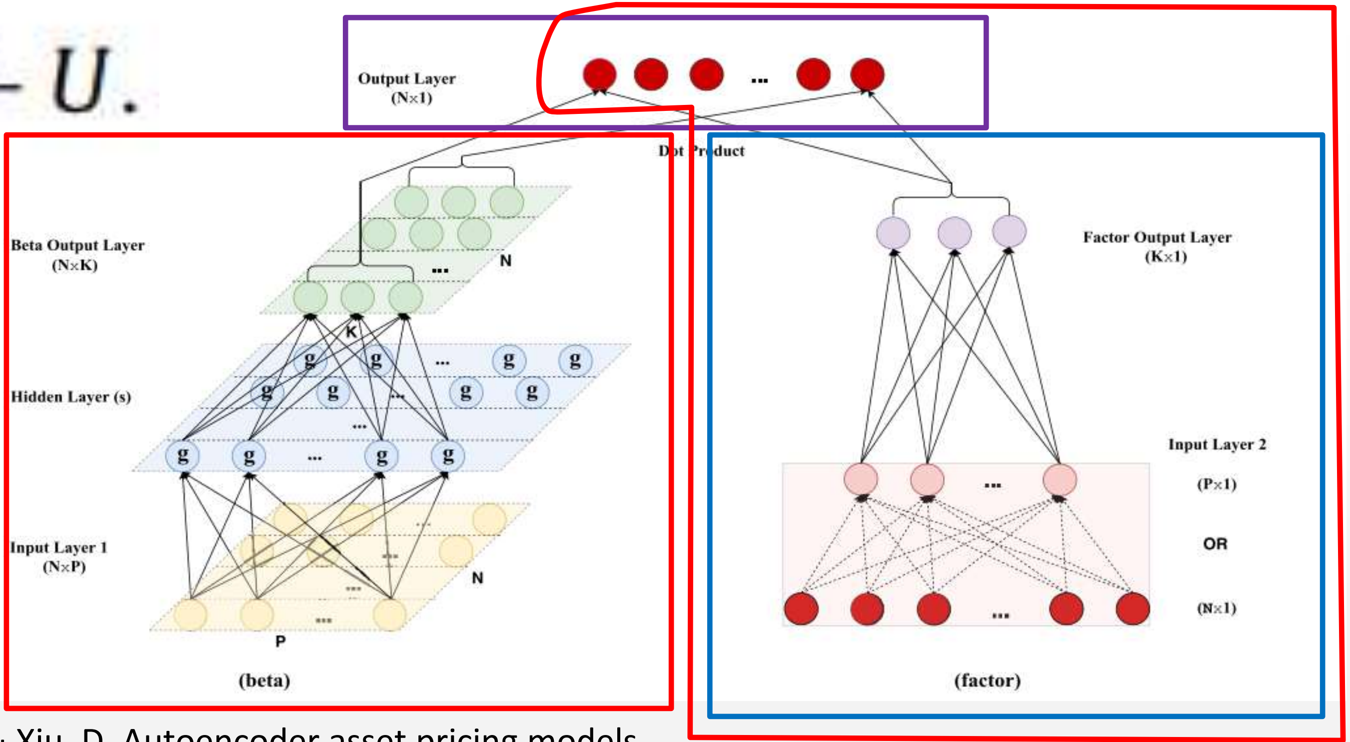


- ✓ Can the model be more interpretable?

$$R = \beta F + U.$$

Conditional Autoencoders

$$\boxed{R} = \boxed{\beta}\boxed{F} + U.$$



Gu, S., Kelly, B., & Xiu, D. Autoencoder asset pricing models

Model Validation and Testing

Framework

- Rolling Window → Train Window (120 months), Validation (36 months), Test (12 months)

Models

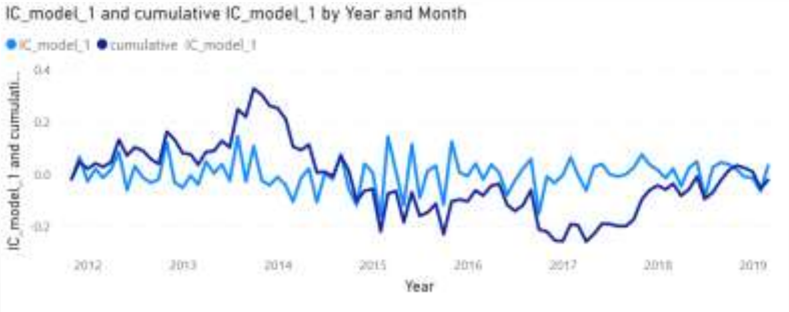
- Reducing factor dimensions: 94 features → 40.
- Beta Network Inputs: 94 features * N Return of Stocks for each month
- Factor Network Inputs: N
- Network Output: N

Implementation

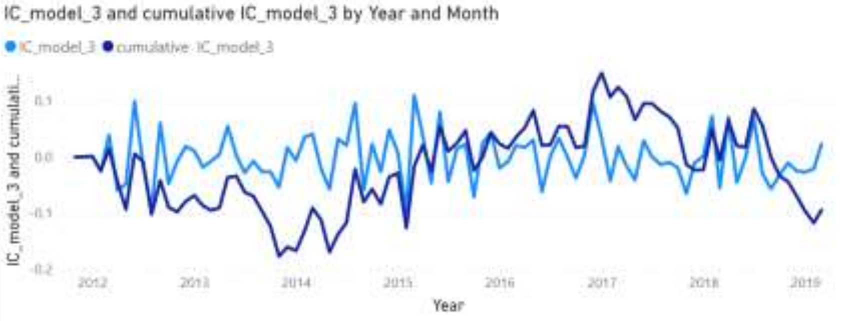
- More hyperparameters should be tuned and different activation function.
- Predictive R2 is very small which means predict return values directly is very hard.
- Therefore, the task can be reframed as trying to predict the rank of stocks based on future return.
- Then, the information coefficient (IC) becomes the measure of interest.

We tuned 4 models with 2 different hyperparameters:

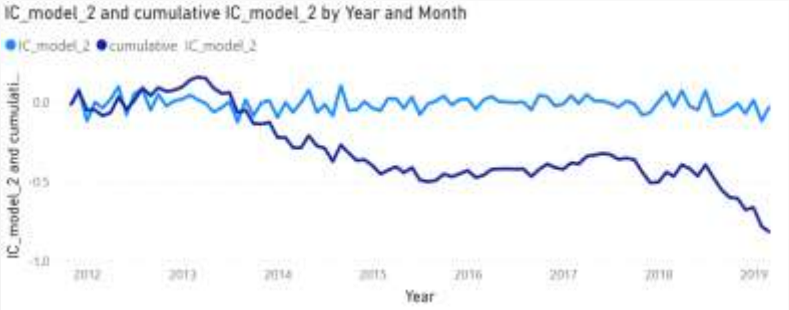
Model 1: learn rate: 0.01 beta layers size: [80, 40]



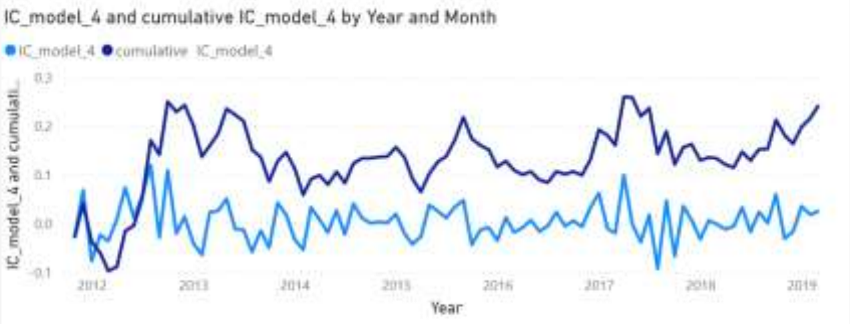
Model 3: learn rate: 0.001 beta layers size: [60,30]



Model 2: learn rate: 0.01 beta layers size: [80, 40]

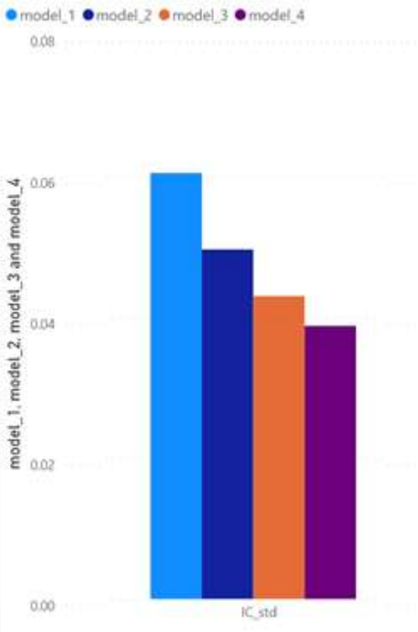


Model 4: learn rate: 0.001 beta layers size: [60,30]

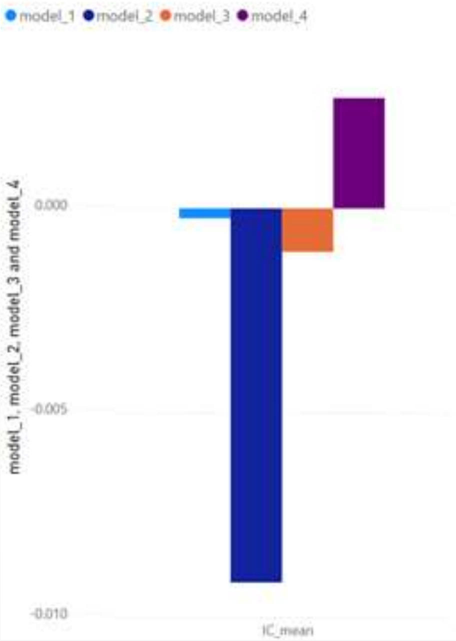


Results explanation

- The 4 ICs show that Model 4 had the best performance and can provide positive cumulative IC.
- We believe a well-tuned model would provide high IC. The beta-autoencoder model appears to be promising for predicting risk premiums.



The standard deviation is decreasing by adding each model



By comparing the mean of each model we can see the progress.

Average of model_1	Average of model_2	Average of model_3	Average of model_4
-0.00001	-0.00015	0.0000026	0.0000002

The average of $R2$ score within each model is getting better by each model.

Concluding words

4

Conclusion

- Both Tree-based and Autoencoder models were developed to forecast future return of securities based on their fundamental factors
- We highlighted the importance of adequate training and validation to prevent "Look-ahead" bias
- Autoencoder models can model multiple covarying latent factors and are less reliant on historical data

Future development

- Forecast models developed could be expanded to portfolio composition and portfolio performance forecast
- Smart Beta Exchange-Traded Funds (ETFs) are rising in popularity among investors
- They can provide a variety of investment and risk-management strategies at lower fees

Thanks!

Any questions?

Sources

Coqueret, G. & Guida, T. (2021). *Machine Learning for Factor Investing*. http://www.mlfactor.com/

Coqueret, G. & Guida, T. (2019). Machine Learning for Factor Investing. [Data set]. https://github.com/shokru/mlfactor.github.io/tree/master/material

Gu, S., Kelly, B., & Xiu, D. (2021). Autoencoder asset pricing models. *Journal of Econometrics* 222:429-450

Gu, S., Kelly, B., & Xiu, D. (2018). Empirical Asset Pricing via Machine Learning *The Review of Financial Studies* 33:2223-2273

Python packages and libraries

- Numpy
- Pandas
- Scipy
- MKL
- Bottleneck

- Scikit-learn
- Xgboost
- LightGBM
- Optuna
- SHAP

- Matplotlib
- Seaborn
- Plotly
- Dash

Appendices

Appendix I – Fundamental Factors

1	average daily volume in amount in USD over 12 months
2	average daily volume in amount in USD over 3 months
3	average daily volume in amount in USD over 6 months
4	total sales on average assets
5	buyback yield
6	book value
7	capital expenditure on price to sale cash flow
8	capital expenditure on sales
9	cash dividends cash flow
10	cash per share
11	cash flow per share
12	debt to equity
13	dividend yield
14	dividend per share
15	EBIT on book value
16	EBIT on non operating asset
17	EBIT on operating asset
18	EBIT on total asset
19	EBITDA margin

20	earnings per share
21	earnings per share basic
22	earnings per share growth
23	earnings per share continuing operations
24	earnings per share diluted
25	enterprise value
26	enterprise value on EBITDA
27	fixed assets on common equity
28	free cash flow
29	free cash flow on book value
30	free cash flow on capital employed
31	free cash flow margin
32	free cash flow on net operating assets
33	free cash flow on operating assets
34	free cash flow on total assets
35	free cash flow on tangible book value
36	free cash flow on total operating assets
37	free cash flow yield
38	free cash flow on price sales
39	intangibles on revenues
40	interest expense coverage
41	average market capitalization over 12 months in USD
42	average market capitalization over 3 months in USD
43	average market capitalization over 6 months in USD

44	price momentum 12 - 1 months in USD
45	price momentum 6 - 1 months in USD
46	price momentum 12 - 1 months in USD divided by volatility
47	price momentum 6 - 1 months in USD divided by volatility
48	net debt on EBITDA
49	net debt
50	net debt on cash flow
51	net margin
52	net debt yield
53	net income
54	net income available margin
55	net income on operating asset
56	net income on total operating asset
57	net operating asset
58	operating asset
59	operating cash flow
60	operating cash flow on book value
61	operating cash flow on capital employed
62	operating cash flow margin
63	operating cash flow on net operating assets
64	operating cash flow on operating assets
65	operating cash flow on total assets
66	operating cash flow on tangible book value
67	operating cash flow on total operating assets
68	operating margin

69	net margin 1Y growth
70	cash flow from operations per share net
71	price to book
72	price earnings
73	margin pretax
74	reccuring earnings on total assets
75	return on capital
76	revenue
77	return on assets
78	return on capital
79	return on capital employed
80	return on equity
81	price to sales
82	average share turnover 12 months
83	average share turnover 3 months
84	average share turnover 6 months
85	total assets
86	total enterprise value less market capitalization
87	total debt on revenue
88	total capital
89	total debt
90	total debt on capital
91	total liabilities on total assets
92	volatility of returns over one year
93	volatility of returns over 3 years