Feature Selection is the new factor modeling

Ernie Chan, Ph.D.
Founder and CEO
PredictNow.ai Inc.

PredictNow.ai

Traditional Factors

Cross-sectional	Time-series
Stock-specific	Market-wide
E.g. P/E, B/M, DivYld, …	E.g. HML, SMB, WML, …
Use that to explain/predict stocks' returns	Use that explain/predict portfolio/strategy's returns
Linear regression model	
Duality between CS and TS factors via hedge portfolios (predictnow.ai/blog)	

Factor modeling vs machine learning

Factors = **Predictors** = **Features** = **Independent Variables**

1. Many trading strategies' returns not attributable to traditional factors (e.g. the Gamestop mania, Archegos liquidation.)
2. Needs hundreds of potential new factors: but many are likely non-informative or redundant.
3. Traditional factor models are linear: unable to account for conditional co-dependence.
4. Linear regression cannot handle redundant ("collinear") factors.
5. Risk management by ML ("metalabeling") requires classification, not regression.

- **Feature selection** is a procedure to select only an informative subset of all features for training.

Benefits of Feature Selection

Including uninformative features may decrease bias (***in-sample error***) but can typically increase variance (***out-of-sample error***).

1. Feature selection is also found to improve OOS **predictive accuracy**.
   - See our research papers.

2. Feature importance ranking also improves ***human interpretability*** *or* ***explanability***.
   - ML not black box anymore!

PredictNow.ai

Mean Decrease Accuracy (MDA)

If we randomize the rows of a feature *f* in train set, how much does the average predictive accuracy (cross) validation set decrease?

- The more it decreases, the more important feature *f* is!
- Alternative feature selection methods are LIME and SHAP.

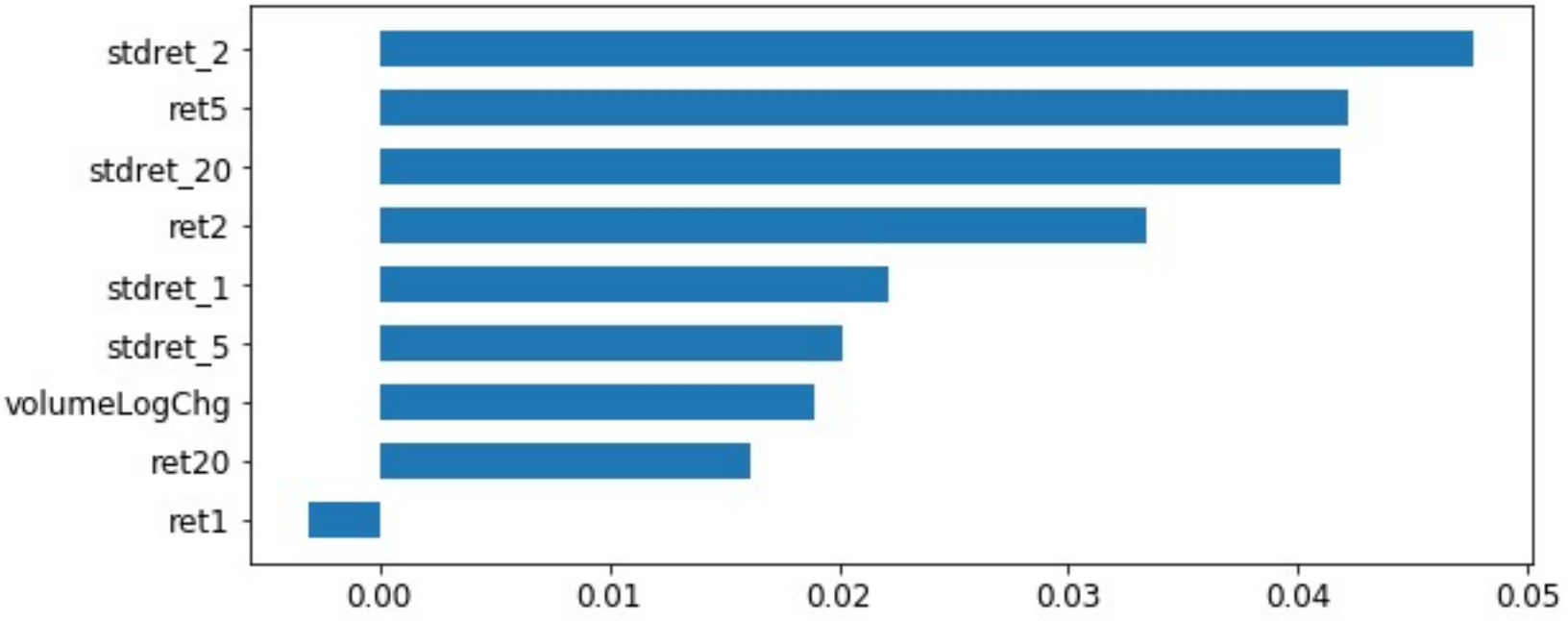
Difficulty: Important Features Vary!

Vary with random seed during cross validation and random shuffle of features.

Lost all interpretability if rankings are not stable and reproducible!

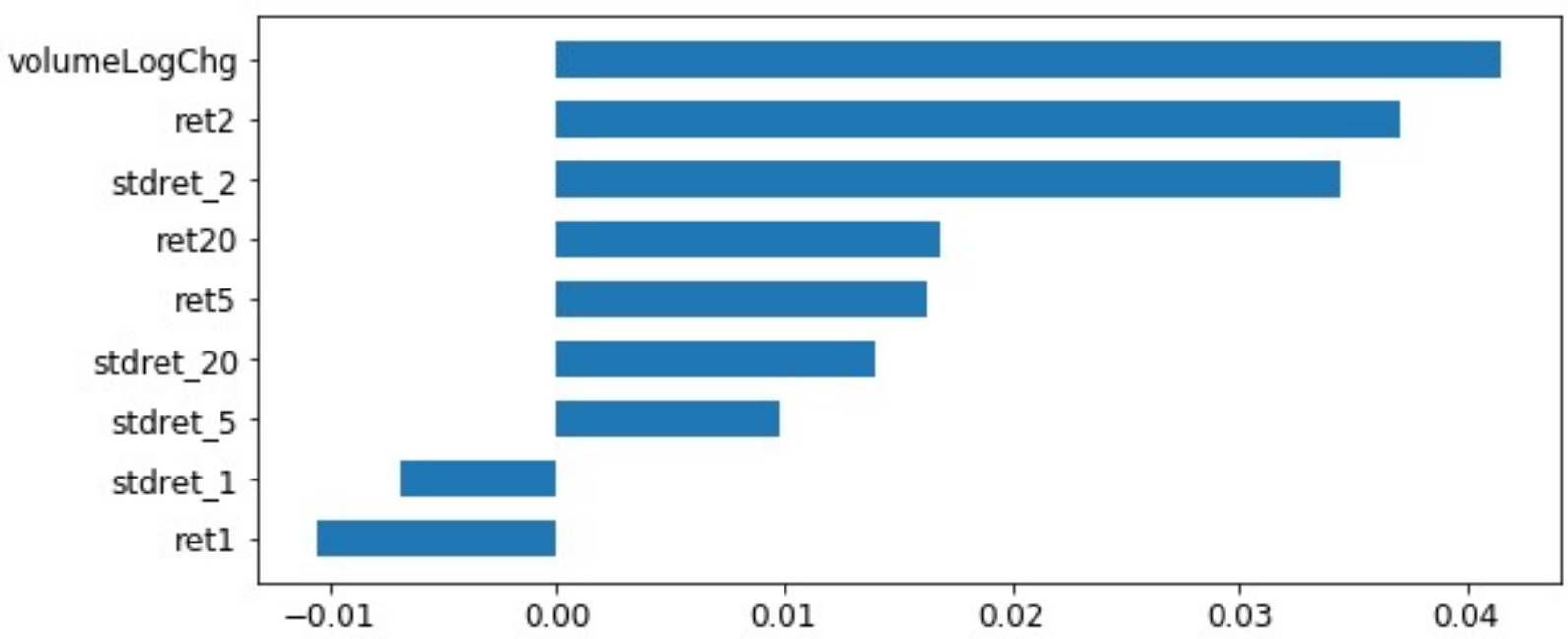
Ex: Use technical indicators such as past returns and volatility to predict USO's (oil ETF) return.

Seed=0 Importance Rankings

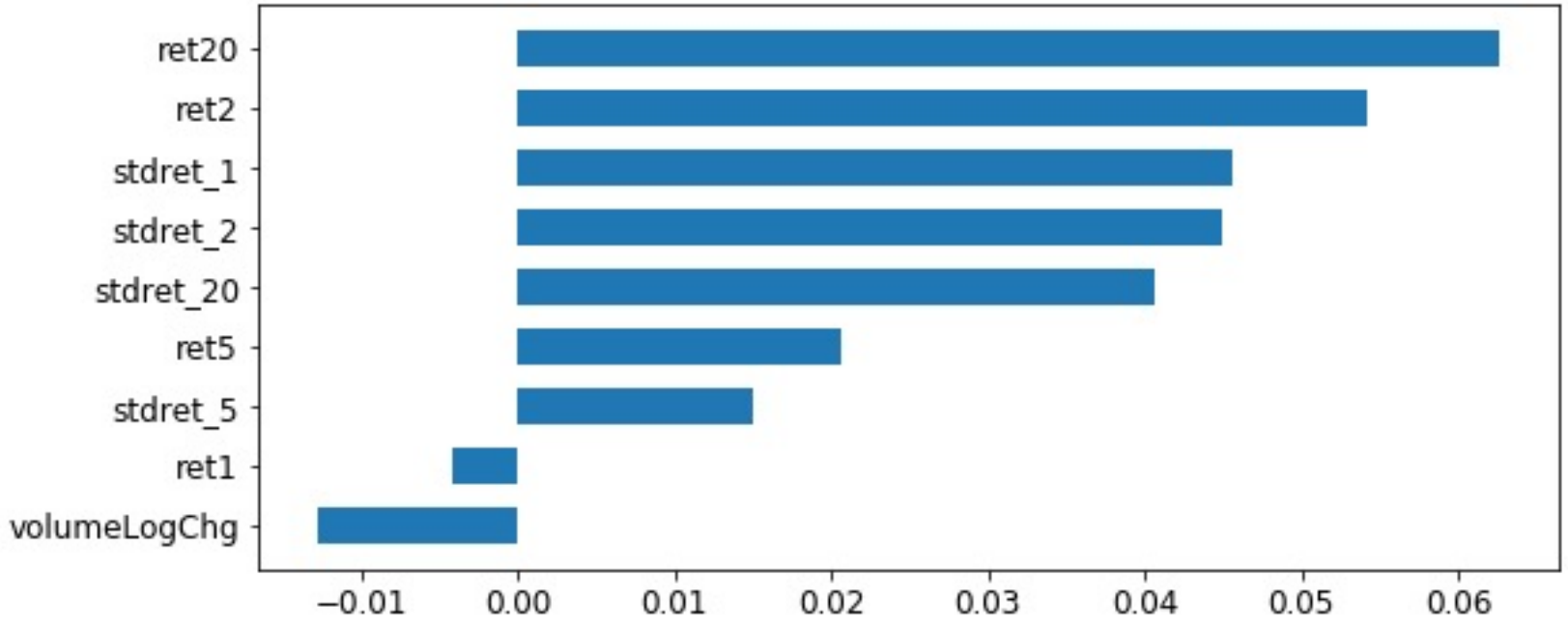


PredictNow.ai

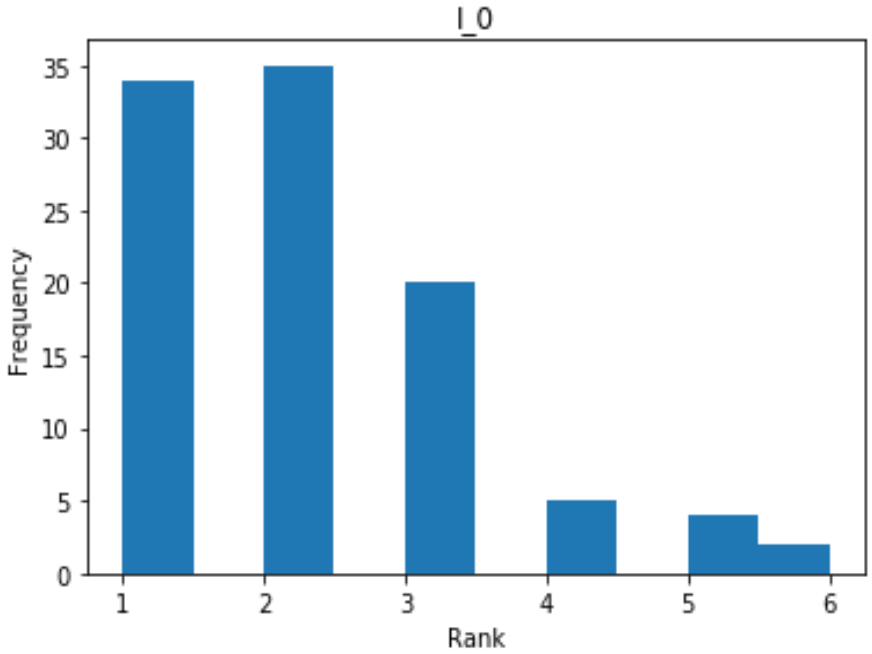
Seed=1 Importance Rankings



Seed=2 Importance Rankings



“Most Important Feature” Changes!



From Man and Chan, “The Best Way to Select Features?” JFDS Winter 2021

PredictNow.ai

How to Reduce Ranks Instability?

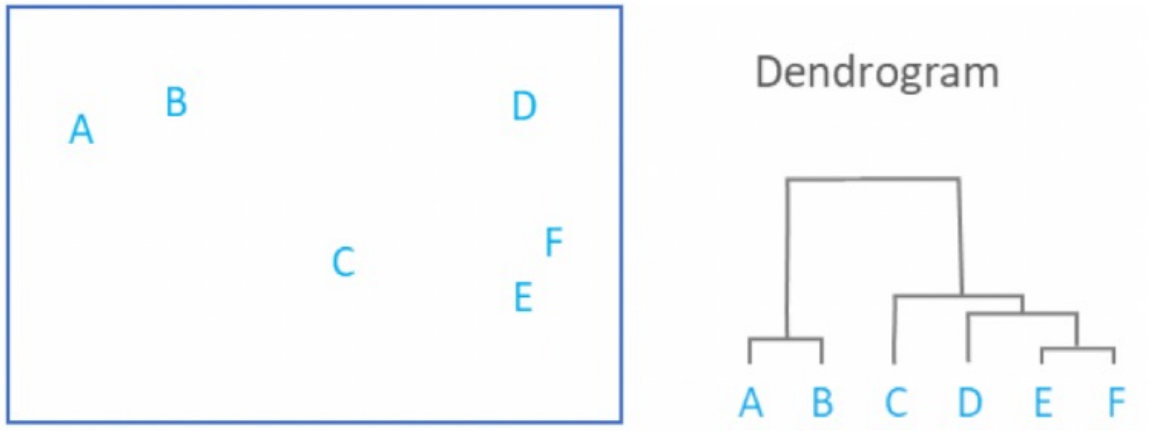
1. Average over many random iterations of a selection algorithm.
   - Ex: Increase number of permutations in MDA.

2. Use *clusters* to group features first.
   - See Man and Chan, 2021b, “Cluster-based Feature Selection”.
   - Email info@predictnow.ai for a copy.

PredictNow.ai

Cluster-Based Feature Selection

- Use *hierarchical clustering* to group features.
- Bottom–up: start with clusters with 1 feature.



From www.displayr.com/what-is-hierarchical-clustering/

Feature Selection on Clusters

MDA is applied to clusters, not individual features.

- All features within a cluster are permuted at the same time.
- Importance score $\propto$ 1/rank of cluster among all clusters.

Ex: Breast Cancer Data

Binary classification of tumors: benign vs malignant[1].

30 features based on medical images.

Clusters have clear interpretability.

- We can even apply a “topic” to them.

[1] “Wisconsin Breast Cancer Dataset” https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).

Top Cluster

Topic	Cluster Importance Scores	Features
1. Geometry summary	0.360	'mean radius', 'mean perimeter', 'mean area', 'mean compactness', 'mean concavity', 'mean concave points', 'radius error', 'perimeter error', 'area error', 'worst radius', 'worst perimeter', 'worst area', 'worst compactness', 'worst concavity', 'worst concave points'

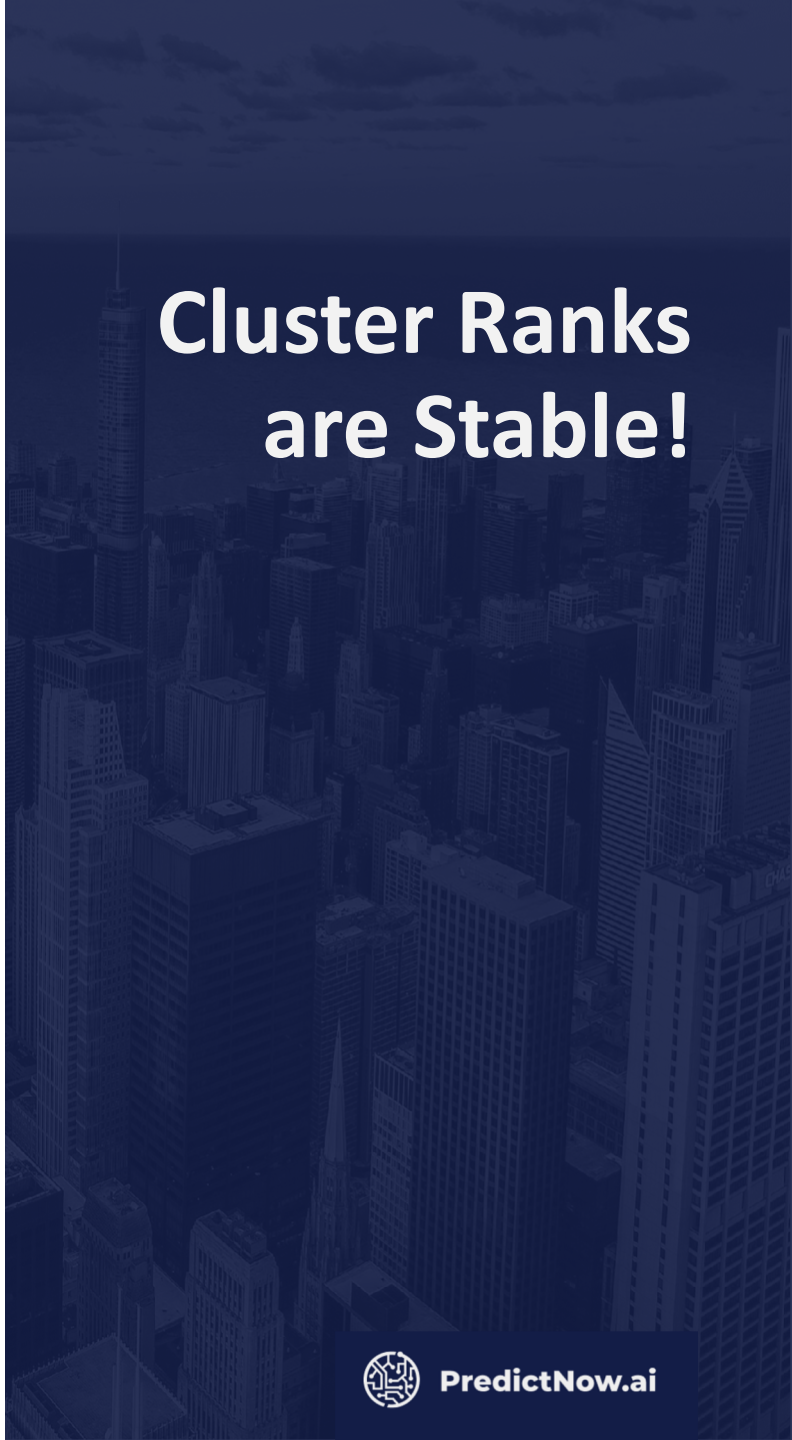
Other Clusters

2. Texture summary	0.174	'mean texture’, 'worst texture'
3. Geometry error	0.112	'compactness error', 'concavity error', 'concave points error', 'fractal dimension error'
4. Smoothness error	0.092	'smoothness error'
5. Symmetry error	0.062	'symmetry error'
6. Texture error	0.056	'texture error'
7. Symmetry summary	0.055	'mean symmetry’, 'worst symmetry'
8. Fractal dimension	0.049	'mean fractal dimension', ' worst fractal dimension'
9. Smoothness summary	0.042	'mean smoothness’, ‘worst smoothness'

Cluster Ranks are Stable!

The ***first*** cluster remains ***first*** irrespective of random seed.

The ***second*** cluster remains ***second*** irrespective of random seed.



Ex: Predict SPX Excess Returns

Predict monthly SPX index return – *risk free rate*[1].

Features are **fundamental** and **technical** factors such as dividend price ratio (d/p), earning price ratio (e/p), stock return variance (svar), etc.

[1] http://www.hec.unil.ch/agoyal/docs/PredictorData2019.xlsx.

PredictNow.ai

Stable & Interpretable Clusters

Topic	Cluster Scores	Features
Fundamental	0.667	d/p, d/y, e/p, b/m, ntis, tbl, lty, dfy, dfr, infl
Technical	0.333	d/e, svar, ltr, tms

Ex: Tail Reaper Strategy

- QTS Tail Reaper strategy (www.qtscm.com).
- Our proprietary tail hedge strategy.
- Use 160 features to predict whether strategy will be profitable each day ("metalabeling").
- First and second ranked clusters *remain same* w.r.t. all random seeds.
- 8 out of 44 clusters have above-average importance scores – selected.

Predictive Performance Improved!

	F1	AUC	Acc
cMDA	0.658	0.672	0.614
cMDA (top 2)	0.595	0.640	0.571
MDA	0.602	0.537	0.529
Full	0.481	0.416	0.414

Implementation

- cMDA is already implemented as part of PredictNow.ai SaaS!

  -Thanks to Nancy Xin Man and Radu Ciobanu

- We also created **hundreds of ready-made features**, many proprietary, for our subscribers.

Request a free trial at:
py.predictnow.ai/register

PredictNow.ai