Financial Data Professional Institute

A Conversation with Tony Guida

"Long Term Machine Learning Predictions for US equity"

Mehrzad Mahdavi, Executive Director, FDP Institute
Kathy Wilkens, Senior Advisor, FDPI Curriculum
Mirjam Dekker, Project Manager, FDP Institute

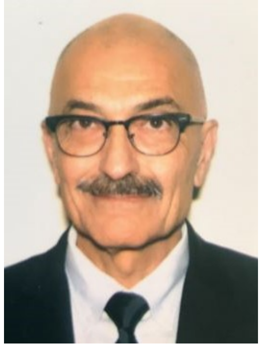
www.fdpinstitute.org
March 5, 2020

Agenda

- Welcome
- Introductions

Tony Guida
Executive Director,
Sr. Quant Research
Ram Active Investment

Katherine Wilkens
Sr. Curriculum
Advisor FDPI



Mehrzad Mahdavi
Executive Director
FDPI

Mirjam Dekker
Project Manager
FDPI

- User case *"Long Term ML Predictions for US Equity"*
- FDP Curriculum
- Q & A

Long Term ML predictions for US Equity

Tony Guida

Executive Director – Senior Quant Research

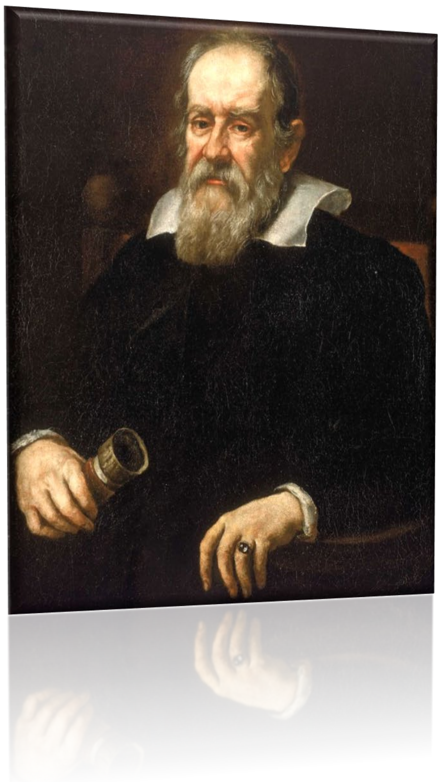
Editor - Journal of Machine Learning in Finance

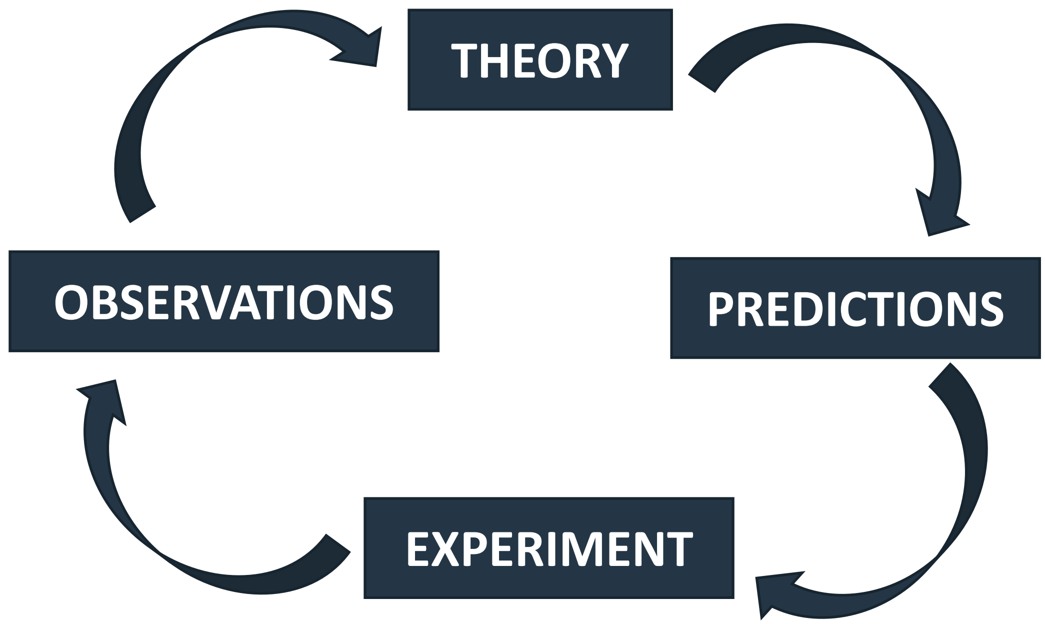
ACTIVE INVESTMENTS

[LT ML predictions for EQ]

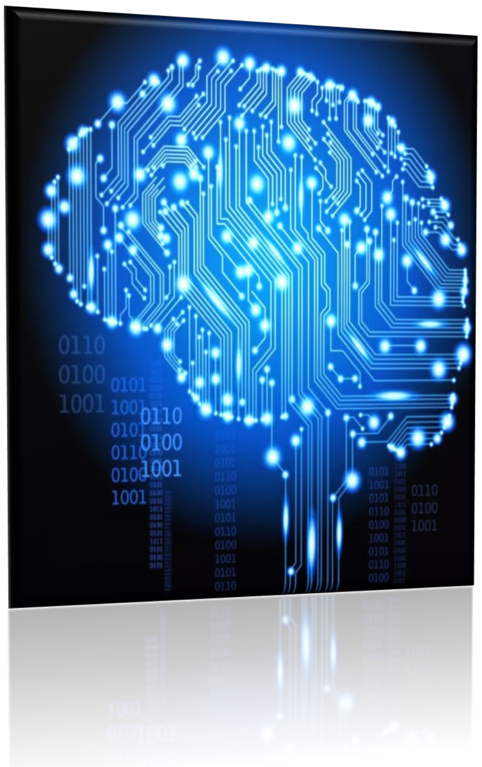
An empirical exercice

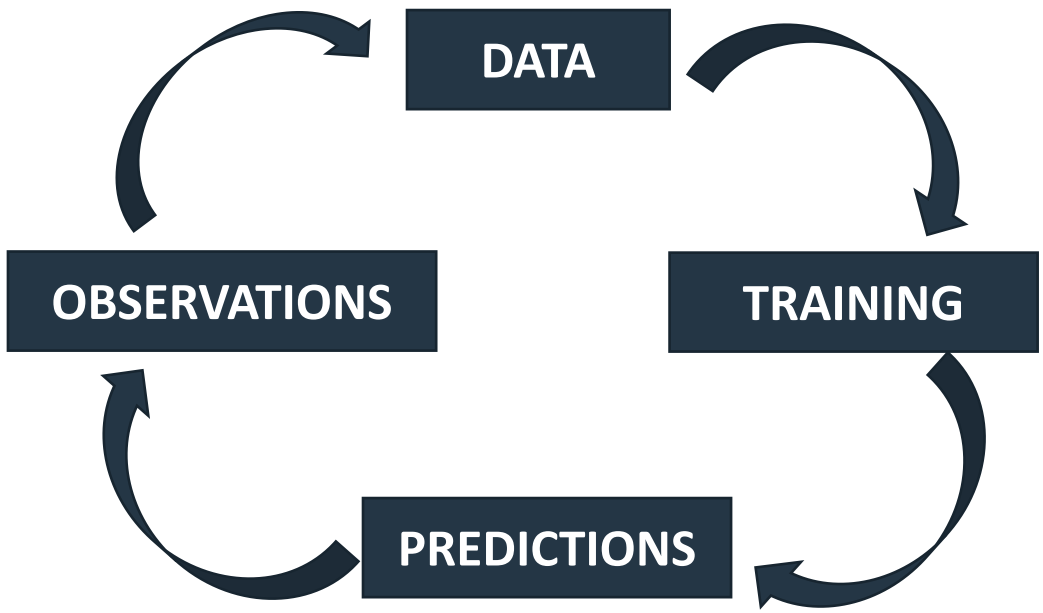
A bit of Epistemology





A New Way for Research





ACTIVE INVESTMENTS

Asset pricing vs Empirical Asset Pricing

- Econometrics vs machine learning
- Share a **common goal**: build a predictive model
- Radical **difference** remains in the “**learning**” part
- Econometrics is a **beta question** while ML is an **alpha answer**
- From a practitioner standpoint **ML** more suited to **high dimensional non-linear signals’ space**
- Poses the problem of **maximizing “factor zoo**”

[LT ML predictions for EQ] definitions and concepts

eXtreme Gradient Boosting : quick introduction

General objective of tree ensemble for K trees

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

$$Obj = \underbrace{\boxed{\sum_{i=1}^{n} l(y_i, \hat{y}_i)}}_{\text{Training on loss}} + \underbrace{\boxed{\sum_{k=1}^{K} \Omega(f_k)}}_{\text{Complexity of the trees}}$$

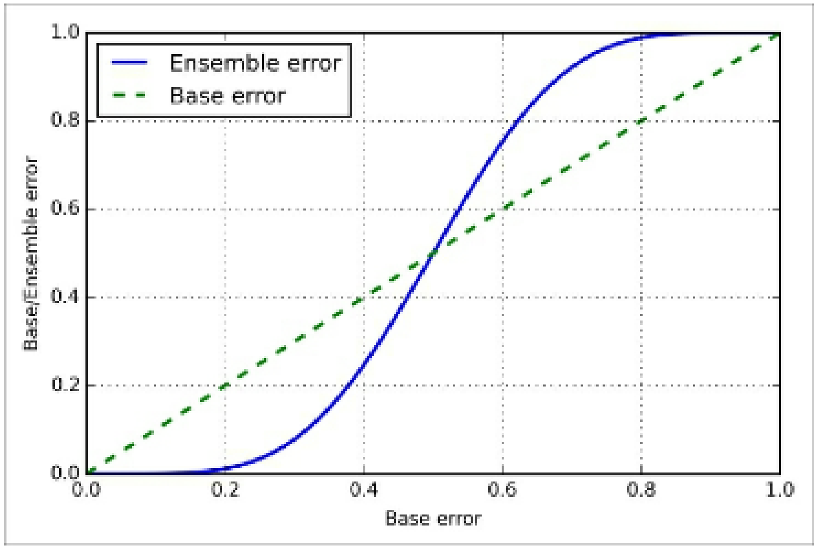
$$\left.\begin{aligned}
\hat{y}_i^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
\hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\dots \\
\hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned}\right\} \text{Additive training}$$

http://xgboost.readthedocs.io/en/latest/model.html#

Wisdom of the crowd in ML

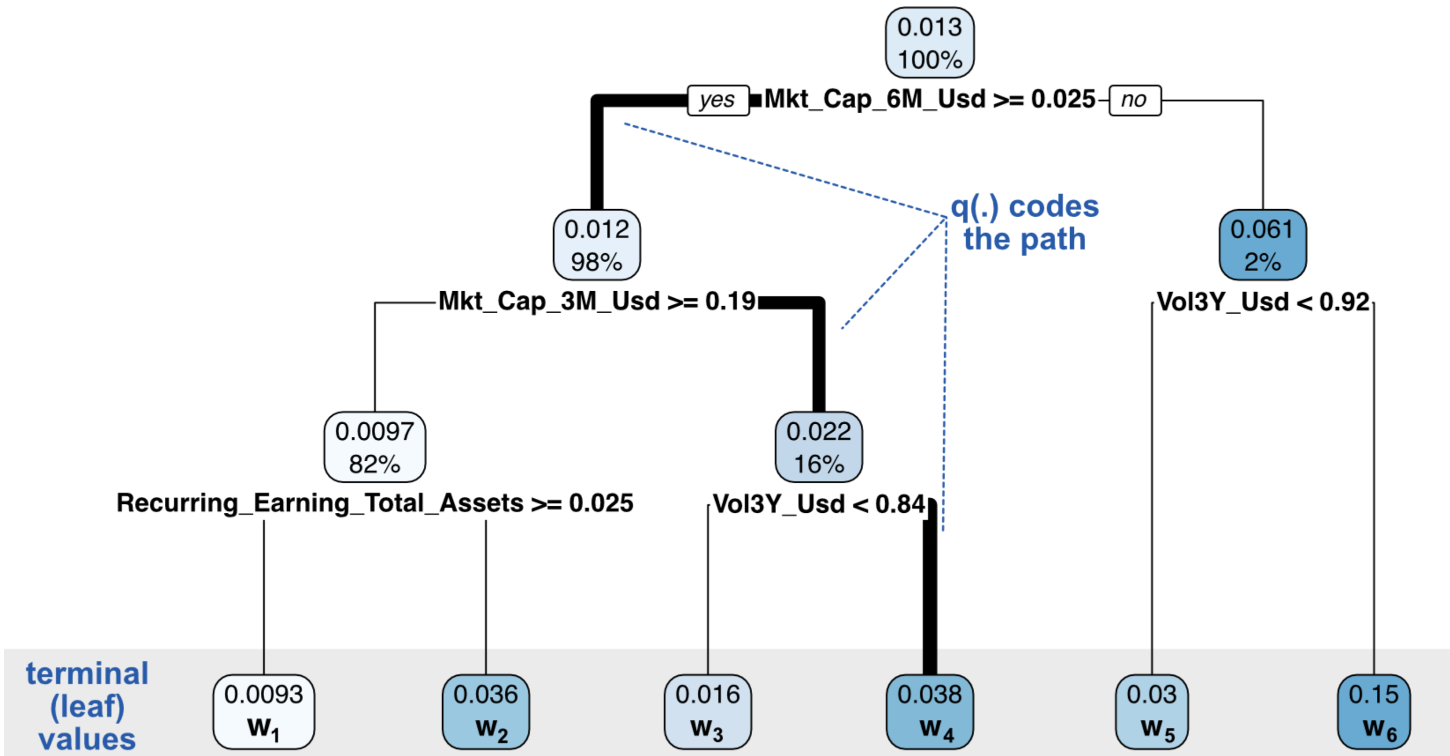
Simple example:

- Assuming independent classifiers
- Classifier has an error rate $\varepsilon < 0.5$
  - Ensemble prediction better than random guess
- If $\varepsilon > 0.5$ for each classifier, ensemble wrong prediction will increase



Source: Raschka, Sebastian. Python Machine Learning (p. 202). Packt Publishing.

Boosted Tree example



Source: "Machine Learning for Factor Investing" Coqueret., Guida (2020 Chapman & Hall)

Measuring the Quality of a ML model

	OUTPERFORMED	UNDERPERFORMED
OUTERPERFORM	**True Positive:** Stock WAS classified as outperforming and DID outperformed	**False negative:** Stock was NOT classified as outperforming and DID outperformed
UNDERPERF.	**False Positive:** Stock WAS classified as outperforming and did NOT outperformed	**True Negative:** Stock was NOT classified as outperforming and DID not outperformed

- **Left axis (vertical)** of the matrix shows **Actual**
- **Top axis (horizontal)** shows **predicted**

Beyond confusion matrix

- **Fp** : false positive. Stock predicted to outperform and that did not outperform out of sample.
- **Fn** : false negative. Stock predicted to underperform that outperform out of sample.
- **Tp**: true positive. Stock predicted to outperform which outperform out of sample.
- **Tn**: true negative. Stock predicted to underperform which underperform out of sample.

<u>Precision</u>: **Tp / (Tp + Fp)**
Precision could be defined as a rate of successful prediction for sector neutral outperforming stocks.

<u>Recall</u>: **Tp / (Tp + Fn)**
Recall could be defined as a true rate, since we include the instances that have been wrongly classified in negative.

<u>Accuracy</u>: **(Tp + Tn) / (Tp+Tn+Fp+Fn)**
This is the accuracy level used in the cross validation part.

<u>F1 score</u>: **2 * (Precision * Recall / Precision + Recall)**

[LT ML predictions for EQ]

dataset & E.D.A

Objective, data and protocol

- We will compare different labels corresponding to different prediction horizon for cross sectional returns
  - (1M, 3M, 6M, 9M, 12M, 18M, 36M)
- Investment universe is **US stocks (~1500)**
- **Full** dataset from Dec-1999 until Dec-2019
- **(~ 100) features**, monthly normalised in percentrank.
- Dataset pre-processed, outliers removed, focussing on training on the tails of the distribution **(top and bottom 25%)** excluding the top 1% **avoiding to train on high vol.**
- Split the dataset between **Training** (80%) and **Testing** (20%)
- Rolling window of **60 months**

Features engineering: Training on tails

Springer Link

Original Research | Published: 20 February 2020

Training trees on tails with applications to portfolio choice

Guillaume Coqueret & Tony Guida

Annals of Operations Research (2020) | Cite this article

Abstract

In this article, we investigate the impact of truncating training data when fitting regression trees. We argue that training times can be curtailed by reducing the training sample without any loss in out-of-sample accuracy as long as the prediction model has been trained on the *tails* of the dependent variable, that is, when 'average' observations have been discarded from the training sample. Filtering instances has an impact on the features that are selected to yield the splits and can help reduce overfitting by favoring predictors with monotonous impacts on the dependent variable. We test this technique in an out-of-sample exercise of portfolio selection which shows its benefits. The implications of our results are decisive for time-consuming tasks such as hyperparameter tuning and validation.

FDP INSTITUTE™

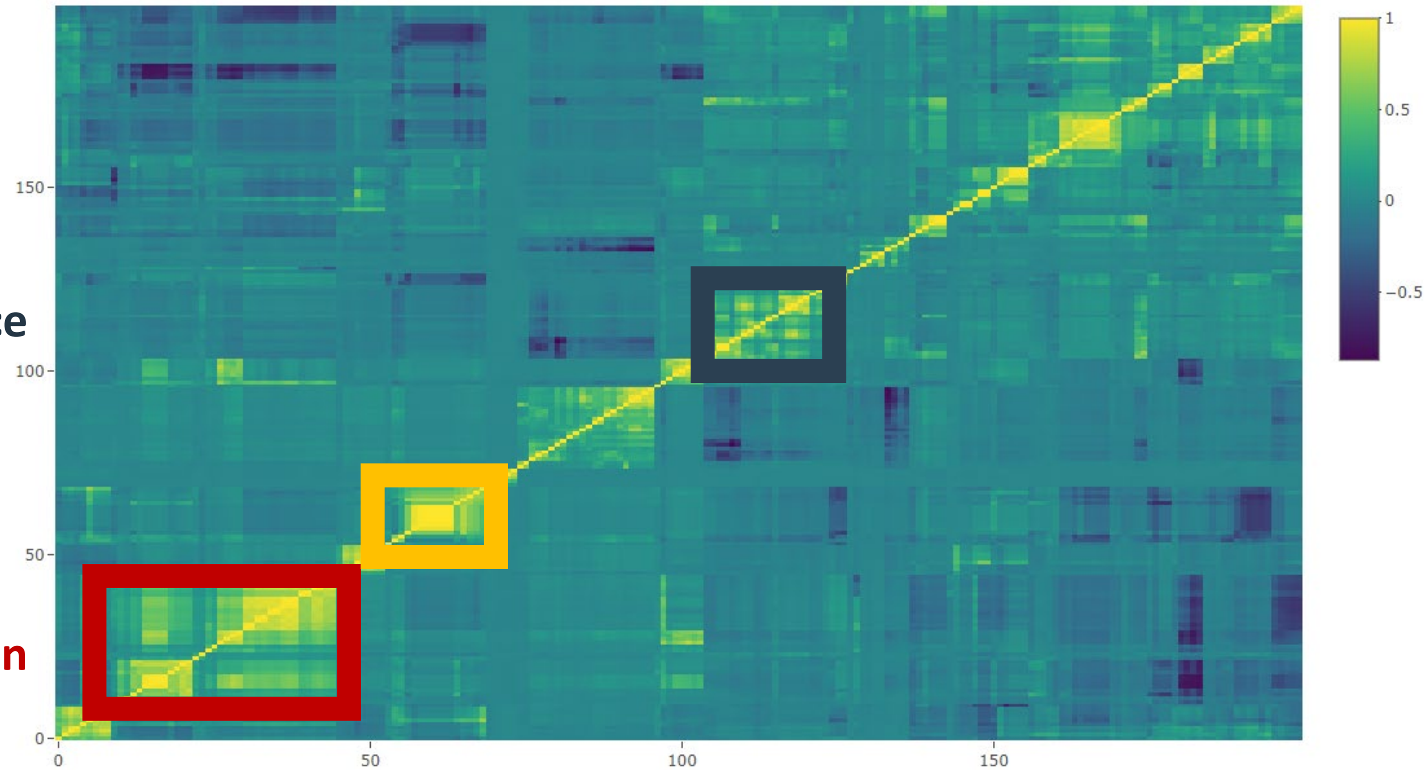
ram ACTIVE INVESTMENTS

Features correlation example

Risk Price based

Volume

Valuation



Source: Guida, Coqueret. Chapter 7, Ensemble Learning Applied to Quant Equity – Big Data and Machine Learning in Quantitative Investment

Creating the dataset

Instances >>>> ~ 600 000

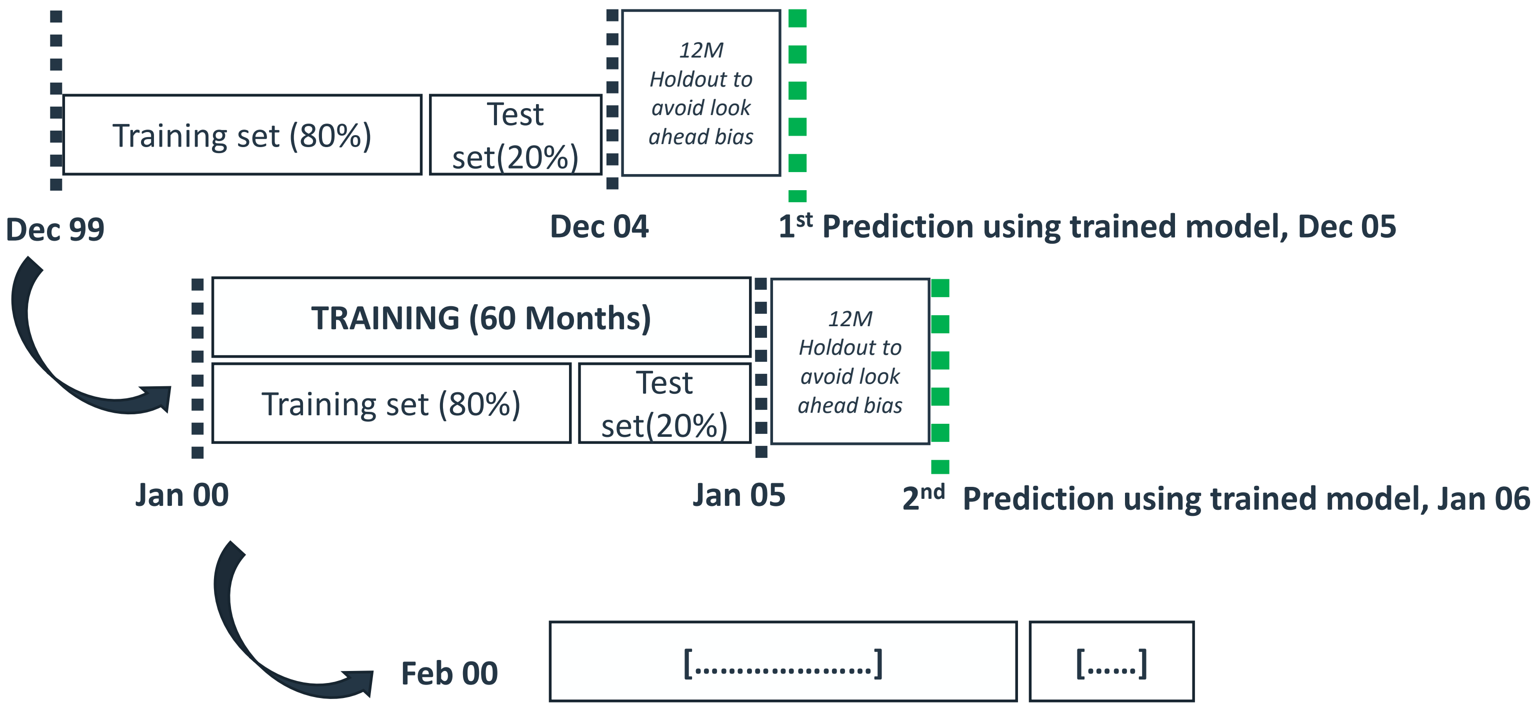
	DataDate	dateReturnPerf	SecID	Fact2	Fact3	Fact4	Fact7	Fact8	Fact9	Fact10	Fact11	Fact12	Fact13	Fact14	Fact16	Fact17	Fact18	Fact20	Fact21	Fact22	Fact23	Fact25	Fact26	Fact27
1	2010-12-31	2011-12-31	1195515867	19	70	94	94	13	99	15	34	55	65	42	32	28	21	27	31	49	38	53	53	76
2	2014-08-31	2015-08-31	570191681	44	NULL	83	48	11	7	19	72	72	28	33	32	31	21	30	22	31	51	13	14	49
3	2012-01-31	2013-01-31	290849864	55	66	98	15	3	100	7	2	2	58	43	14	38	25	19	22	11	11	10	11	65
4	2012-02-29	2013-02-28	324622763	69	69	99	27	4	100	7	2	1	65	42	14	39	16	18	21	6	6	13	15	62
5	2012-08-31	2013-08-31	1528063821	73	70	99	61	4	99	14	2	2	47	41	19	38	16	25	26	43	45	83	84	50
6	2012-03-31	2013-03-31	1850474528	61	67	99	27	4	100	11	2	1	60											
7	2012-11-30	2013-11-30	2021474168	75	74	99	60	4	NULL	13	2	10	47											
8	2017-01-31	2018-01-31	408403206	42	31	61	68	10	2	26	94	92	14											
9	2013-01-31	2014-01-31	1593812596	65	73	99	53	4	NULL	18	3	10	47											
10	2016-05-31	2017-05-31	352206094	NULL	NULL	NULL	NULL	NULL	74	NULL	77	77	33											
11	2009-03-31	2010-03-31	1251799406	55	65	49	30	10	100	11	3	1	83											
12	2007-02-28	2008-02-29	1092089186	36	69	21	97	8	90	10	74	63	33											
13	2016-09-30	2017-09-30	1171513832	30	32	60	67	10	100	27	89	83	13											
14	2003-07-31	2004-07-31	1668572152	48	30	70	35	8	91	30	3	1	41											
15	2010-08-31	2011-08-31	1317881264	20	4	87	96	15	99	10	29	56	62											
16	2008-02-29	2009-02-28	31085621	46	81	19	24	7	63	5	77	62	51											
17	2009-04-30	2010-04-30	512957258	52	62	50	32	10	100	9	3	1	84											
18	2012-04-30	2013-04-30	2143460900	67	68	99	27	4	100	10	2	1	60											
19	2016-10-31	2017-10-31	1415589206	20	NULL	NULL	64	11	58	35	79	73	15											
20	2011-01-31	2012-01-31	486144046	21	70	95	95	13	99	12	35	55	65											
21	2013-03-31	2014-03-31	156902714	54	64	99	67	1	100	19	5	35	36											
22	2009-08-31	2010-08-31	290508352	29	3	53	94	18	100	19	1	1	71											
23	2010-09-30	2011-09-30	1527687499	20	4	87	96	15	99	12	29	56	62											
24	2016-06-30	2017-06-30	2006847672	NULL	NULL	NULL	NULL	NULL	80	NULL	75	74	23	NULL	41	56	37	30	26	74	64	87	82	50
25	2007-12-31	2008-12-31	156190601	47	84	12	83	7	100	8	90	77	53	72	57	60	62	64	72	31	30	18	15	29
26	2014-07-31	2015-07-31	1573508350	48	NULL	89	84	12	6	19	71	71	31	32	31	32	25	31	23	35	61	51	61	53
27	2007-06-30	2008-06-30	1049253878	38	70	26	97	6	96	10	81	70	32	18	67	23	79	75	75	34	42	71	71	30
28	2007-03-31	2008-03-31	400280762	36	70	21	97	8	90	11	74	62	32	17	68	53	81	81	74	34	51	71	77	30
29	2016-08-31	2017-08-31	1510213818	32	30	60	70	10	100	24	89	83	14	23	44	58	29	36	28	83	56	94	88	51
30	2011-02-28	2012-02-29	231369713	19	65	95	58	15	100	11	23	45	63	48	31	34	26	24	31	57	59	60	62	78
31	2008-04-30	2009-04-30	474472098	45	81	20	24	6	65	6	78	62	51	77	54	34	53	59	67	14	10	38	26	26

Some features examples

- *Fundamental trailing*
- *Price based*
- *Volume based*
- *Risk based*
- *Composites*

Rolling Windows for training (case for 12M forward)

In this example we use a rolling window of **60 months** to predict the **12M forward performance** of a stock.



[LT ML predictions for EQ]
Building & Training models

Hyperparameters:

- **The learning rate**, η: it is the step size shrinkage used in update to prevents overfitting. After each boosting step, we can directly get the weights of new features and η actually shrinks the feature weights to make the boosting process more conservative.
- **The maximum depth**: it is the longest path (in terms of node) from the root to a leaf of the tree. Increasing this value will make the model more complex and more likely to be overfitting.
- **Regression $\boldsymbol{\lambda}$**: it is the L^2 regularization term on weights (mentioned in the technical section) and increasing this value will make model more conservative.
- **gamma:** minimum loss reduction required to make a further partition on a leaf node of the tree. The larger, the more conservative the algorithm will be.

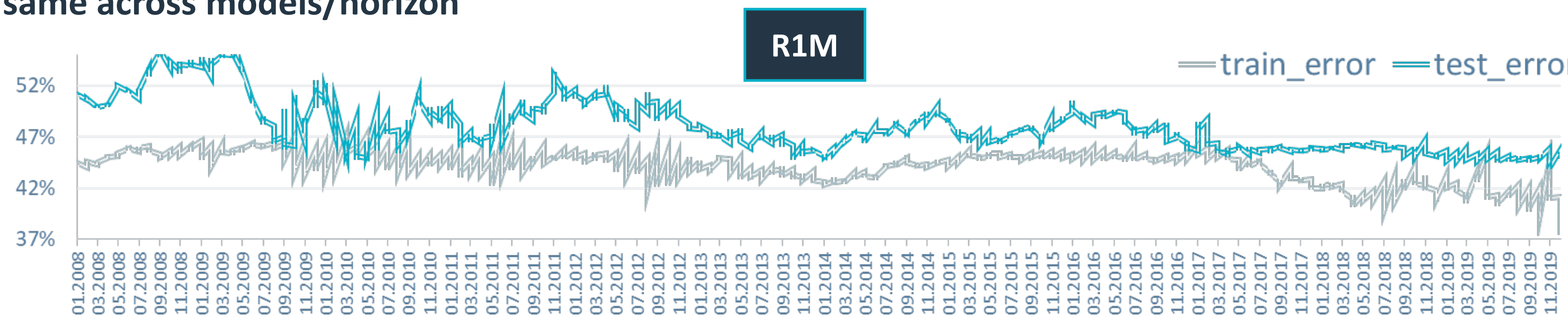
model	max_depth	eta	round	eval_metric	subsample	col_by_sample
XGB	5	1%	150	error	0.8	0.8

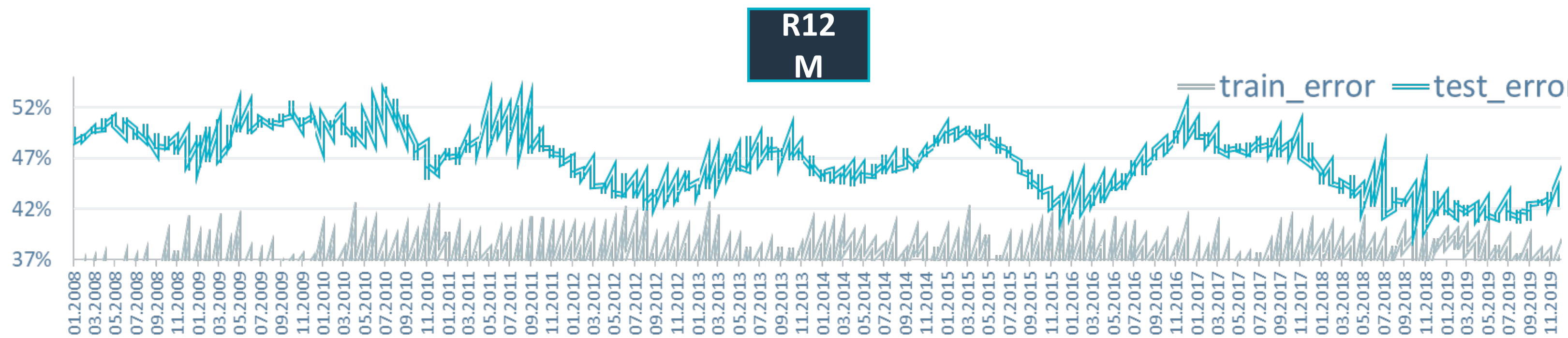
LT vs the rest: impact on training

We compare **the accuracy** in training and test for each rebalancing. **Training parameters are kept the same across models/horizon**

LT vs the rest: impact on training

We compare **the accuracy** in training and test for each rebalancing. **Training parameters are kept the same across models/horizon**





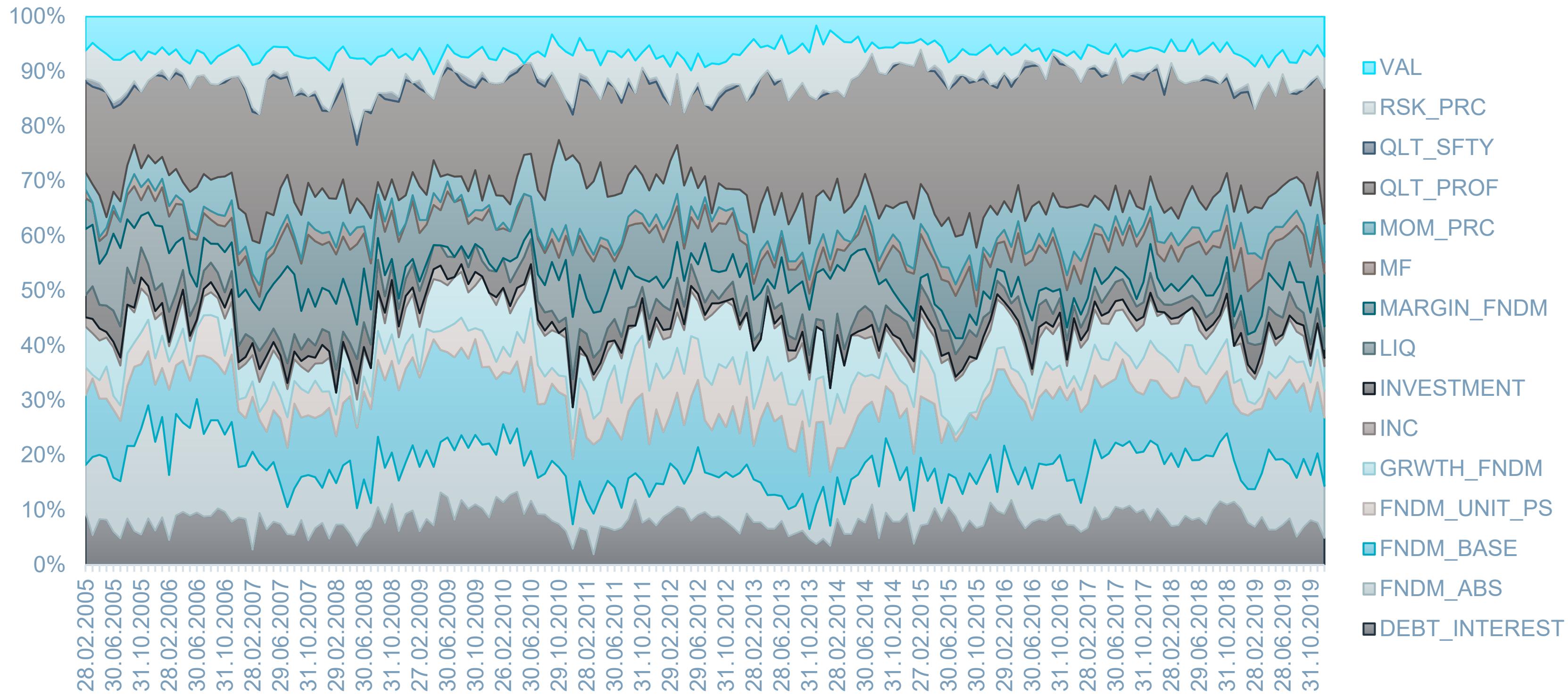
FDP INSTITUTE

ram ACTIVE INVESTMENTS

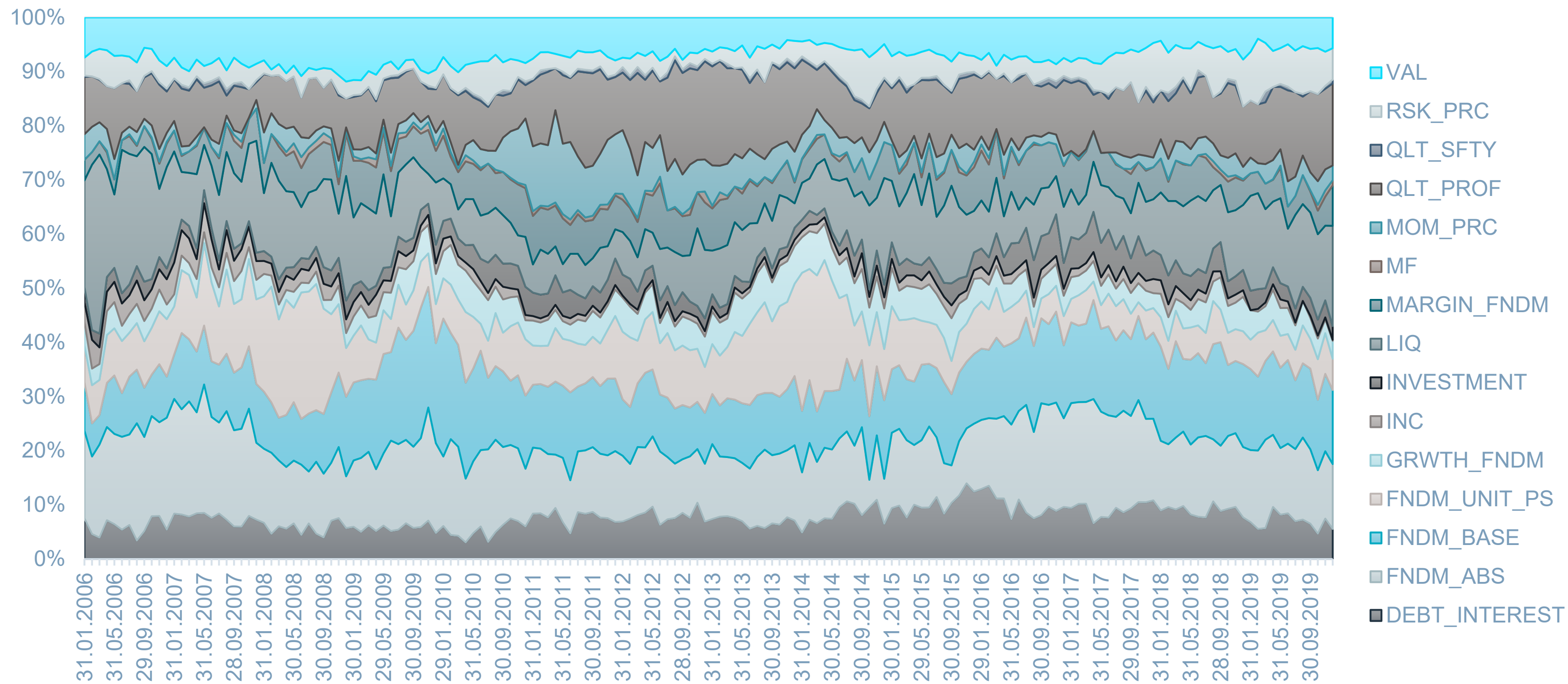
Training model: quality measures

Model	[train]	[test]
R1M	43.7%	47.6%
R3M	40.7%	48.0%
R6M	38.9%	47.7%
R9M	37.6%	46.5%
R12M	36.1%	46.4%
R18M	32.8%	46.2%
R24M	30.6%	42.8%
R36M	27.4%	40.3%

Interpretability breakdown – 1M preds.



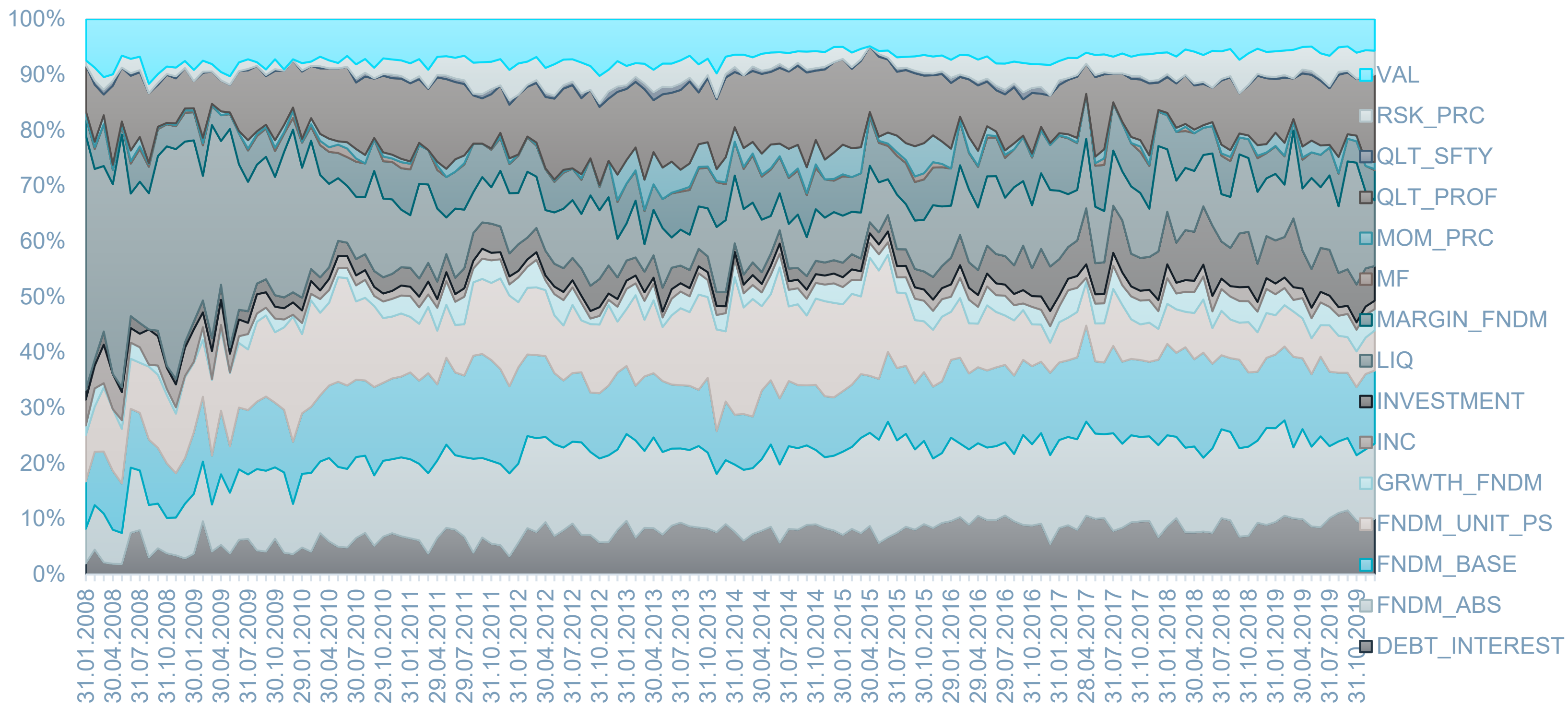
Interpretability breakdown – 12M preds.



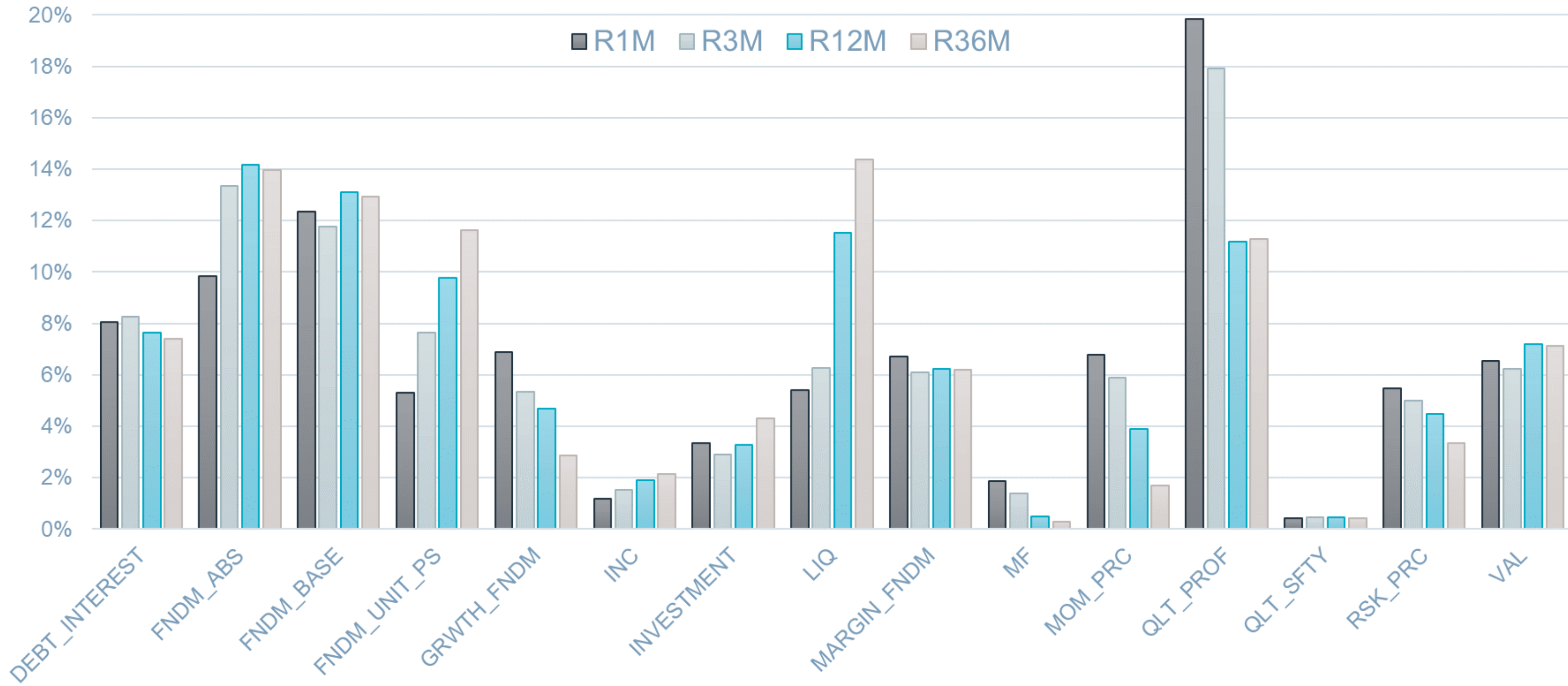
FDP INSTITUTE

ram ACTIVE INVESTMENTS

Interpretability breakdown – 36M preds.



Interpretability: simple avg feature importance



FDP INSTITUTE

ram ACTIVE INVESTMENTS

[LT ML predictions for EQ]

Analysing portfolios results

Decile performance's analysis: monotonicity

Avg annual net performance: net of TC gross of mc	**R1M**	**R3M**	**R6M**	**R9M**	**R12M**	**R18M**	**R24M**	**R36M**
D1	9.06%	11.98%	11.38%	11.77%	10.09%	10.60%	10.07%	9.61%
D2	8.64%	11.59%	11.67%	12.03%	12.53%	11.94%	12.33%	11.25%
D3	9.08%	10.28%	9.76%	12.39%	12.52%	12.38%	14.01%	12.57%
D4	9.89%	11.39%	10.37%	11.69%	13.40%	10.42%	14.73%	13.44%
D5	12.44%	12.61%	12.54%	12.39%	12.12%	13.71%	14.27%	13.49%
D6	11.73%	13.61%	13.68%	13.10%	11.90%	14.97%	16.25%	15.03%
D7	11.74%	13.58%	12.17%	13.93%	13.28%	15.00%	17.02%	15.19%
D8	11.61%	13.39%	13.10%	11.96%	15.41%	16.86%	19.33%	18.17%
D9	11.93%	15.30%	16.17%	16.39%	17.27%	17.89%	22.42%	21.39%
D10	13.20%	20.00%	20.28%	21.69%	23.47%	25.60%	27.20%	26.49%

FDP INSTITUTE™

ram ACTIVE INVESTMENTS

Decile turnover's analysis: look for the tails...

avg monthly turnover (buy + sell)	R1M	R3M	R6M	R9M	R12M	R18M	R24M	R36M
D1	63.7%	48.2%	43.7%	41.2%	39.6%	37.2%	35.5%	32.4%
D2	80.9%	71.4%	67.2%	64.6%	63.6%	59.7%	58.9%	55.1%
D3	84.8%	77.7%	74.0%	70.8%	70.2%	67.8%	67.8%	64.7%
D4	86.9%	80.7%	77.1%	73.9%	73.6%	72.1%	70.8%	68.4%
D5	86.9%	81.0%	78.5%	75.5%	75.1%	73.9%	72.0%	69.7%
D6	86.9%	81.6%	78.5%	75.4%	74.7%	73.3%	72.9%	69.3%
D7	86.0%	80.7%	77.0%	73.5%	72.7%	72.7%	71.8%	67.7%
D8	83.5%	78.5%	73.4%	70.0%	68.9%	69.0%	67.9%	64.8%
D9	80.3%	72.9%	67.6%	64.3%	63.2%	61.6%	60.3%	57.9%
D10	62.3%	53.1%	47.8%	45.5%	44.8%	42.1%	41.0%	39.2%

Comparison accross portfolios

from Feb 08 until Dec 19	Avg perf p.a. Net of tc (USD)	Vol p.a.	risk/perf ratio	Turnover avg monthly (B+S)	avg annual trading cost
D10 port R1M	13.2%	**12.34%**	1.07	62%	1.87%
D10 port R3M	20.0%	16.51%	1.21	53%	1.59%
D10 port R6M	20.3%	17.72%	1.14	48%	1.43%
D10 port R9M	21.7%	18.68%	1.16	46%	1.37%
D10 port R12M	24.5%	18.58%	1.32	45%	1.34%
D10 port R18M	25.6%	19.17%	**1.34**	42%	1.26%
D10 port R24M	**27.1%**	20.44%	1.33	41%	1.23%
D10 port R36M	26.5%	20.70%	1.28	**39%**	**1.18%**
Universe EW	13.4%	12.03%	1.11	NA	NA
SP500	9.8%	14.90%	0.66	NA	NA

FDP INSTITUTE

ram ACTIVE INVESTMENTS

Conclusion

[1] Machine learning is not new but a “**new**” way for doing research today.

[2] ML used with traditional data proved to add a non-linear adaptative component to alpha prediction

[3] Long term predictions seems to give higher risk-adjusted performance with less turnover than the usual 1M forward horizon.

ram

ACTIVE INVESTMENTS

Important Information

FDP INSTITUTE™

ram ACTIVE INVESTMENTS

FDP Curriculum

1. Introduction to Data Science & Big Data
2. DM & ML: Introduction
3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models
4. DM & ML: Classification & Clustering
5. DM & ML: Performance Evaluation, Backtesting & False Discoveries
6. DM & ML: Representing & Mining Text
7. Big Data, DM & ML: Ethical & Privacy Issues
8. Big Data and Machine Learning in the Financial Industry

Source: FDP Institute Study guide March 2020 Exam

***Sample of the* Reading(s):**

Guida, T. (2019). Big Data and Machine Learning in Quantitative Investments. West Sussex, UK: John Wiley & Sons Ltd.

- Topic 1 – Reading 1.4: Chapters 2, 4 & 5.
- Topic 8 - Reading 8.9 : Chapter 10.

Sample Keywords *(of the Guida reading):*

Mainstream (p. 336)
Part of Speech Tagging (p. 349)
Primary source (p. 336)
Stemming (p. 350)
Social media (p. 337)
Lemmatization (p. 350)
Sentiment analysis (p. 339)
Naïve Bayes (p. 355)
Natural language processing (p.347)
FNN (p. 363)
Tokenization (p. 348)
RNN (p. 363)
Word filter (p. 348)
CNN (p. 363)

FDP INSTITUTE

FDP Curriculum

1. Introduction to Data Science & Big Data

2. DM & ML: Introduction

3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models

4. DM & ML: Classification & Clustering

5. DM & ML: Performance Evaluation, Backtesting & False Discoveries

6. DM & ML: Representing & Mining Text

7. Big Data, DM & ML: Ethical & Privacy Issues

8. Big Data and Machine Learning in the Financial Industry

Source: FDP Institute Study guide March 2020 Exam

Sample Learning Objectives *(provided for reading 8.9.1)*

Demonstrate proficiency in the following areas:

8.9.1 Natural language processing of financial news. For example:

A. Describe the three categories of sources of news data.
B. Explain the advantages and disadvantages of using the new category of social media.
C. Describe sentiment analysis.
D. Describe the word list approach to sentiment analysis.
E. Describe the three challenges associated with sentiment analysis.
F. Describe the four steps — pre-processing, feature representation, inference and evaluation — in applying NLP to texts.
G. Understand aspects of pre-processing: tokenization, vocabulary, part of speech, stemming and lemmatization.
H. Understand aspects of representation of words as features: bag of words, N-gram, distributed representation

Sample Question:

According to "Natural Language Processing of Financial News," by Sesen et al., what is the description of a "word list" approach to sentiment analysis?

a) Words appearing in an article are manually labeled as positive or negative
b) A data set that associates words with different sentiments is created
c) The predictive power of a news item is used to assign sentiment labels to words

Answer: b

Source: LO 8.9.1, Reading 8.9, pp 340-341

FDP INSTITUTE

Q & A

Kind reminders of upcoming webinars as we go through the Q & A.
Add your questions in the chat room please.

WEBINAR SERIES
A Conversation With…

Ganesh Mani
Adj. Faculty Carnegie Mellon
“Data Supply Chain Mgmt.”

March 10, 2020
1pm EST

WEBINAR SERIES
A Conversation With…
Rick Roche, CAIA
Man. Dir. Little Harbor Advisors
“Evolution of Machine Learning in Investment Mgmt.”
March 17 - 11am EST

WEBINAR SERIES
A Conversation With…
Michael Oliver Weinberg
Managing Dir. , Head of Hedge Funds & Alt. Alpha
APG
Autonomous Learning Investment Strategies
March 25th – 4pm EST

WEBINAR SERIES
A Conversation With…

George Mussalli, CFA
PanAgora

Mike Chen, Ph.D.
PanAgora

“An integrative Approach to Quantitative ESG Investing”
April 1, 2020 @ 12noon

In Closing

- Registration for the October 26 – November 8th exam opens May 10th
- For a recent candidate webinar go to www.fdpinstitute.org/webinars

Learn more about the FDP Institute at www.fdpinstitute.org