Machine Learning for Stock Selection

Keywan Rasekhschaffe, PhD

Robert C. Jones

Agenda

- Issues of overfitting and maximizing the signal to noise ratio
- Data prep and feature engineering: often overlooked
- Evaluating your algorithm choice: what do you want to achieve?
- Pitfalls

Rasekhschaffe, Jones (2019) provide a concrete example

What is Machine Learning

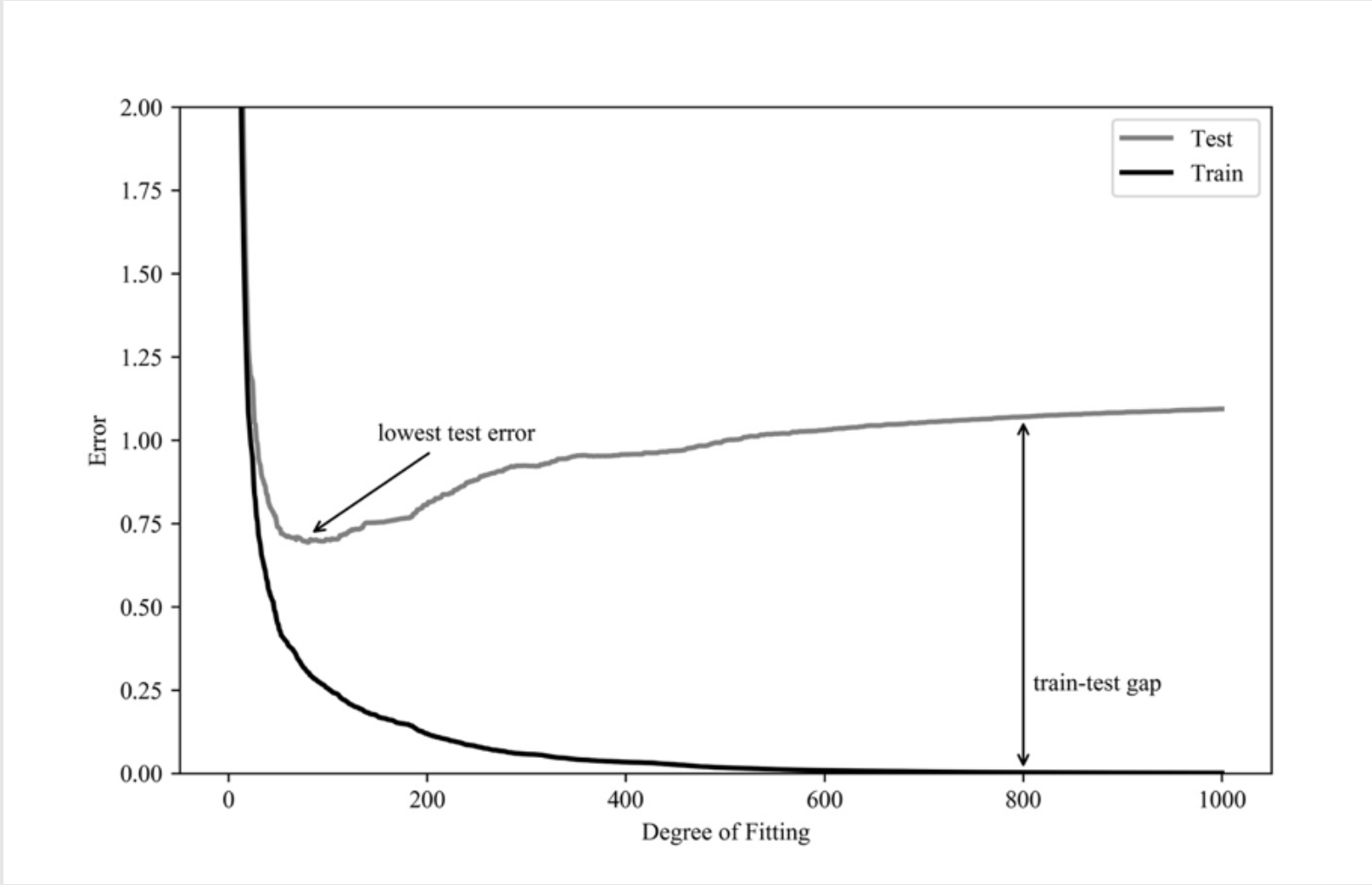
- **A branch of computer science concerned with the design of algorithms that learn from examples to make predictions**
- **Examples:**
  - Neural Networks, Deep Learning, CART, Random Forests, Gradient Boosting, Support Vector Machines
- **Many potential benefits:**
  - Can recognize complex, often non-linear patterns in large data sets
  - Many techniques work well with correlated features
  - Self correcting with additional training
  - Can detect dynamic relationships and is robust to factor decay
  - Potentially Less crowded (smart alpha)
- **One BIG problem:**
  - Overfitting

Two Ways to Overcome Overfitting

- **Feature engineering**
  - Use domain knowledge to increase the signal-to-noise ratio
  - Requires expertise in finance, statistics and data
- **Forecast ensembles**
  - The wisdom of crowds
  - Different algorithms
  - Different training sets
  - Different forecast horizons
  - Use averaging to cancel offsetting biases
  - Result: more signal, less noise

Overfitting and Underfitting

- More complexity initially leads to improved out-of-sample performance, but then results deteriorate
- Not optimal to overfit or underfit
- Train-test gap widens with increasing complexity: never judge results in-sample



Feature Engineering

- What problems do we ask the algo to solve?
- Incorporates domain knowledge
- Seldom discussed, but at least as important as prediction algorithms
- What type of data do we allow it to consider?
- Success requires expertise in investing, data and statistics

Examples of Feature Engineering

Investment Expertise	▪ How to classify stocks (region, industry, style) ▪ How to define success (returns, excess returns, alphas) ▪ Which factors to consider
Data Expertise	▪ Factor calculations ▪ Adjustments for errors and missing data ▪ Proper lags and adjustments to avoid look-ahead bias
Statistical Expertise	▪ Clustering ▪ Normalization ▪ Dimensionality reduction

Ensembles: Combing Diverse ML Forecasts

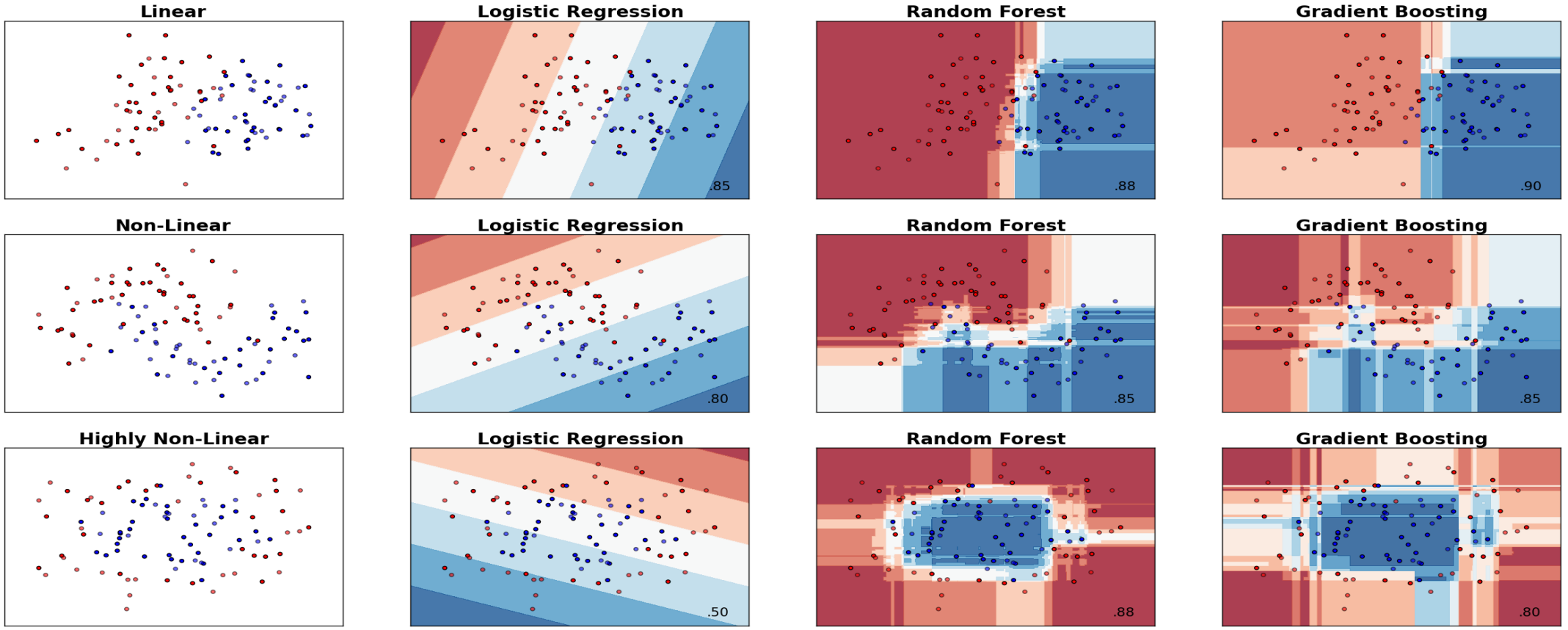
Diversification (I)	▪ Forecast combinations from **different algorithms** ▪ Different algos detect different attributes in the data
Diversification (II)	▪ Forecast combinations from **different forecast horizons** ▪ Different factors are important for different forecast horizons ▪ Technical factors are more important for shorter horizons ▪ Fundamental factors are more important for longer horizons
Diversification (III)	▪ Forecast combinations from **different training windows** ▪ Different factors are impotant in different market environments ▪ Training widows based on time, seasonality, or economic / market conditions

Sample Algorithms

Gradient Boosted Classification and Regression Trees	▪ Ensemble technique that combines weak learners (algorithms) into a strong learner (ensemble)
Random Forests	▪ Ensemble technique that overcomes overfitting by equal weighting forecasts from many trees
Deep Learning	▪ Recent evolution of neural network (back propogation) algorithms ▪ Less likely to overfit
Support Vector Machines	▪ Learns by analogy ▪ Efective in high dimensional spaces

Why Use Multiple Algorithms?

Different algorithms detect different attributes in the data



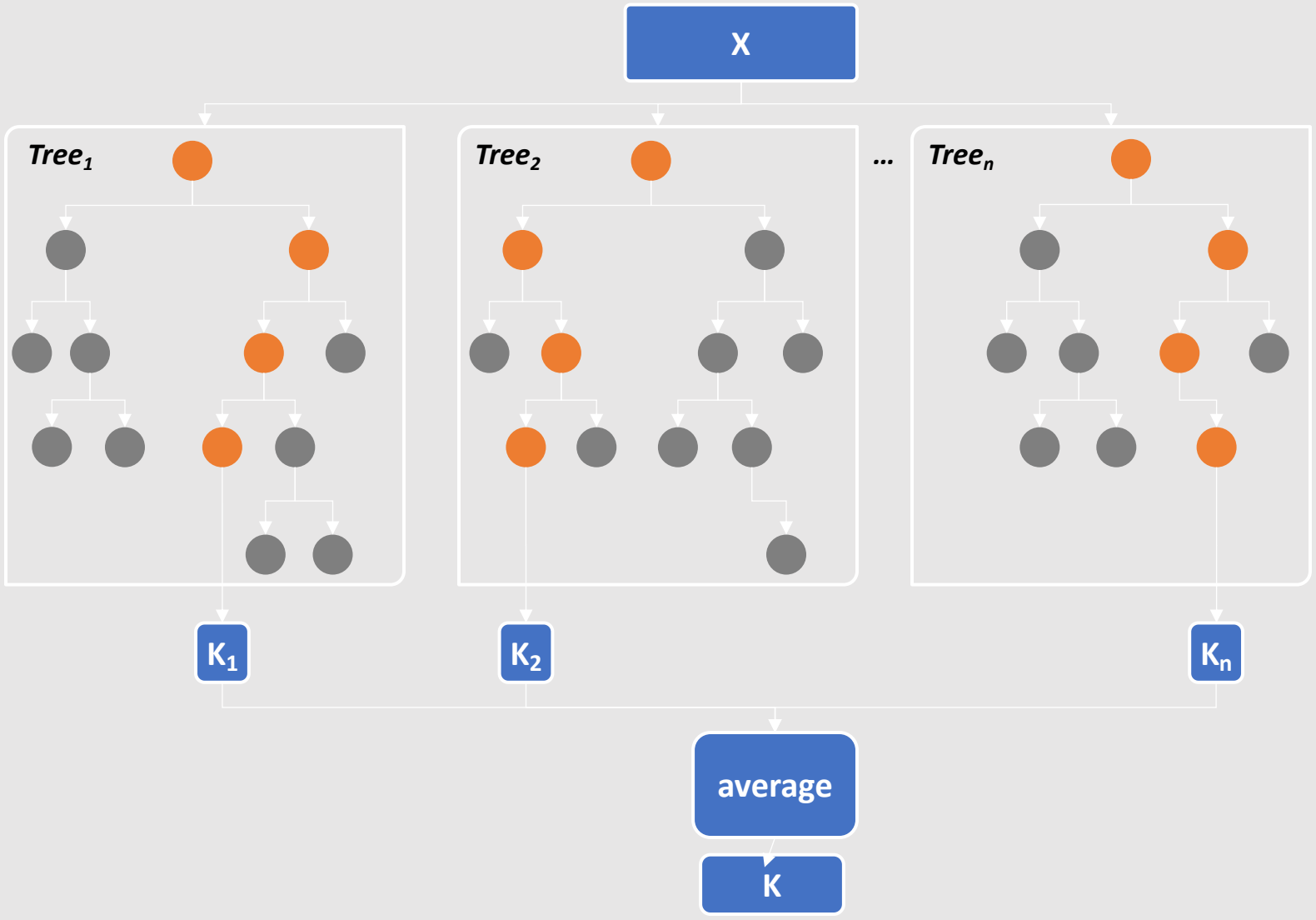
CART Models

- **CART** = **C**lassification **A**nd **R**egression **T**ree
- Classifies data into binary branches via hierarchical splits
- Easy to overfit
- “Pruning” can reduce overfitting
- Important building block for ensemble algos
- Although still popular in finance, single trees are no longer widely used in serious ML applications

Random Forests

- Increase number of trees and construct a forest
- Randomly selects subsets of features
- Ensemble “bagging” (bootstrap aggregating) algorithm
- Uses majority decision or average of base estimators
- Model variance decreases

Random Forest Diagram



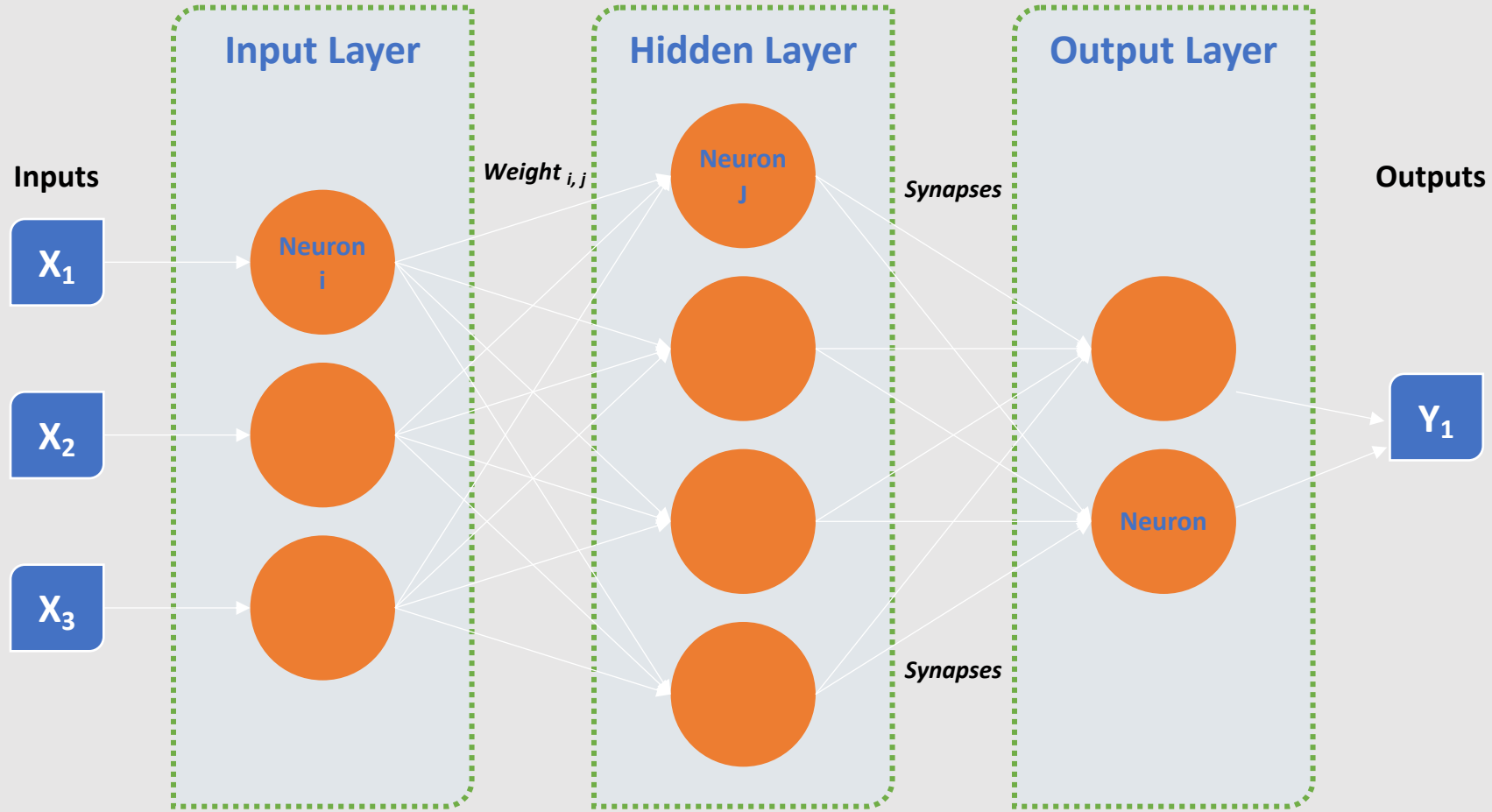
Boosting

- Combines weak learners (factors or simple models) into a strong learner (ensemble model)
- Ensemble technique like Random Forests
- Iteratively re-weight observations
- Put more emphasis on misclassified observations
- Each subsequently chosen weak learner gets assigned a decreasing weight
- Base learners can be trees, decision stumps, etc.
- Examples are AdaBoost and **Gradient Boosted Regression Trees**

Artificial Neural Networks

- Interconnected group of “neurons”
- Inspired by biology
- Input layer <---> hidden layer <---> output layer is the most common architecture
- Shallow architectures work quite well for forecasting returns

Artificial Neural Network Example



Deep Learning

- We can avoid overfitting using regularization and randomization techniques
- Dropout: randomly disable neurons in the learning process
- Forces the network to generalize by learning multiple representations of the same data
- Gold standard for computer vision / audio
- The algorithm (sometimes) takes care of feature engineering
- Requires carefully tuned architecture / parameters
- Can be computationally expensive

Support Vector Machines

- Finds hyperplane with largest distance between classes in the training data
- Can use linear or non-linear kernel
- Linear kernel related to logistic regression, but more effective in high-dimensional spaces
- “Soft margin” often used in practice when classes aren’t perfectly separable
- Flexible and powerful, but often slow

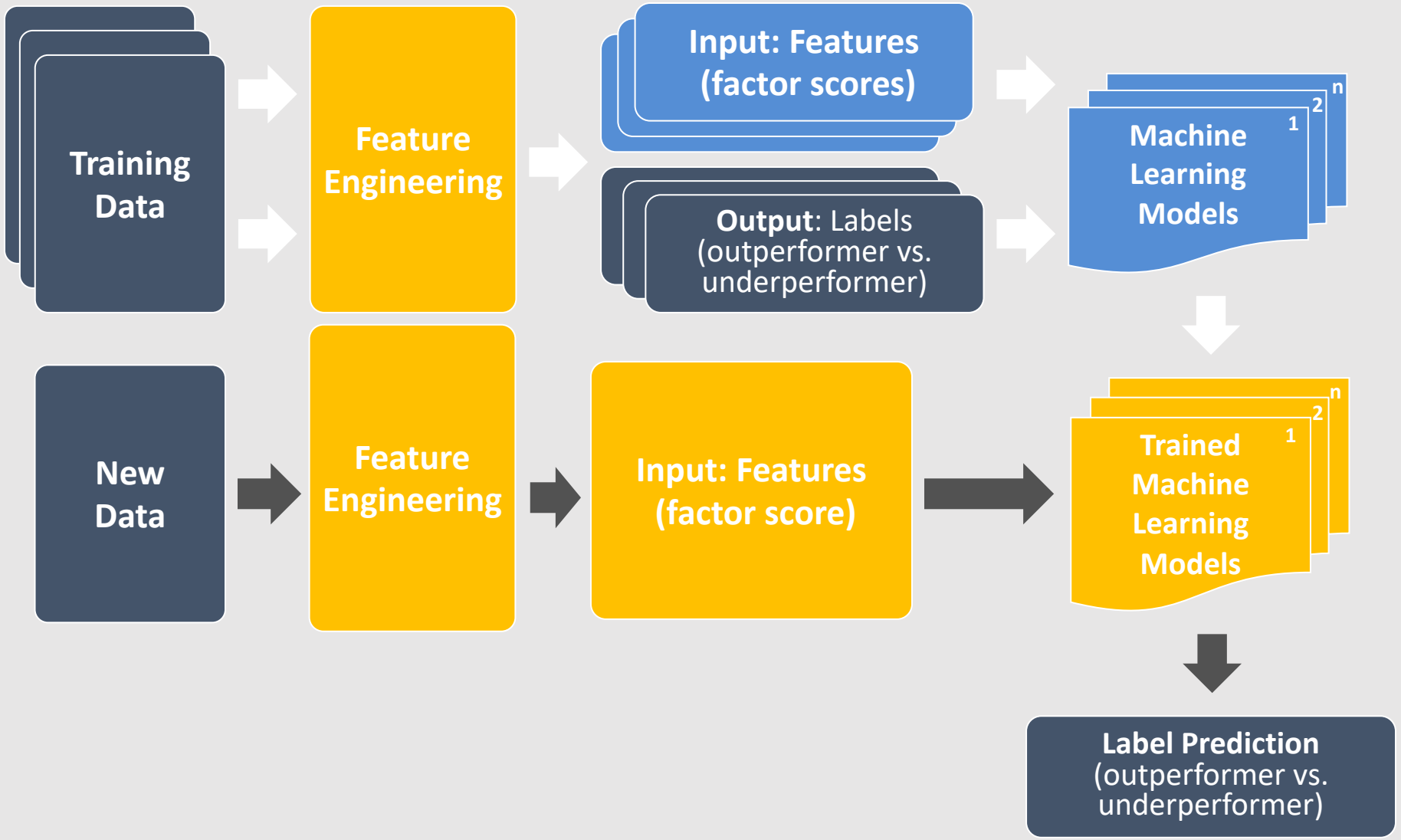
Forecast Combinations and Diversification

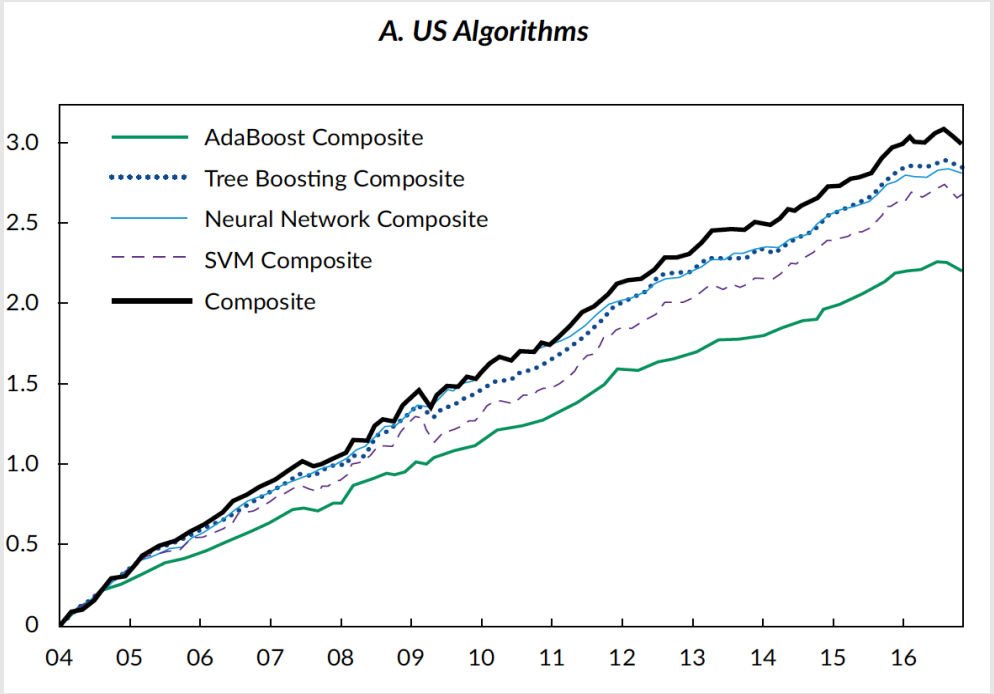
- Algos
- Training windows
- Forecast horizons
- Factor library (or libraries)

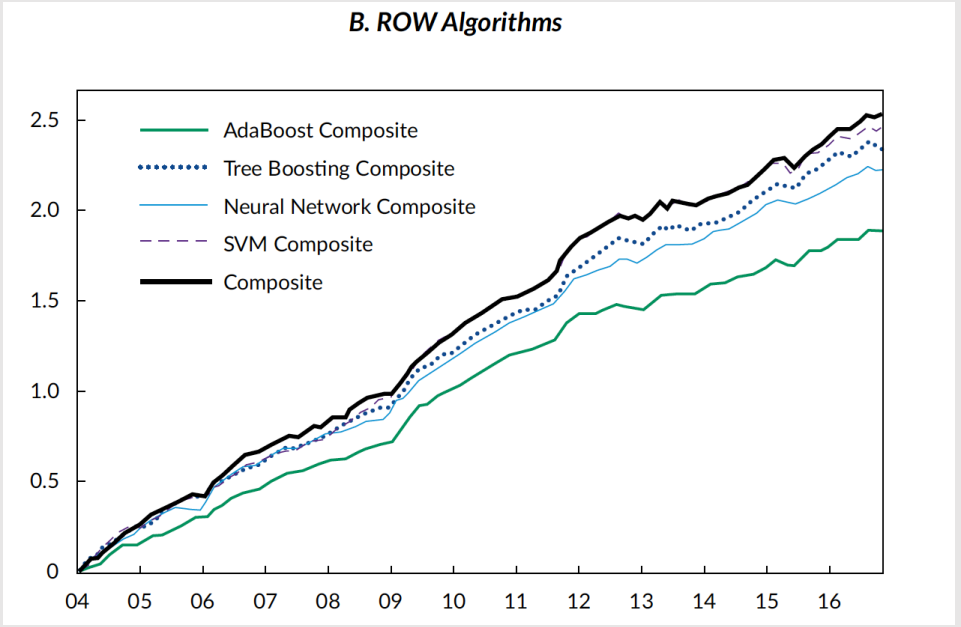
Different Training Windows

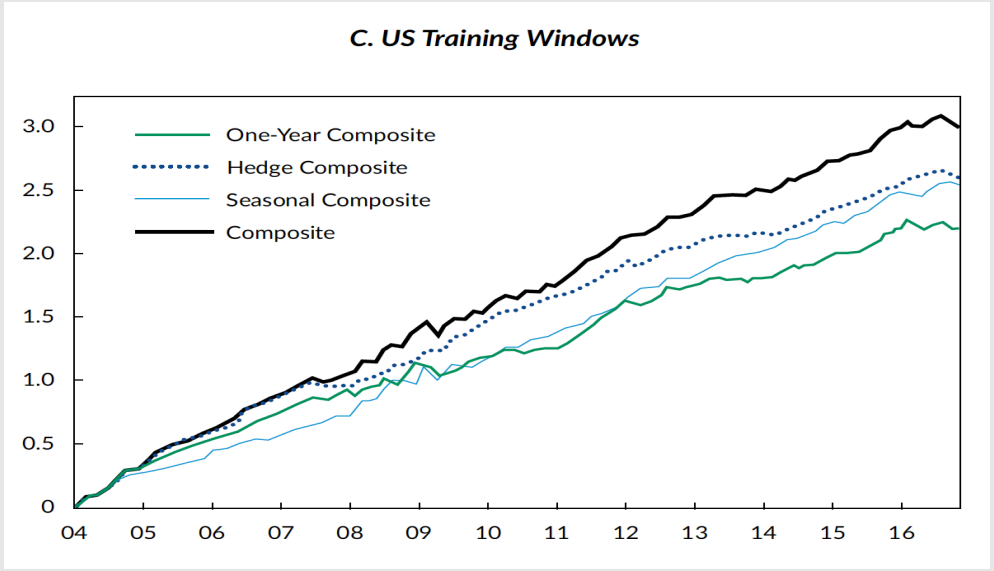
- Best to train algos using data from periods that are “similar” to the forecast period
- **Recency**: More recent data likely to be more relevant
- **Seasonality**: Data from same season (month) likely to be more relevant
- **Context**: Data from periods with similar macro conditions likely to be more relevant
  - ✓ Similar risk environment (VIX, credit spreads)
  - ✓ Similar monetary conditions (level of short rates, slope of the yield curve)
  - ✓ Similar valuation levels (dividend yield, E/P)
- **Hedging**: Data from periods where other forecasts are noisy are likely to improve overall accuracy

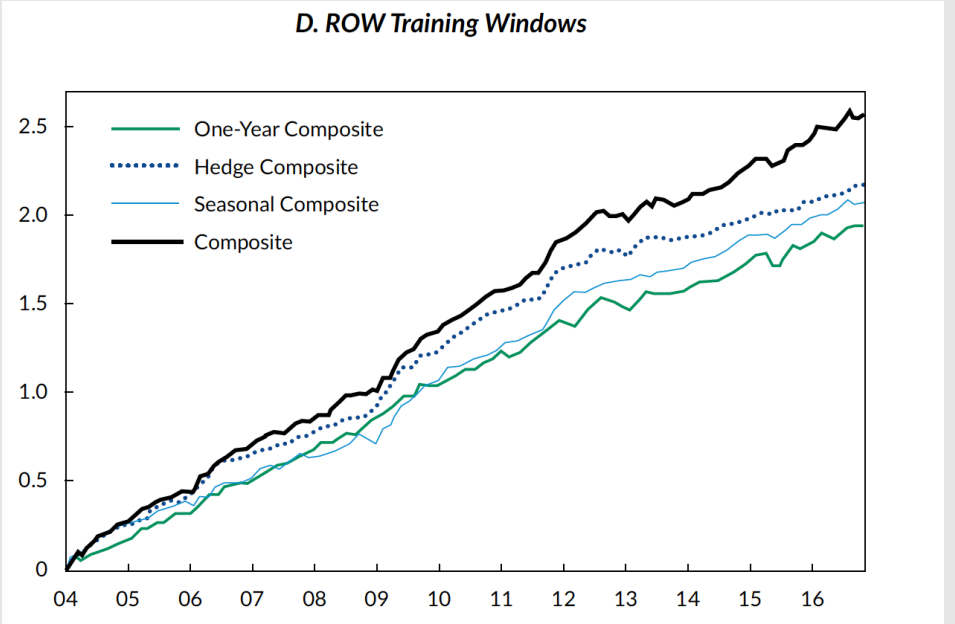
Overview of Our Approach




A. US Algorithms
AdaBoost Composite
Tree Boosting Composite
Neural Network Composite
SVM Composite
Composite
3.0
2.5
2.0
1.5
1.0
0.5
0
04
05
06
07
08
09
10
11
12
13
14
15
16


B. ROW Algorithms
AdaBoost Composite
Tree Boosting Composite
Neural Network Composite
SVM Composite
Composite
2.5
2.0
1.5
1.0
0.5
0
04
05
06
07
08
09
10
11
12
13
14
15
16


C. US Training Windows
One-Year Composite
Hedge Composite
Seasonal Composite
Composite
3.0
2.5
2.0
1.5
1.0
0.5
0
04
05
06
07
08
09
10
11
12
13
14
15
16


D. ROW Training Windows
One-Year Composite
Hedge Composite
Seasonal Composite
Composite
2.5
2.0
1.5
1.0
0.5
0
04
05
06
07
08
09
10
11
12
13
14
15
16

Results I

Table 2. Information Coefficients for Models, Composite, and Benchmarks

	US Zone		ROW	
Model	Rank IC	*t*-Statistic	Rank IC	*t*-Statistic
ML composite	6.48%	15.87	6.43%	16.37
12-Month period				
AdaBoost	3.19	9.07	3.54	10.98
SVM	4.61	8.00	5.08	10.35
Tree boosting	4.66	9.59	4.79	11.12
Neural network	3.99	12.94	3.17	11.97
Seasonal				
AdaBoost	3.17	11.53	3.25	11.75
SVM	5.01	9.65	4.79	10.56
Tree boosting	5.00	12.04	4.53	11.89
Neural network	5.20	15.39	4.41	11.92
Hedge				
AdaBoost	3.92	14.06	3.53	13.52
SVM	4.94	13.21	5.23	14.36
Tree boosting	5.00	14.14	4.78	16.16
Neural network	4.58	14.24	4.47	14.61
Top 10 factors benchmark	2.81	6.33	4.49	9.44
OLS benchmark	3.36	5.78	2.71	4.72

Correlations

Table 3. Correlations between Machine Learning Portfolios and the Market Factor

Portfolio	US Equal Weighted	US Risk Weighted	ROW Equal Weighted	ROW Risk Weighted	MKT
US equal weighted	1.00	0.81	0.32	0.29	−0.57
US risk weighted		1.00	0.28	0.30	−0.20
ROW equal weighted			1.00	0.96	−0.16
ROW risk weighted				1.00	−0.05
MKT					1.00

Note: MKT refers to the Fama–French (1992) market risk factor.

Table 4. Regression Results vs. Fama–French–Carhart Factors (*t*-statistics in parentheses)

A. Regression of MLA decile spreads on Fama–French–Carhart factors

	US Zone				ROW			
	Equal Weighted	Risk Weighted	Top 10 Benchmark	OLS Benchmark	Equal Weighted	Risk Weighted	Top 10 Benchmark	OLS Benchmark
Excess return	1.60	1.95	1.23	1.10	1.50	1.64	1.06	0.79
	(6.04)	(11.32)	(6.32)	(3.56)	(9.32)	(11.61)	(5.91)	(5.23)
Alpha	1.84	2.01	1.17	1.03	1.53	1.65	1.11	0.85
	(9.48)	(11.93)	(7.05)	(3.79)	(9.48)	(11.53)	(6.19)	(6.18)
MKT	−0.38	−0.09	0.07	0.02	−0.04	0.01	−0.07	−0.12
	(−6.93)	(−1.91)	(1.53)	(0.28)	(−0.80)	(0.14)	(−1.44)	(−3.03)
SMB	−0.23	−0.07	0.03	−0.08	−0.07	−0.05	−0.1	0.02
	(−2.44)	(−0.84)	(0.38)	(−0.63)	(−0.93)	(−0.67)	(−1.22)	(0.27)
HML	−0.06	0.00	0.37	0.05	−0.09	−0.11	0.1	0.01
	(−0.67)	(0.05)	(4.97)	(0.34)	(−1.25)	(−1.70)	(1.18)	(0.19)
MOM	0.18	0.08	−0.11	0.22	−0.06	−0.04	0.05	0.14
	(3.76)	(1.90)	(2.71)	(3.43)	(−1.44)	(−1.15)	(1.16)	(4.10)
Adjusted R^2	0.47	0.07	0.29	0.32	0.01	0.00	0.03	0.20
No. of observations	154	154	154	154	154	154	154	154

(continued)

Table 4. Regression Results vs. Fama–French–Carhart Factors (*t*-statistics in parentheses) (continued)

B. Regression of MLA spreads with transaction costs on Fama–French–Carhart factors

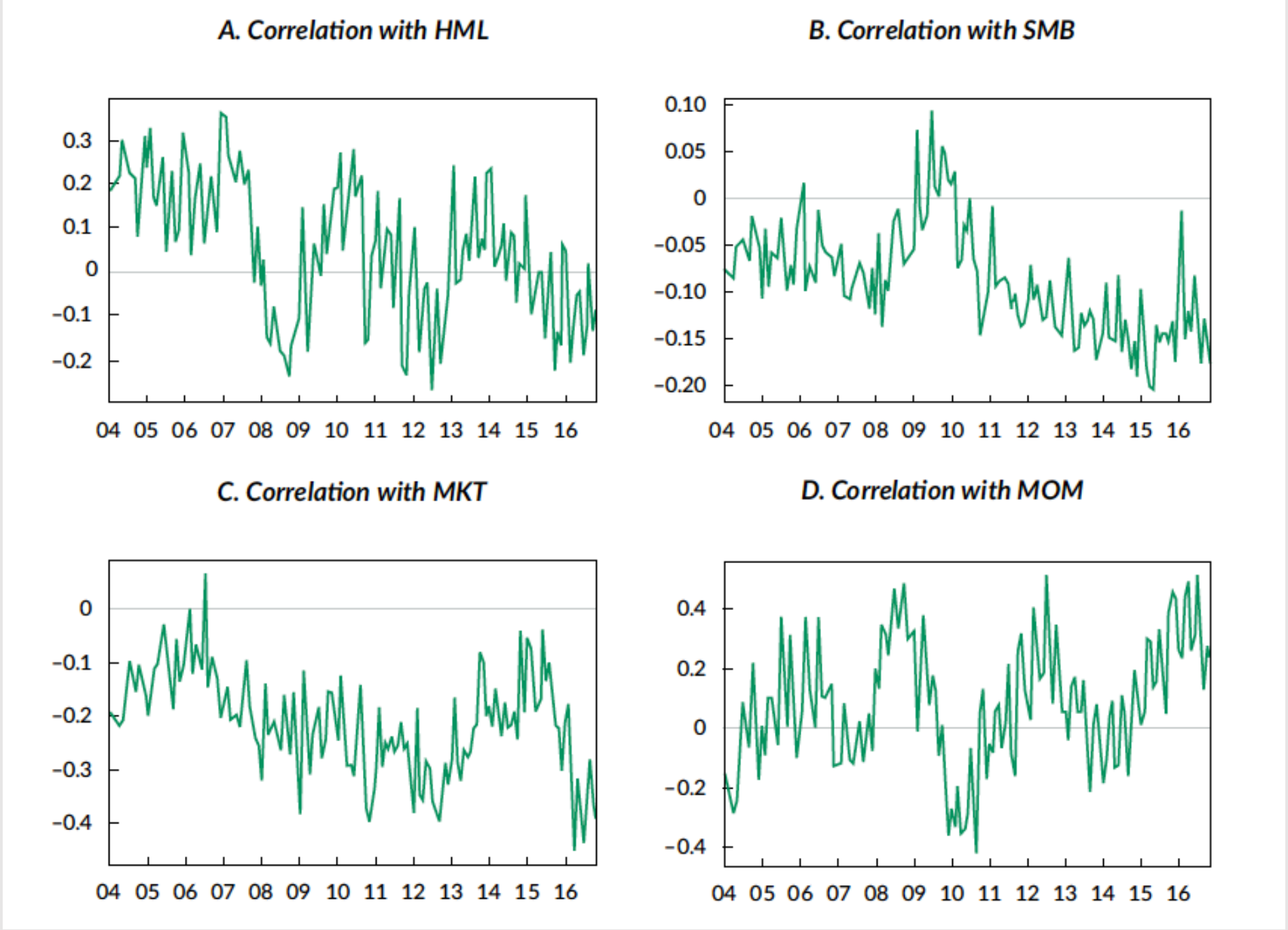
	US Zone				ROW			
	Equal Weighted	Risk Weighted	Top 10 Benchmark	OLS Benchmark	Equal Weighted	Risk Weighted	Top 10 Benchmark	OLS Benchmark
Excess return	1.35	1.67	1.00	0.85	1.24	1.38	0.93	0.61
	(5.15)	(9.76)	(5.16)	(2.77)	(7.74)	(9.83)	(5.17)	(4.04)
Alpha	1.61	1.74	0.94	0.77	1.27	1.39	0.98	0.67
	(8.34)	(10.37)	(5.69)	(2.90)	(7.90)	(9.74)	(5.45)	(4.89)
MKT	−0.38	−0.09	0.07	0.02	−0.04	0.01	−0.07	−0.12
	(−6.95)	(−1.99)	(1.59)	(0.21)	(−0.86)	(0.16)	(−1.41)	(−3.06)
SMB	−0.23	−0.06	0.03	−0.08	−0.06	−0.04	−0.1	0.02
	(−2.51)	(−0.75)	(0.36)	(−0.61)	(−0.85)	(−0.57)	(−1.23)	(0.29)
HML	−0.06	0.00	0.37	0.05	−0.09	−0.11	0.1	0.01
	(−0.66)	(0.02)	(4.99)	(0.38)	(−1.26)	(−1.70)	(1.2)	(0.21)
MOM	0.18	0.08	−0.11	0.22	−0.05	−0.04	0.05	0.14
	(3.76)	(1.97)	(−2.70)	(3.45)	(−1.37)	(−1.07)	(1.19)	(4.15)
Adjusted R^2	0.48	0.07	0.29	0.32	0.01	0.00	0.03	0.20
No. of observations	154	154	154	154	154	154	154	154

Table 5. Fama–MacBeth Regressions (*t*-statistics in parentheses)

	US Zone		ROW	
Forecast/Characteristic	Equal Weighted	Risk Weighted	Equal Weighted	Risk Weighted
ML composite	1.71	1.77	1.00	1.12
	(10.04)	(12.05)	(9.22)	(10.44)
Earnings revision	0.21	0.15	0.33	0.32
	(2.17)	(1.57)	(4.46)	(4.25)
Dividend yield	0.11	0.09	0.16	0.19
	(0.94)	(0.91)	(1.38)	(1.71)
Return on equity	0.28	0.10	−0.00	−0.04
	(1.81)	(0.80)	(−0.03)	(−0.48)
Book-to-price ratio	0.29	0.19	0.44	0.37
	(1.47)	(1.32)	(3.17)	(2.78)
Momentum	−0.44	−0.32	0.27	0.33
	(−1.21)	(−1.29)	(1.08)	(1.45)
Growth in earnings per share	0.12	0.05	−0.02	−0.04
	(1.08)	(0.64)	(−0.35)	(−0.65)
1-Month reversal	0.21	0.12	0.16	0.12
	(0.87)	(0.65)	(0.83)	(0.67)
Low volatility	−0.10	0.16	−0.13	0.05
	(−0.30)	(0.80)	(−0.58)	(0.28)
Earnings yield	−0.38	−0.17	0.09	0.25
	(−1.94)	(−1.34)	(0.86)	(2.52)
Accounting accruals	0.21	0.15	0.29	0.28
	(2.24)	(1.93)	(3.77)	(3.39)
Intercept	0.06	0.01	−0.31	−0.35
	(0.09)	(0.02)	(−0.66)	(−0.80)

Table 6. Long and Short Portfolios (*t*-statistics in parentheses)

	US Zone		ROW	
Measure	Long	Short	Long	Short
Excess return	1.90	−0.03	1.50	−0.13
	(3.54)	(−0.05)	(4.18)	(−0.33)
Fama–French–Carhart four-factor alpha	1.13	−0.96	0.95	−0.69
	(4.43)	(−3.28)	(4.72)	(−2.97)


A. Correlation with HML
0.3
0.2
0.1
0
−0.1
−0.2
04 05 06 07 08 09 10 11 12 13 14 15 16
B. Correlation with SMB
0.10
0.05
0
−0.05
−0.10
−0.15
−0.20
04 05 06 07 08 09 10 11 12 13 14 15 16
C. Correlation with MKT
0
−0.1
−0.2
−0.3
−0.4
04 05 06 07 08 09 10 11 12 13 14 15 16
D. Correlation with MOM
0.4
0.2
0
−0.2
−0.4
04 05 06 07 08 09 10 11 12 13 14 15 16

Pitfalls

- The algos only see the data they are presented with: no solution for overfit inputs
- If there is look-ahead bias any algo worth its salt will suggest to trade on that!
- How is internal and external data checked for errors?
- Does your data team apply the same process in production and on historical data?
- Are costs accounted for? Cross-sectional variation in hidden costs? Asymetric between long and short?

Conclusion

- Feature Engineering can Increase the signal-to-noise ratio before any estimation takes place
- Forecast combinations can lead to more signal and less noise because a portion of errors cancels out
- ML only effective if the past is informative about the future: Data needs to be point-in-time
- Inputs ideally should have low and comparable selection bias

Q&A


SUPPORT VECTOR MACHINES
REPEA
POWE
LAWS
END
DUALITY
GAP
Map Reduce
Map Reuse
Map Recycle
BAYESIANS AGAINST DISCRIMINATION
FREE VARIABLES!
BAN GENETIC ALGORITHMS