Ensemble Approaches for Financial Prediction Using Stacking and Rolling Window Method

Jaeyoon Han

Master's Course
Dept. of Social Network Science
Kyung Hee University

CONTENTS

1. Introduction

Introduction

주식시장 예측 연구는 지속적으로 이루어져 왔다.

학술적 관점

시계열에 대한 분석으로
주어진 시스템이 어떤 법칙에 의해
움직이는가에 대한 질문을
이해할 수 있게 된다.

실무적 관점

주식시장을 예측할 수 있다면
높은 수익률을 보장할 수 있기 때문에
매우 중요하다.

Introduction

효율적 시장 가설
(Efficient Market Hypothesis, EMH)

주식시장은 **무작위 행보** 특성을 가지고 있으며,
과거 데이터를 이용해 **초과수익을 창출하기 매우 어려움**
(Malkiel & Fama, 1970)

다양한 요소에 영향을 받는다

정치적 사건, 기업의 정책, 경제적 상황, 투자자의 예측, 원자재의 가격,
은행 금리, 환율, 주식 시장의 움직임, 투자자들의 심리 상태 등
다양한 거시경제적 요소에 영향을 받는다.
(Miao, Chen & Zhao, 2007; Tan, Quake & Ng, 2007)

Introduction

기계학습의 도입

기계학습은 현재 데이터를 기반으로
컴퓨터가 스스로 학습을 통해
일반적인 패턴을 찾아낼 수 있다는 장점이 있음

인공신경망 (Artificial Neural Network, ANN)
서포트 벡터 머신 (Support Vector Machine, SVM)
로지스틱 회귀분석 (Logistic Regression)
k-최근접 이웃 (k-Nearest Neighbor)
랜덤 포레스트 (Random Forest)

→ **앙상블 모델에 관한 연구가 매우 적음**

2. Related Works

Literature Review

Authors	Targets	Algorithms
Balling *et al.*, 2001	유럽의 주가 등락 예측	Logistic Regression, ANN, k-NN, SVM, RF, AdaBoost with technical indicators
Kumar & Thenmozhi, 2006	인도 주식시장의 등락 예측	LDA, Logistic Regression, ANN, RF, SVM with technical indicators
Ou & Wang, 2009	항셍 지수 등락 예측	10 data mining techniques with Log-return
Guresen, Kayakutlu & Daim, 2011	NASDAQ 지수 예측	MLP, RNN, DAN2, GARCH to extract new inputs
Kara, Boyacioglu & Baykan, 2011	이스탄불 주식 시장 예측	ANN, SVM with technical indicators
Roy *et al.*, 2015	골드만 삭스의 주가지수 예측	LASSO, Ridge, Bayesian regularized ANN with daily information of index
Patel *et al.*, 2015	인도 주식 시장의 등락과 지수 예측	ANN, SVM, RF, Fusion approach **(Ensemble)** with technical indicators
Yang, B., Gong Z. & Yang W., (2017)	중국 상하이 종합지수, SZSE 종합지수 예측	DNN with bagging approach **(Ensemble)** using open, high, low, close price

Literature Review

대부분의 연구에서 입력 데이터로 **Technical indicator 사용**

Fusion method 외의 ensemble approach는 찾기 힘듦

XGBoost와 같이 **최근 좋은 성능을 보이는 알고리즘**에 대한 연구 부족
(Kaggle)

↓

기존 연구에서 사용하지 않은 다양한 Regressor를 사용하여 성능 향상

최근 폭넓게 사용되고 있는 Stacking 을 이용하여 모델 앙상블

Algorithms

최소제곱법 (Ordinary Least Squares)

베이스라인 모델로 자주 사용

매우 단순한 회귀분석 알고리즘으로
잔차제곱합을 최소화하는 회귀계수를 찾는다.

$$\hat{\beta} = \arg\min_{\beta}\{(y - X\beta)^T(y - X\beta)\}$$

$$= (X^TX)^{-1}X^Ty$$

Algorithms

능형 회귀 (Ridge Regression)

분산이 큰 OLS의 성능 저하를 해결하기 위해
기존의 비용 함수에 L_2 페널티를 추가하여 성능 향상

$$Obj(\theta) = L(\theta) + \lambda \|\hat{\beta}\|^2$$

Lasso

기존의 비용 함수에 L_1 페널티를 추가하여 성능 향상
L1 페널티의 특성으로 인해 변수 선택(feature selection)이 가능

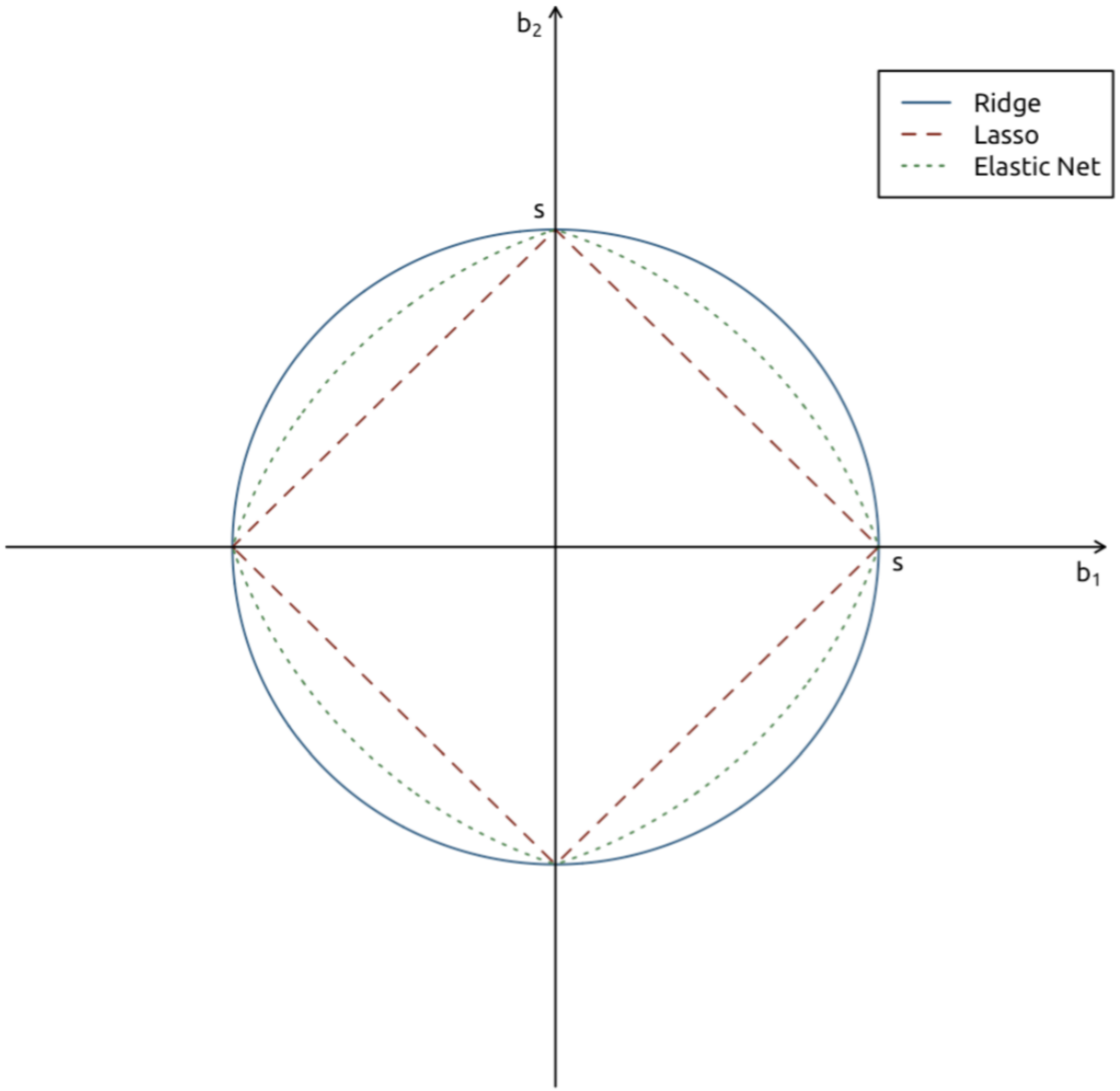
$$Obj(\theta) = L(\theta) + \lambda \|\hat{\beta}\|_1$$

Elastic Net Regression

두 페널티를 선형결합하여 사용

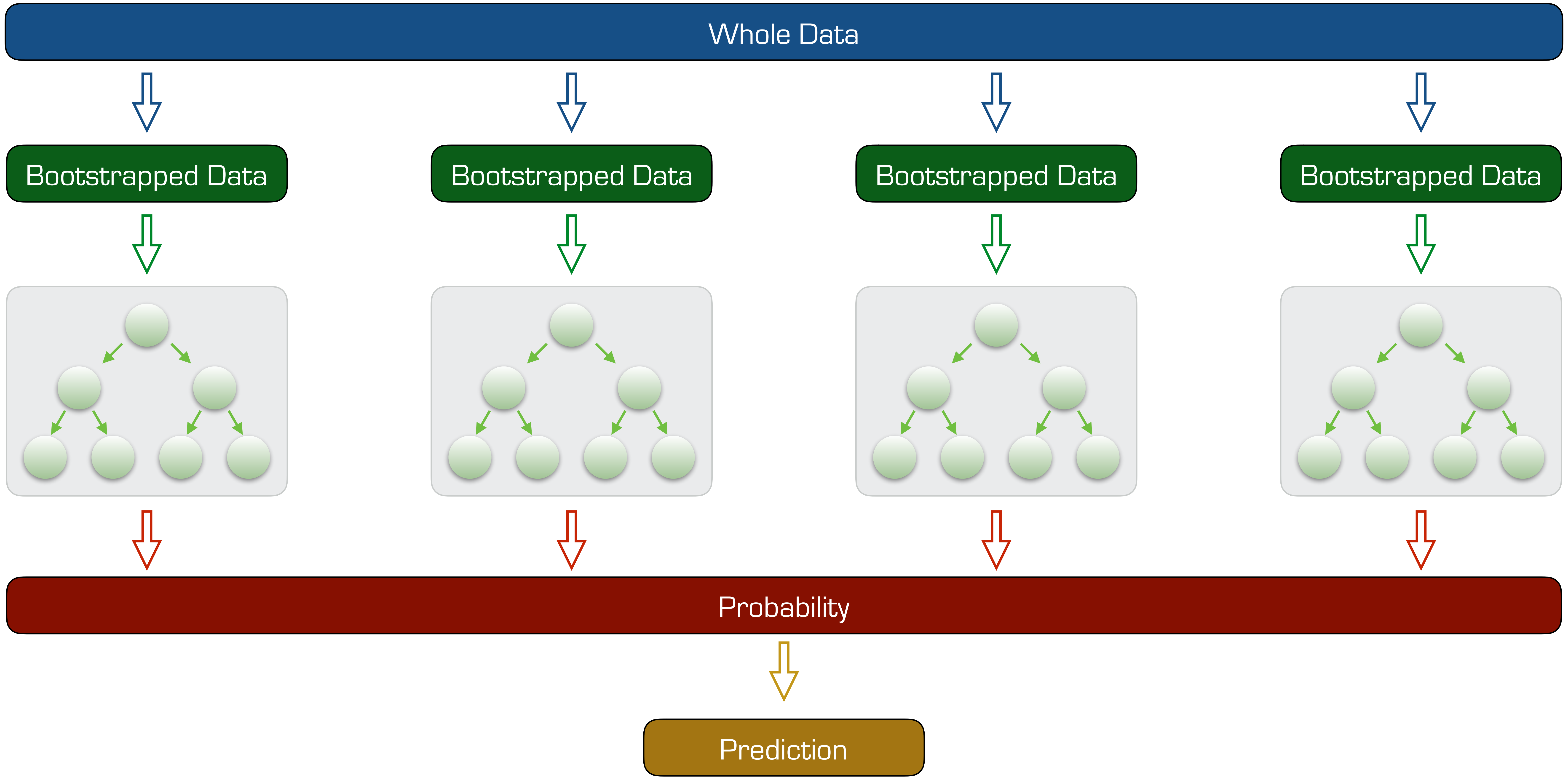
$$Obj(\Theta) = L(\Theta) + \lambda(\alpha|\hat{\beta}| + (1-\alpha)\|\hat{\beta}\|^2)$$

Algorithms



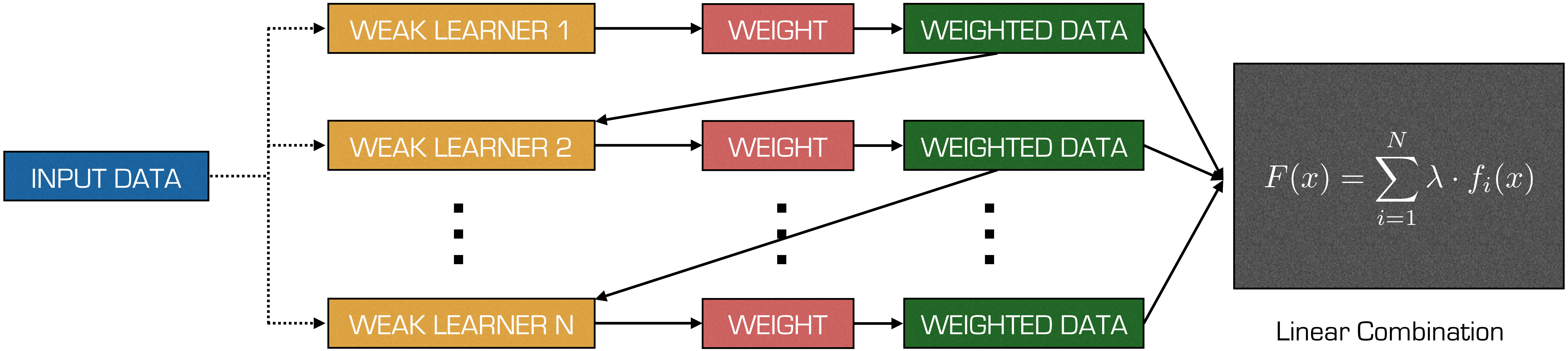
Algorithms

랜덤 포레스트 (Random Forest)



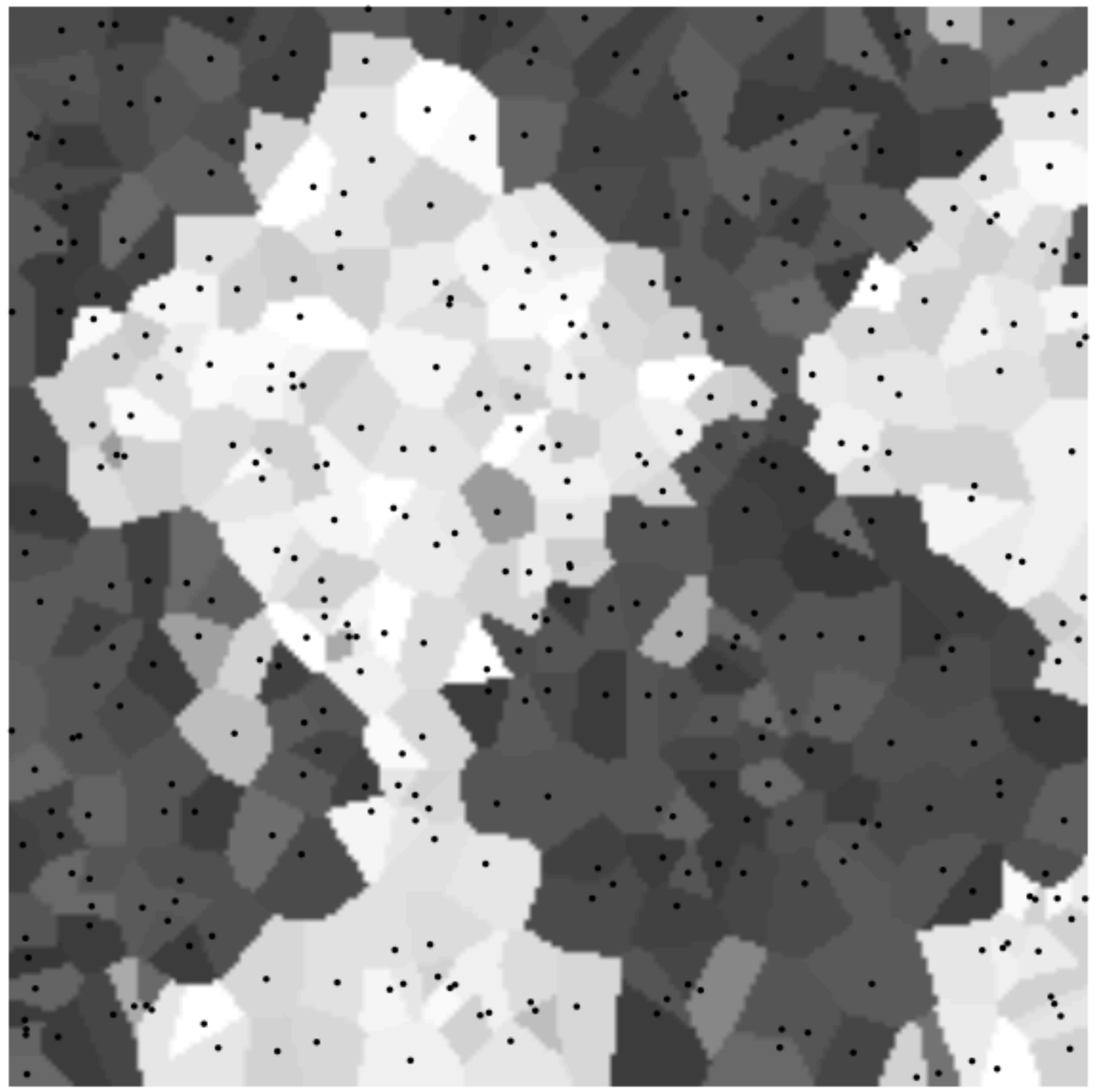
Algorithms

XGBoost

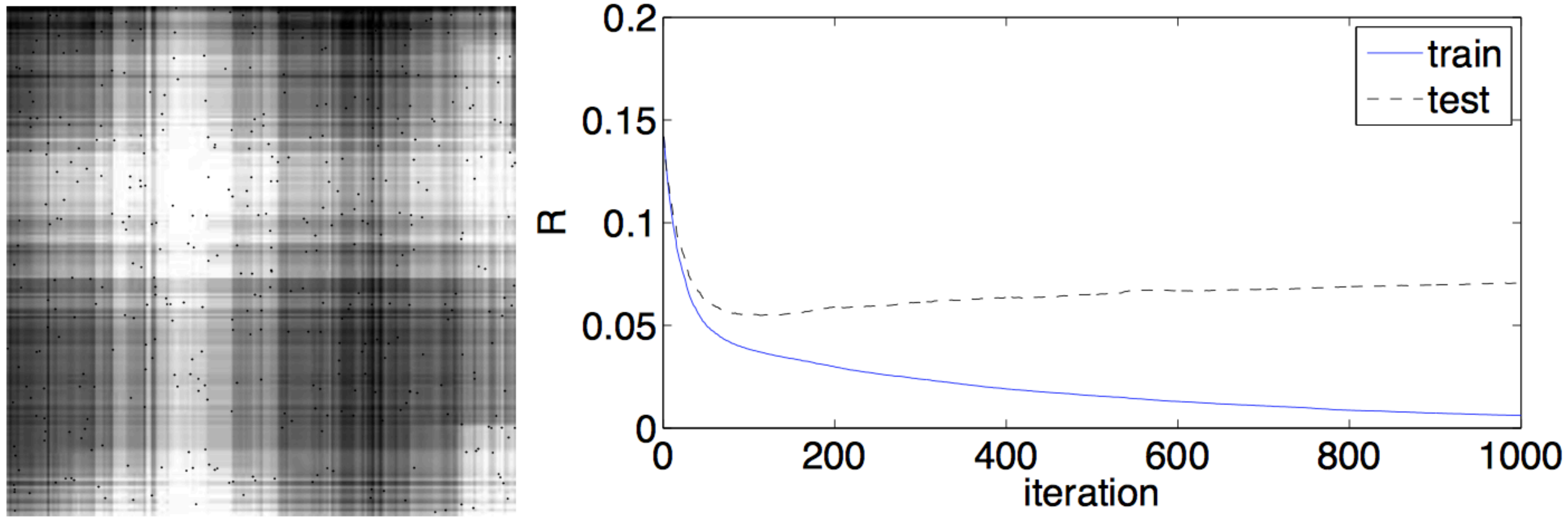


트레이닝 데이터에 대하여 모델링을 하여,
잘못 분류된 데이터에 가중치를 주어 가산적으로 모델을 학습하는 방식

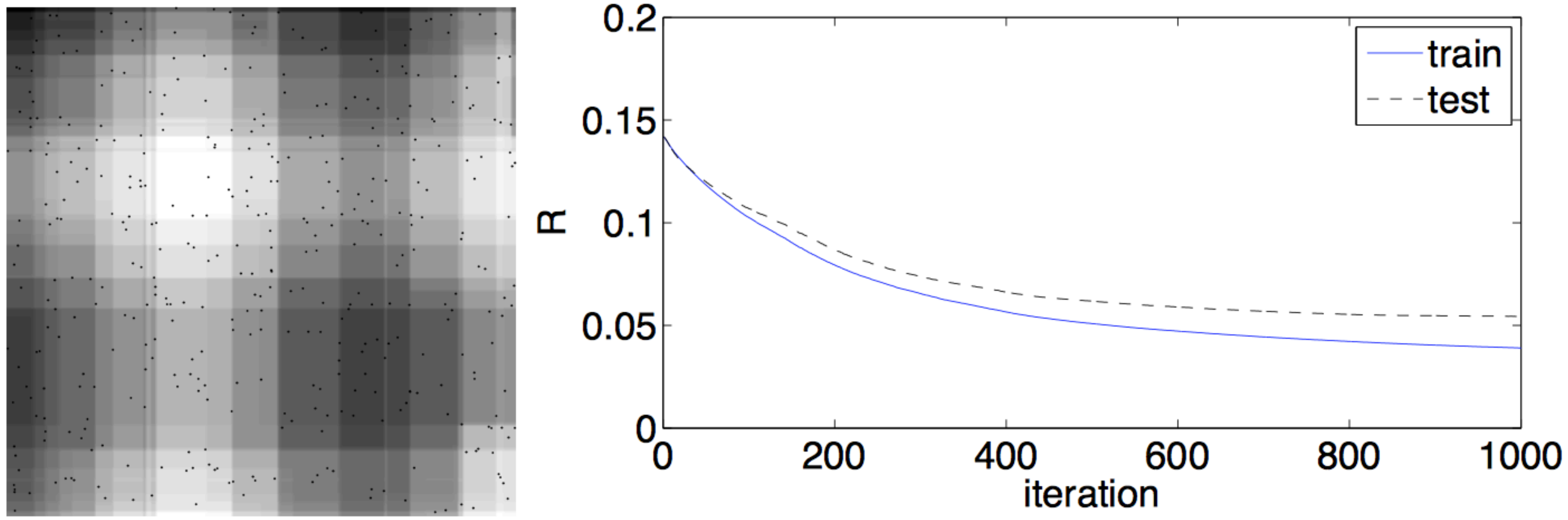
Algorithms



Gradient boosting, least-squares loss, $v = \frac{1}{20}$



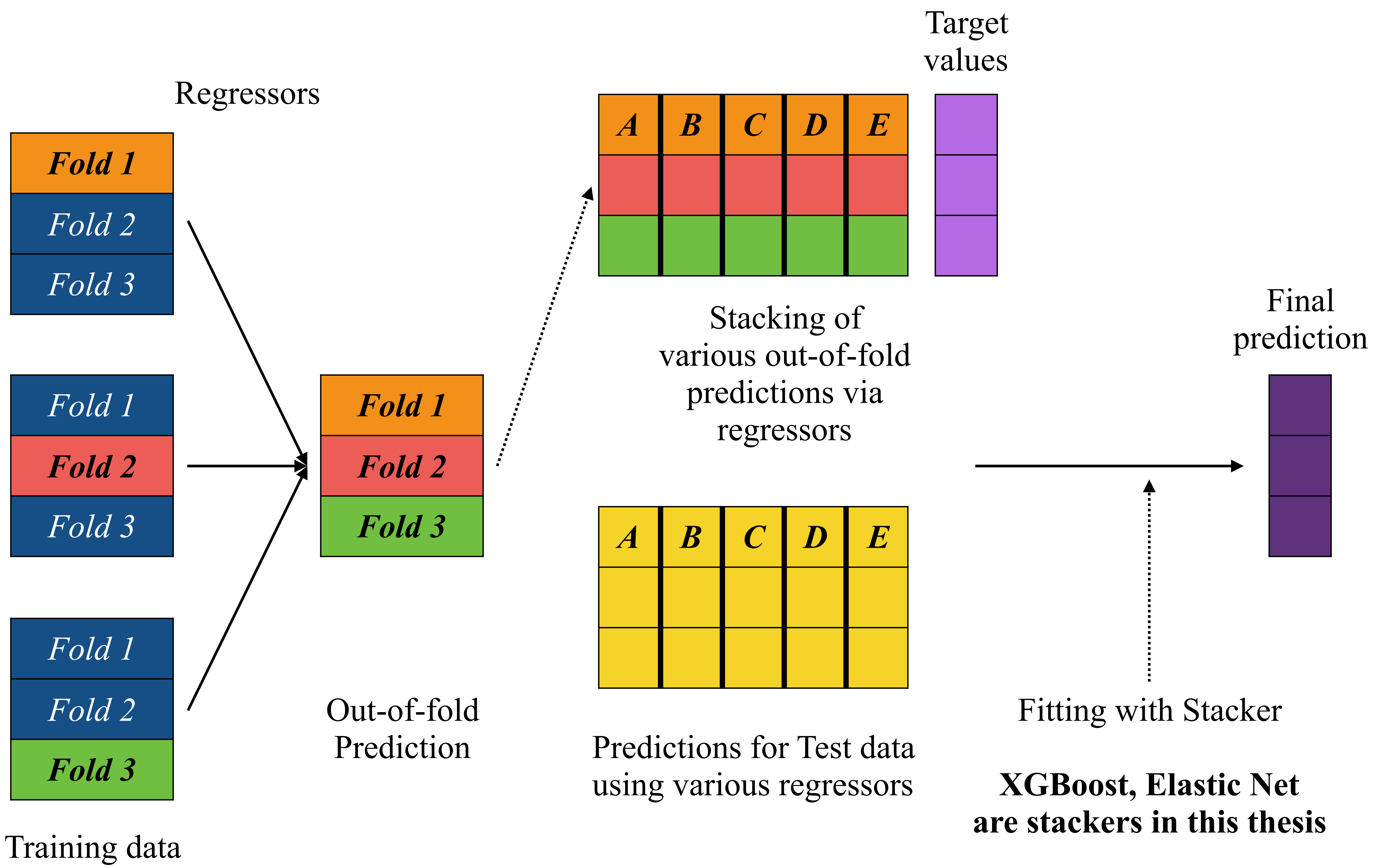
Gradient boosting, least-squares loss, $v = \frac{1}{200}$



Stacking

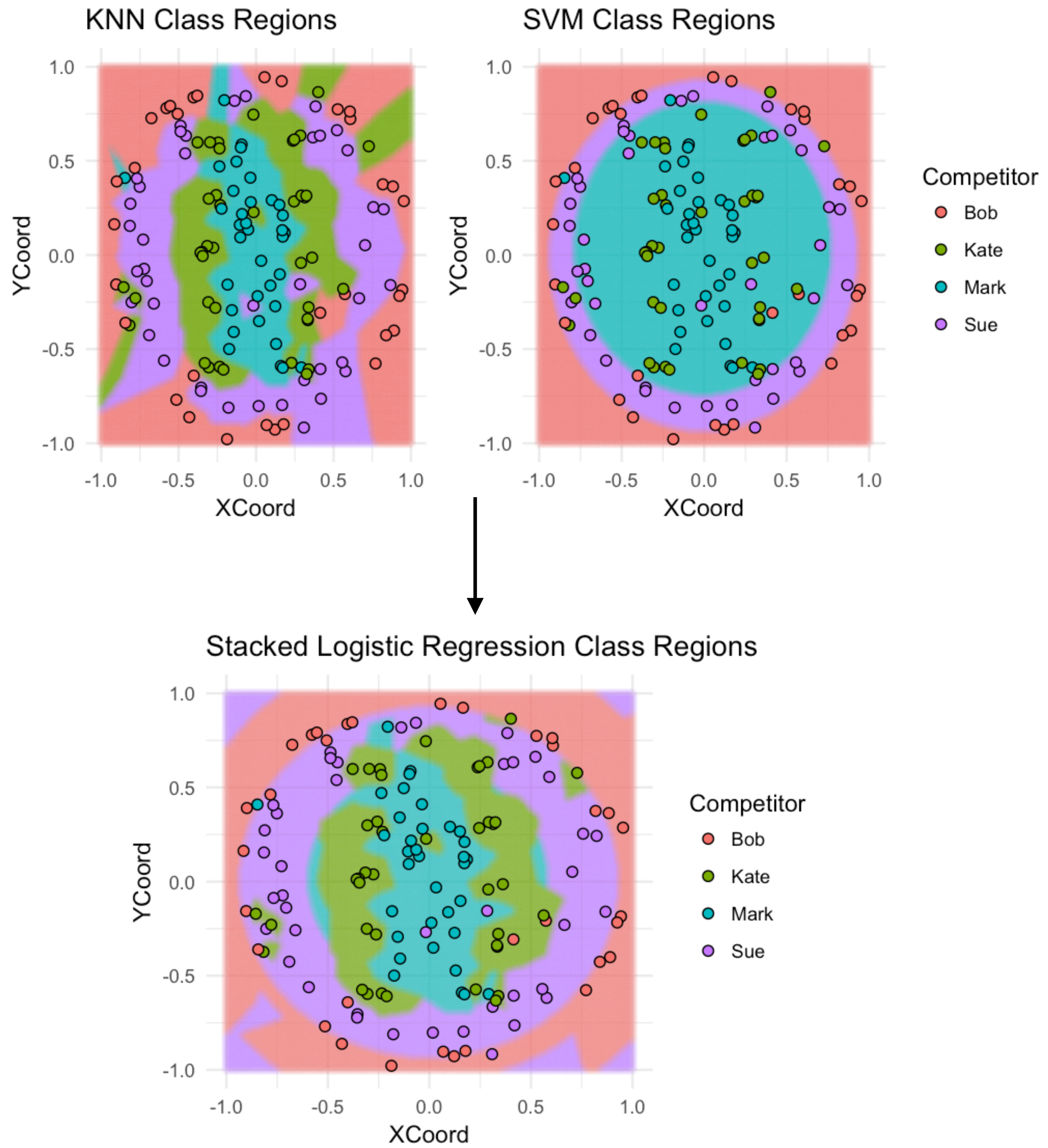
여러 알고리즘의 고유 특성을 그대로 반영하여 앙상블하는 기법

메타 학습 (meta learning)이라고도 부름



일반화 오류를 줄일 수 있음

여러 알고리즘의 특성을 **모두 반영**할 수 있음



http://blog.kaggle.com/2016/12/27/a-kagglers-guide-to-model-stacking-in-practice/

3. Experimental Design

Data Description

Type	Variable name	Description
Stock Market Indices	DOWJONES	Dow Jones Industrial Average
	SNP	S&P 500
	DAX	DAX (Deutscher Aktien IndeX)
	KOSPI	Korea Stock Exchange KOSPI Index
Exchange Rates	USDKRW	US Dollar / KOR Won Exchange Rate
	EURUSD	Euro / US Dollar Exchange Rate
	GBPUSD	British Pound / US Dollar Exchange Rate
	AUDUSD	Australian Dollar / US Dollar Exchange Rate
	DXY	Dollar Index Spot
Commodity	WTI	WTI Oil
	XAUUSD	Gold Spot ($/Oz)
	XAGUSD	Silver Spot ($/Oz)
	COOPER	Cooper
Bond Futures	KE1	Korea Bond Futures 3Y
	KAA1	Korea Bond Futures 10Y
	FV1	US Bond Futures 5Y
	JB1	Japan Bond Futures 10Y
	RX1	Germany Bond Futures 10Y

Training : 2010/01/01 - 2015/12/31
Test : 2016/01/01 - 2016/11/30

Data Description

Index	Lagged 1-Day				Lagged 5-Day			
	Training (10/01/01 ~ 15/12/31)		Test (16/01/01 ~ 16/11/30)		Training (10/01/01 ~ 15/12/31)		Test (16/01/01 ~ 16/11/30)	
	Increase	Decrease	Increase	Decrease	Increase	Decrease	Increase	Decrease
DOWJONES	50.6%	49.4%	55.0%	45.0%	59.3%	40.7%	54.1%	45.9%
SNP	51.7%	48.3%	55.4%	44.6%	60.5%	39.5%	56.7%	43.3%
DAX	52.8%	47.2%	50.2%	49.8%	57.1%	42.9%	48.1%	51.9%
KOSPI	51.2%	48.8%	54.1%	45.9%	54.4%	45.6%	53.2%	46.8%
USDKRW	45.8%	54.2%	47.6%	52.4%	46.8%	53.2%	51.9%	48.1%
EURUSD	48.5%	51.5%	47.6%	52.4%	46.6%	53.4%	43.7%	56.3%
GBPUSD	48.1%	51.9%	45.9%	54.1%	50.0%	50.0%	50.6%	49.4%
AUDUSD	49.0%	51.0%	52.8%	47.2%	49.1%	50.9%	52.4%	47.6%
DXY	48.6%	51.4%	51.5%	48.5%	52.3%	47.7%	58.9%	41.1%
WTI	47.3%	52.7%	48.1%	51.9%	49.4%	50.6%	54.1%	45.9%
XAUUSD	50.6%	49.4%	51.1%	48.9%	51.8%	48.2%	55.0%	45.0%
XAGUSD	50.8%	49.2%	54.5%	45.5%	48.1%	51.9%	52.4%	47.6%
COOPER	47.0%	53.0%	49.4%	50.6%	47.5%	52.5%	55.0%	45.0%
KE1	46.3%	53.7%	47.1%	52.9%	54.9%	45.1%	48.5%	51.5%
KAA1	45.3%	54.7%	49.4%	50.6%	56.0%	44.0%	50.6%	49.4%
FV1	48.6%	51.4%	48.1%	51.9%	53.0%	47.0%	51.1%	48.9%
JB1	49.6%	50.4%	50.6%	49.4%	59.1%	40.9%	55.8%	44.2%
RX1	49.1%	50.9%	53.7%	46.3%	57.3%	42.7%	58.0%	42.0%

Research Design

휴장일로 인해 결측값이 존재하는 경우,
해당 결측값을 이전 지수 값으로 대체

(Kumar & Deo, 2012)

입력 변수의 형태를 1일 전, 5일 전의 로그 수익률로 변경

$$R(t) = \log(S(t)) - \log(S(t - \Delta t))$$

채권 선물의 경우, 음수 값을 갖는 경우가 있기 때문에
100에서 각 데이터 값을 빼 단위 수익률로 데이터 전처리

Research Design

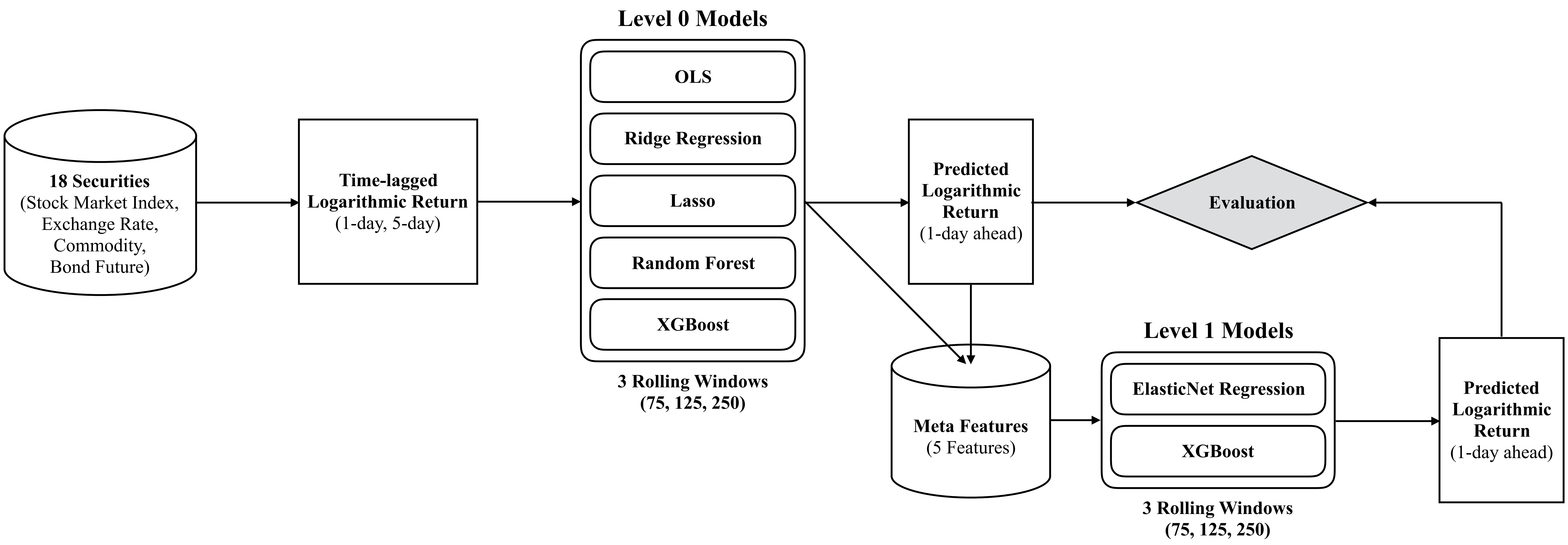
Timestamp	Var A	Var B	Var C	Var D	Var E	…	Var Q	Var R	Target
1	A_1	B_1	C_1	D_1	E_1	…	Q_1	R_1	T_1
2	A_2	B_2	C_2	D_2	E_2	…	Q_2	R_2	T_2
3	A_3	B_3	C_3	D_3	E_3	…	Q_3	R_3	T_3
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮
75	A_{75}	B_{75}	C_{75}	D_{75}	E_{75}	…	Q_{75}	R_{75}	T_{75}
76	A_{76}	B_{76}	C_{76}	D_{76}	E_{76}	…	Q_{76}	R_{76}	T_{76}
77	A_{77}	B_{77}	C_{77}	D_{77}	E_{77}	…	Q_{77}	R_{77}	T_{77}
78	A_{78}	B_{78}	C_{78}	D_{78}	E_{78}	…	Q_{78}	R_{78}	T_{78}
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮
d	A_d	B_d	C_d	D_d	E_d	…	Q_d	R_d	T_d

1-day ahead prediction

장기간의 추세가 오히려 모델의 분산을 과도하게 높이면서
발생하는 낮은 예측력 문제를 완화할 수 있음

75일 (4개월), 125일 (6개월), 250일 (1년)

Research Design



Research Design

Parameter space

Model	Parameters	Level(s)	Note
Ridge	Strength of Regularization (λ)	$2.0 \times 10^{-9} \sim 4.5 \times 10^{-5}$	200 candidates in the range
LASSO	Strength of Regularization (λ)	$2.0 \times 10^{-9} \sim 4.5 \times 10^{-5}$	200 candidates in the range
ELN	Strength of Regularization (λ)	$2.0 \times 10^{-9} \sim 1.0$	300 candidates in the range
	The ratio of L1 penalty (α)	0.1, 0.2, …, 0.9	-
RF	The number of estimators (n)	100, 200, 300	Fixed to 200
	The maximum depth of the tree (d)	2, 3, 4, 5	Fixed to 3
	The number of features when splitting (f)	4, 17	Fixed to 4
XGB	The maximum depth of the tree (d)	2, 3, 4, 5	Fixed to 4
	Learning Rate (η)	0.1, 0.2, 0.3, 0.4	Fixed to 0.3
	The number of estimators (n)	300	Set by early stopping
XGB	The maximum depth of the tree (d)	2, 3, 4	Fixed to 3
(Stacked)	Learning Rate (η)	0.01, 0.05, 0.1	Fixed to 0.01
	The number of estimators (n)	1500	Set by early stopping

4. Results

Result

Average of the prediction performance of each algorithm in terms of time-lagged data

Models	Lagged	Measures			
		Accuracy	R^2	rRMSE	MAPE
OLS	1-Day	0.486	0.000739	0.00859	0.6116
	5-Day	0.824	0.674622	0.01079	0.8373
Ridge	1-Day	0.488	0.001003	0.00846	0.6020
	5-Day	0.823	0.675341	0.01069	0.8300
LASSO	1-Day	0.488	0.001596	0.00829	0.5837
	5-Day	0.814	0.643632	0.01079	0.8398
RF	1-Day	0.486	0.021385	0.00823	0.5843
	5-Day	0.810	0.605804	0.01129	0.8749
XGB	1-Day	0.486	0.007559	0.00851	0.6100
	5-Day	0.807	0.602628	0.01147	0.8776
XGB (Stacked)	1-Day	0.497	0.005446	0.00905	0.6603
	5-Day	**0.828**	**0.692136**	0.01055	0.8087
ELN (Stacked)	1-Day	0.484	0.002283	**0.00821**	**0.5820**
	5-Day	0.826	0.677381	0.01039	0.8096

Stacking models outperforms other single models with respect to every metrics

It shows the difference of characteristic between XGBoost and Elastic Net Regression

Result

Average performance for each type of data and window size

Lagged	Window Size	Measures			
		Accuracy	R^2	rRMSE	MAPE
1-Day	75	**0.491**	**0.007612**	0.0087	0.6219
	125	0.489	0.005973	0.0084	0.6042
	250	0.483	0.003562	**0.0083**	**0.5885**
5-Day	75	0.812	0.639724	0.0110	0.8576
	125	0.821	0.655049	0.0108	0.8377
	250	**0.824**	**0.664460**	**0.0107**	**0.8329**

To predict the value of index in financial market, 1-day logarithmic return and window size of 250-day should be used

To predict the trend of financial market, 5-day logarithmic return and window size of 250-day should be used

The performance with 250 window size is improved significantly than other options,
regardless of the input type of logarithmic return.

Result

Best performance for each metric and each security

Index	Measures			
	Accuracy	R^2	rRMSE	MAPE
DOWJONES	0.939	0.950945	0.0039	0.2999
SNP	0.935	0.955393	0.0038	0.2934
DAX	0.818	0.629791	0.0147	0.9809
KOSPI	0.827	0.661463	0.0079	0.5700
USDKRW	0.831	0.699477	0.0064	0.4774
EURUSD	0.965	0.942838	0.0030	0.2184
GBPUSD	0.810	0.605786	0.0076	0.5293
AUDUSD	0.818	0.661467	0.0075	0.5701
DXY	0.970	0.962573	0.0022	0.1660
WTI	0.753	0.496469	0.0310	2.3794
XAUUSD	0.892	0.784251	0.0104	0.7477
XAGUSD	0.853	0.795648	0.0149	1.0678
COOPER	0.771	0.596170	0.0119	0.8865
KE1	0.866	0.877997	0.0002	0.0131
KAA1	0.913	0.898816	0.0002	0.0189
FV1	0.892	0.773337	0.0004	0.0299
JB1	0.706	0.390252	0.0002	0.0137
RX1	0.874	0.628528	0.0003	0.0254

The accuracy is greater than 0.706 (JB1) for all variable and R^2 is at least greater than 0.6 for the most variable.

WTI is tough to forecast its index value due to the big fluctuations
which overwhelm statistical confidence interval, occasionally happen (Hamilton, 2008).

Result

Type	Models	Measures			
		Accuracy	R^2	rRMSE	MAPE
Stock Market Index	OLS	0.677	0.386081	0.0101	0.7180
	Ridge	0.675	0.387311	0.0100	0.7120
	LASSO	0.671	0.385360	0.0098	0.6958
	RF	0.667	0.374378	0.0102	0.7200
	XGB	0.673	0.372077	0.0104	0.7304
	XGB (Stacked)	**0.680**	0.382656	0.0106	0.7581
	ELN (Stacked)	0.673	**0.392562**	**0.0097**	**0.6796**
Exchange Rate	OLS	0.660	0.363216	0.0074	0.5450
	Ridge	0.663	0.363953	0.0073	0.5392
	LASSO	0.657	0.351397	0.0075	0.5456
	RF	0.656	0.325982	0.0078	0.5749
	XGB	0.650	0.328254	0.0078	0.5648
	XGB (Stacked)	**0.670**	**0.369519**	0.0076	0.5455
	ELN (Stacked)	0.657	0.364436	**0.0071**	**0.5265**
Commodity	OLS	0.626	0.257432	0.0239	1.8281
	Ridge	0.630	0.258418	0.0235	1.8037
	LASSO	0.624	0.240915	0.0234	1.7922
	RF	0.615	0.235250	0.0235	1.8096
	XGB	0.609	0.230777	0.0243	1.8762
	XGB (Stacked)	**0.642**	**0.319813**	0.0234	1.8310
	ELN (Stacked)	0.623	0.263447	**0.0229**	**1.7621**
Bond Futures	OLS	0.656	**0.337622**	**0.0003**	0.0261
	Ridge	0.655	0.336883	**0.0003**	0.0258
	LASSO	0.651	0.308993	0.0004	0.0263
	RF	0.652	0.315256	0.0004	0.0281
	XGB	0.651	0.287799	0.0004	0.0278
	XGB (Stacked)	0.658	0.324154	0.0004	0.0275
	ELN (Stacked)	**0.664**	0.334151	**0.0003**	**0.0250**

← DOWJONES
SNP
DAX
KOSPI

← USDKRW
EURUSD
GBPUSD
AUDUSD
DXY

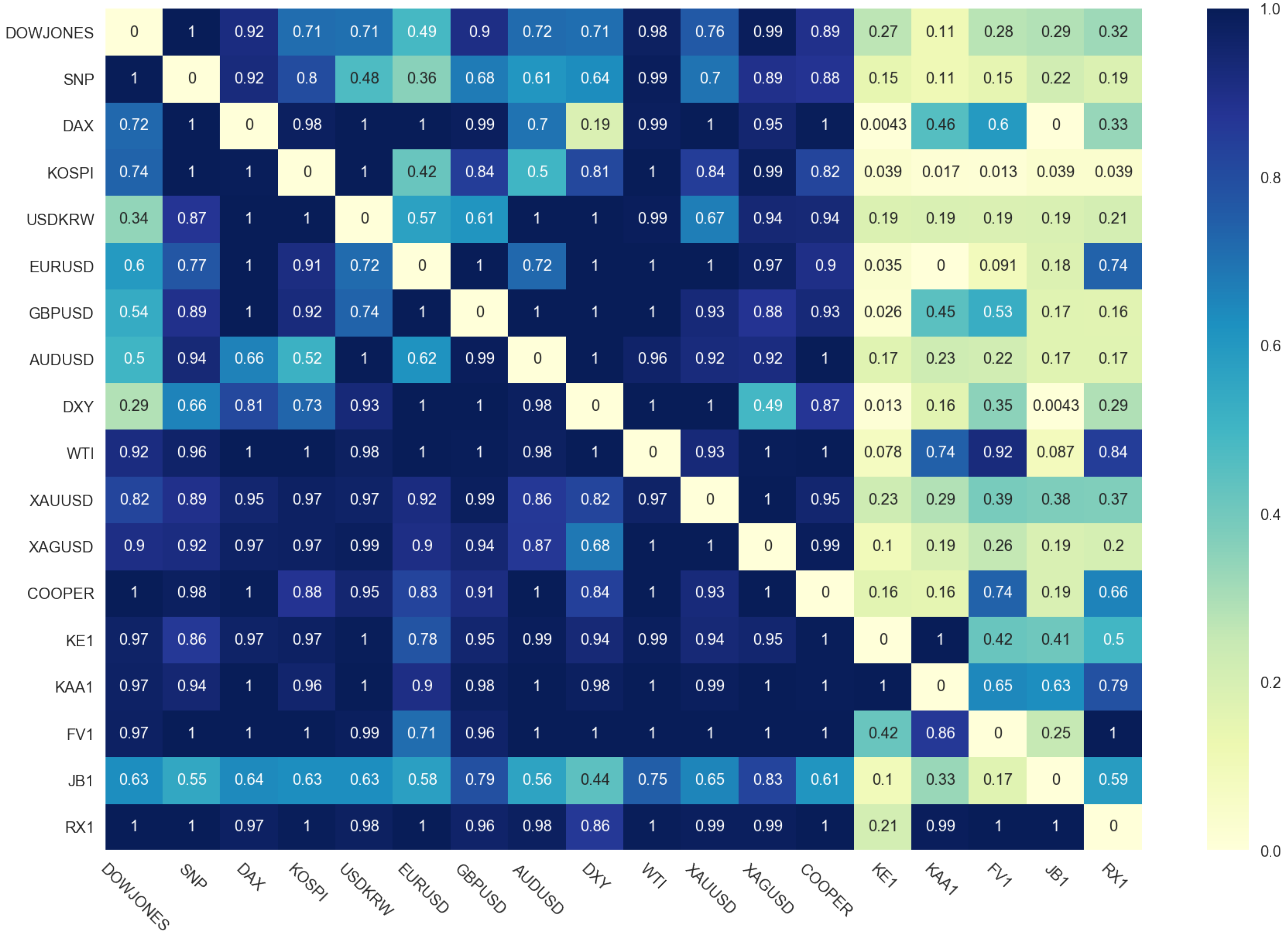
← WTI
XAUUSD
XAGUSD
COOPER

← KE1
KAA1
FV1
JB1
RX1

It is difficult to forecast the index for commodity due to its sensitivity such as substantial swings and sharp price reversals

(Bernanke, 2009; Pierdzioch, Rülke & Stadtmann, 2013; Baumeister & Peersman, 2013)

Feature Importance



1. The result exhibits that variables in bond futures have no significant effect on predicting.

2. Indices for the commodity is essential to forecast the trend of exchange rate.
   This result supports the previous studies that commodity prices and exchange rate are not in the relationship of random walk, and so predictable mutually.
   (Kohlscheen, Avalos & Schrimpf, 2017)

3. The impact of Dow Jones on predicting exchange rate, US dollar in particular, is inconsistent.
   (Bernard & Galati, 2000; Giannarakis *et al.*, 2017)

4. Conclusion

Conclusion

- **Stacked generalization outperforms other single models.**
  - The average prediction performance of XGBoost with stacking with the weekly return, 0.828 for accuracy and 0.692136 for R^2, was significantly better than that of the single models with the weekly return, 0.816 for accuracy and 0.640627 for R^2.
  - In terms of rRMSE and MAPE, which evaluate the error of prediction values, Elastic Net Regression with stacked generalization with the daily return outperforms other models.

- **I have identified the difference of performance between the case of using the daily return and the weekly return.**
  - With respect to forecast the trend of the market, the performance using the weekly return was better, while the daily return performed well on predicting the value of index precisely.

- **I examined the relationship and the influence between used variables by feature selection of LASSO and supports the results of the previous studies.**

Conclusion

The major contributions

- The performance of the proposed approach gives high performance, and it is based on a combination of simple techniques, and so any machine learning algorithm can be applied to the proposed approach.
- It was attempted to mutually predict financial indices based on various country and characteristics without external data.
- I have proposed the analysis guideline for predicting the stock market through extensive model implementation. (time-lagged data, the size of window, etc.)
- Using the automatic variable selection by LASSO, I looked at the relationship between variables and re-verified the previous studies which observed the correlation between financial indices.

References

- Malkiel, Burton G., and Eugene F. Fama, "Efficient capital markets: A review of theory and empirical work," The Journal of Finance, Vol. 25, No. 2, (1970), 383-417.
- Miao, Kai, Fang Chen, and Z. G. Zhao, "Stock price forecast based on bacterial colony RBF neural network [j]," Journal of Qingdao University (Natural Science Edition), Vol. 20, No. 2, (2007), 50-54.
- Tan, Tuan Zea, Chai Quek, and Goek See Ng, "Biological brain-inspired genetic complementary learning for stock market and bank failure prediction," Computational Intelligence, Vol. 23, No. 2, (2007), 236-261.
- Ballings, Michel, et al., "Evaluating multiple classifiers for stock price direction prediction," Expert Systems with Applications, Vol. 42, No. 20, (2015), 7046-7056.
- Kumar, Manish, and M. Thenmozhi, "Forecasting stock index movement: A comparison support vector machines and random forest," In Indian Institute of Capital Markets 9th Capital Markets Conference Paper, (2006). Available at SSRN: https://ssrn.com/abstract=876544
- Ou, Phichhang, and Hengshan Wang, "Prediction of stock market index movement by ten data mining techniques," Modern Applied Science, Vol. 3, No. 12, (2009).
- Guresen, Erkam, Gulgun Kayakutlu, and Tugrul U. Daim, "Using artificial neural network models in stock market index prediction," Expert Systems with Applications, Vol. 38, No. 8, (2011), 10389-10397.
- Kara, Yakup, Melek Acar Boyacioglu, and Ömer Kaan Baykan, "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange," Expert Systems with Applications, Vol. 38, No. 5, (2011), 5311-5319.
- Roy, Sanjiban Sekhar, et al. "Stock market forecasting using lasso linear regression model." Afro-European Conference for Industrial Advancement, Springer, Cham, (2015).
- Hamilton, James D., "Understanding crude oil prices", *National Bureau of Economic Research*, (2008).

References

- Patel, Jigar, et al. "Predicting stock market index using fusion of machine learning techniques." Expert Systems with Applications, Vol. 42, No. 4, (2015), 2162-2172.
- Yang, Bing, Zi-Jia Gong, and Wenqi Yang. "Stock market index prediction using deep neural network ensemble," Control Conference (CCC), 2017 36th Chinese. IEEE, 2017.
- Kaggle, Two Sigma Financial Modeling Challenge, Available at https://www.kaggle.com/c/two-sigma-financial-modeling/
- Kumar, Sunil, and Nivedita Deo, "Correlation and network analysis of global financial indices," Physical Review E, Vol. 86, No. 2, (2012).
- Baumeister, Christiane, and Gert Peersman. "The role of time−varying price elasticities in accounting for volatility changes in the crude oil market." Journal of Applied Econometrics, Vol. 28, No. 7, (2013), 1087-1109.
- Bernanke, Ben S. "Outstanding Issues in the Analysis of Inflation." Understanding Inflation and the Implications for Monetary Policy: A Phillips Curve Retrospective, (2009), 447.
- Pierdzioch, Christian, Jan-Christoph Rülke, and Georg Stadtmann. "Forecasting metal prices: Do forecasters herd?" Journal of Banking & Finance, Vol. 37, No. 1, (2013), 150-158.
- Kohlscheen, Emanuel, Fernando H. Avalos, and Andreas Schrimpf, "When the walk is not random: commodity prices and exchange rates.", International Journal of Central Banking, (2017).
- Bernard, Henri J., and Gabriele EB Galati, "III. Special feature: The co-movement of US stock markets and the dollar," BIS Quarterly Review, (2000).
- Giannarakis, Grigoris, et al., "The impact of Dow Jones Sustainability Index on US dollar value," International Journal of Economics and Financial Issues, Vol. 7, No. 2, (2017), 556-561.

Q & A