금융과 머신러닝

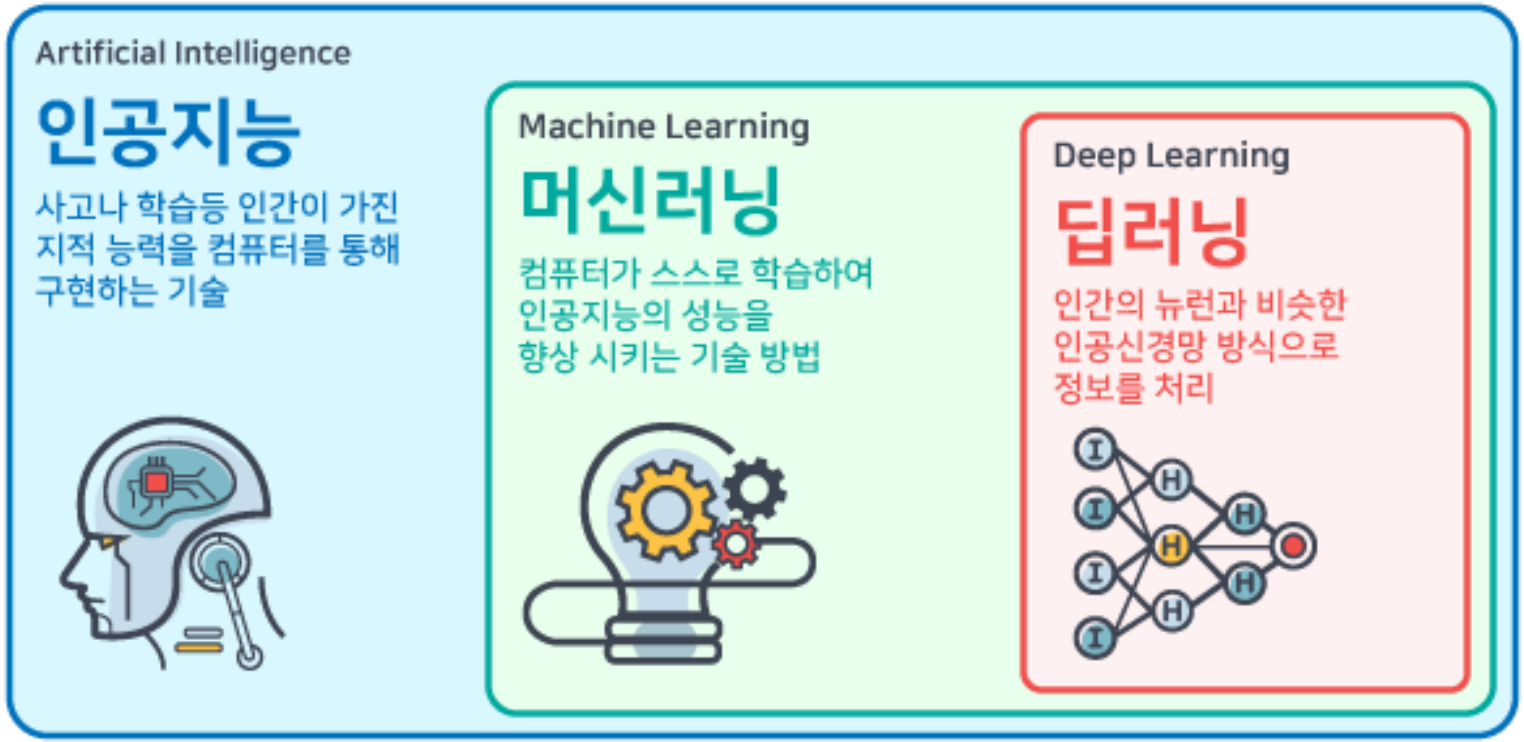
- 미시구조론의 함의를 바탕으로 정보추출 -

금융 내 머신러닝 활용 사례

- FDS(이상거래탐지 시스템)
- 로보어드바이저
- 대안신용평가

머신러닝이란?

컴퓨터가 짜여진 알고리즘에 따라 주어진 데이터가 가진 특징 및 패턴을 스스로 학습하여 도출해내는 것을 의미한다.



왜 머신러닝을 활용할까?

1. 금융데이터의 비선형적 특징

2. 계량경제학의 방법론적 한계

논문 소개

박석진, 정재식 (2019)

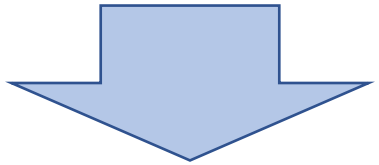
고빈도 자료를 이용한 머신러닝 모형의 예측력 비교·분석

: KOSPI200 선물시장을 중심으로

- 금융학회 우수논문상 수상

연구 주제

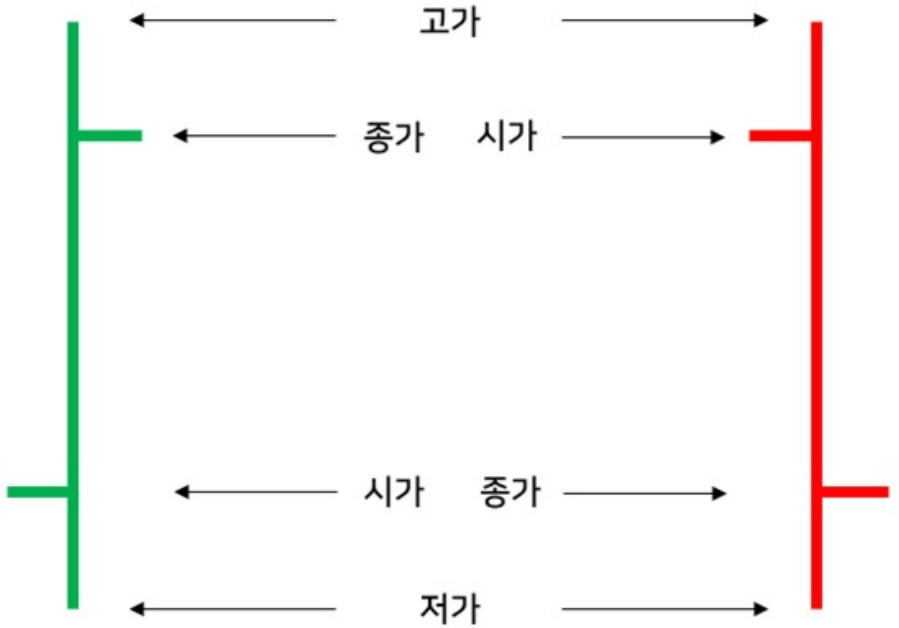
머신러닝에 사용되는 데이터를 어떻게 구성해야 하는가?



미시구조론적 함의를 지닌 VIB(Volume Imbalance Bar)를 활용

바(Bar)는 무엇인가?

- 시간 바
- 틱 바
- 거래량 바



VIB (Volume Imbalance bar)

- 정보 기반 바 (information driven bar)

  주문 흐름이 기대 수준 이상으로 매수 혹은 매도 한쪽으로 몰릴 때마다 바를 구성

- 미시구조론적 함의

  주문 흐름이 사적정보를 가진 정보우월자의 거래행위와 밀접한 관계가 있다는 미시적 토대와 관련됨

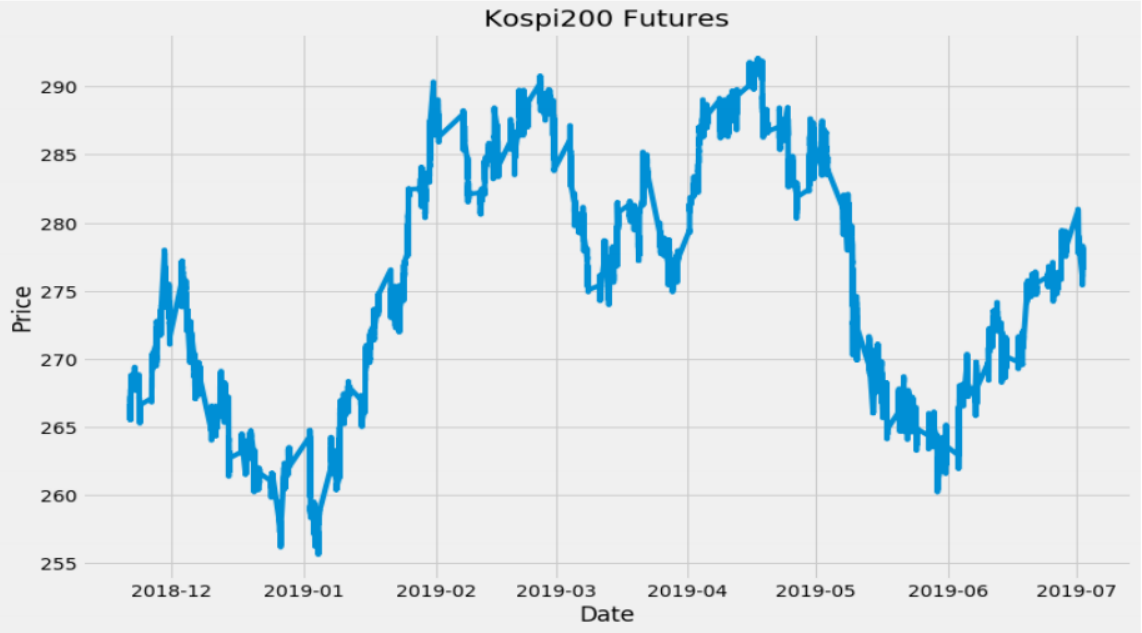
데이터 및 기초통계량

- KOSPI200 지수 선물의 실시간 체결 데이터

  (2018. 11. 21. ~ 2019. 7. 2. – 총 150일)

- 총 데이터 수 : 525만 6743개

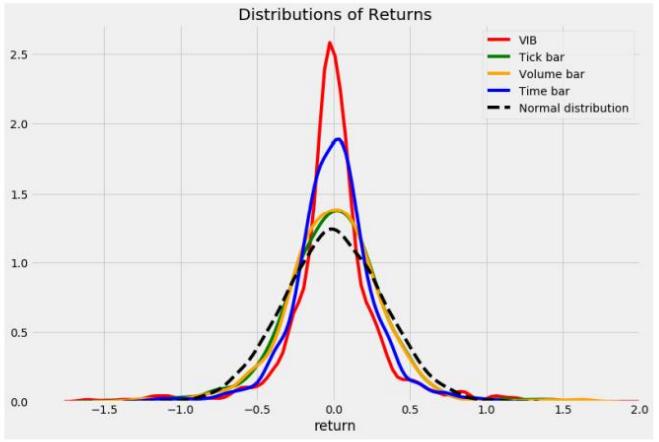
<Figure 5> Time series of Kospi200 Futures



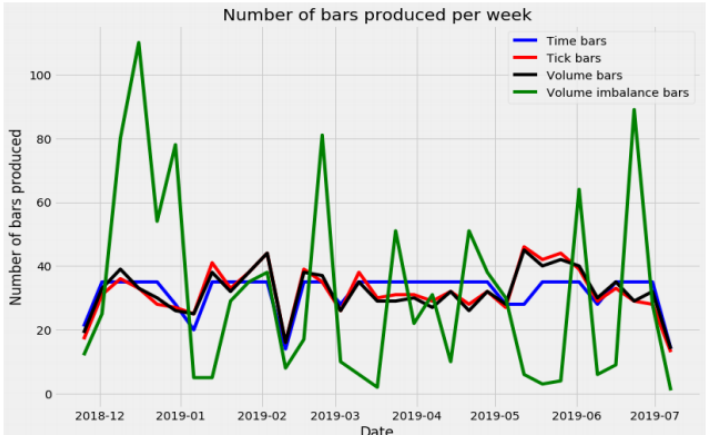
데이터 및 기초통계량

- VIB (1,038개)
- 시간 바 (1,049개)
- 틱 바 (1,050개)
- 거래량 바 (1,049개)

<Figure 7> Distributions of Returns



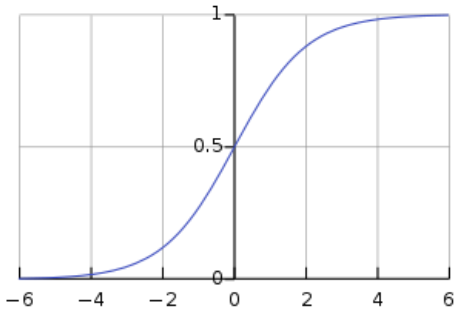
<Figure 6> Number of Bars Produced Per Week



예측 모형

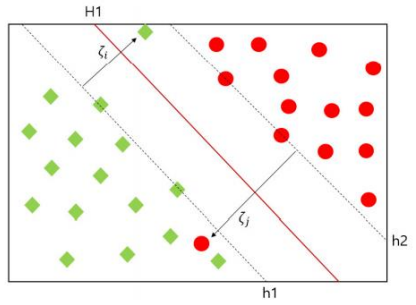
계량경제학 (벤치마크)

- 로지스틱 모형

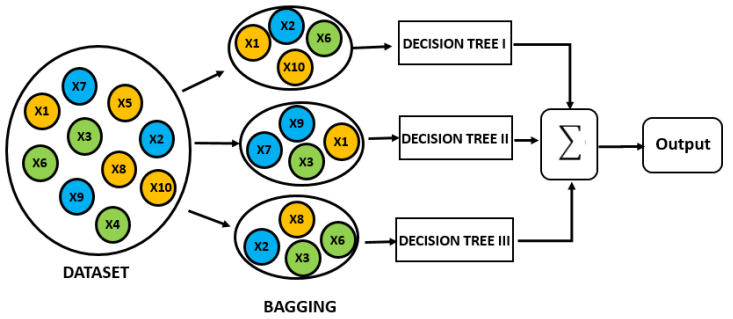


머신러닝

- SVM (Support Vector Model)



- RF (Random forest)



실증 분석

- 현재 바(bar)의 종가와 다음 바(bar)의 고가를 비교하여 가격의 방향성 판단

- 종가 대 종가 비교 시 경로 상의 정보 누락으로 인한 측정 오차 발생 가능성

- 바(bar) 형성 중 발생한 정보에 종가가 오염되어 측정 오차 발생 가능성

- 각 방향성의 비율이 비슷하도록 바(bar)의 임계치 설정

- 주어진 예측 모형에 따라 분석

결과의 평가지표

- 재현율(Recall) : 제1종 오류의 강건성
- 정밀도(Precision) : 제2종 오류의 강건성
- 정확도(Accuracy) : 예측모형의 전반적인 예측 성능

* 제1종 오류 : 실제값이 참인데 거짓으로 예측
* 제2종 오류 : 실제값이 거짓인데 참으로 예측

실증 분석 결과

- VIB외 바(bar)들의 정확도 : 50% 내외
- VIB의 정확도 : 최소 65%

<Table 4> Out of Sample Forecast Results

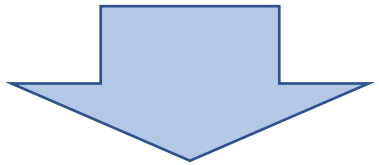
<Table 4> presents the out-of-sample forecast results of Logistic regression, SVM and Random forest. Forecast ability of each model is evaluated by Precision, Recall and Accuracy.

Bar type	Model	Precision	Recall	Accuracy
Time bar	Logistic regression	0.54	0.69	0.54
	SVM	0.52	0.59	0.52
	Random forest	0.54	0.58	0.54
Tick bar	Logistic regression	0.51	0.76	0.49
	SVM	0.51	0.79	0.50
	Random forest	0.49	0.54	0.48
Volume bar	Logistic regression	0.53	0.70	0.52
	SVM	0.52	0.76	0.51
	Random forest	0.50	0.52	0.48
VIB	Logistic regression	0.67	0.73	0.65
	SVM	0.70	0.66	0.65
	Random forest	0.69	0.76	0.68

실증 분석 결과

- VIB의 높은 정확도가 정보의 반영보다는 바(bar)의 형성 시간이 길기 때문일 수도 있다는 의심

- 바 형성 시간과 '고가-시가 ' 의 상관계수가 0.57
- 바 형성 시간에 패턴이 존재하면 정보 반영보다는 형성 패턴 파악으로 인한 예측



목적 변수 재설정 : 종가 대 종가

결과 비교 · 분석

- 데이터 사이즈가 작을 때는 로지스틱 모형과 머신러닝 모형의 정확도에서 큰 차이가 나지 않음

- 데이터 사이즈가 커질수록 머신러닝의 정확도가 상대적으로 더 커짐

<Table 7> Out of sample forecast results based on rolling window size.

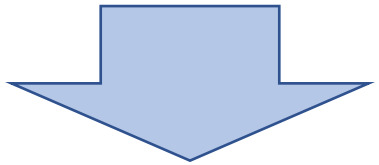
<Table 7> presents out of sample forecast results of different rolling window size. All results are based on VIB(Volume imbalance bar).

Window size	Model	Precision	Recall	Accuracy
50	Logistic regression	0.62	0.62	0.61
	SVM	0.60	0.56	0.59
	Random forest	0.63	0.63	0.62
100	Logistic regression	0.62	0.65	0.62
	SVM	0.63	0.56	0.60
	Random forest	0.64	0.63	0.62
200	Logistic regression	0.65	0.67	0.63
	SVM	0.66	0.64	0.64
	Random forest	0.66	0.67	0.64
300	Logistic regression	0.67	0.69	0.65
	SVM	0.69	0.64*	0.65
	Random forest	0.68	0.69	0.66
400	Logistic regression	0.67	0.73	0.65
	SVM	0.70	0.66	0.65
	Random forest	0.69	0.76	0.68
500	Logistic regression	0.68	0.71	0.65
	SVM	0.72	0.67	0.67
	Random forest	0.72	0.74	0.70
600	Logistic regression	0.69	0.75	0.68
	SVM	0.74	0.70	0.70
	Random forest	0.72	0.73	0.70
700	Logistic regression	0.68	0.76	0.67
	SVM	0.73	0.71	0.69
	Random forest	0.72	0.71	0.69
800	Logistic regression	0.72	0.79	0.69
	SVM	0.78	0.76	0.73
	Random forest	0.77	0.74	0.71
900	Logistic regression	0.66	0.81	0.70
	SVM	0.74	0.78	0.75
	Random forest	0.74	0.72	0.74

결 론

- 미시구조론의 함의를 활용하여 데이터를 가공하였을 때 더욱 정확한 예측이 가능함

- 원자료를 사용하더라도 어떤 기준에 따라 바를 구성하냐에 따라 머신러닝의 예측력이 유의미하게 차이남



미시구조론 + 머신러닝

감사합니다.