빅데이터마이닝:머신러닝 과제 3
머신러닝의 구성요소

데이터테크전공

20173204 곽명빈

서론

국가 · 기업 · 가계가 보유한 자산 중에서 가장 큰 비중을 차지하는 것이 부동산이다. 부동산에 편중된 자산 구조로 인해 부동산 가격 변동은 국가 · 기업 · 가계의 경제상황에 큰 영향을 미치게 된다. 이로 인해 부동산 가격의 상승 또는 하락 여부는 주요 관심사항이며, 부동산 가격 변화에 대비하기 위해 다양한 방법을 이용한 부동산 시장 예측이 시도되고 있다. 부동산 시장 예측은 주로 시계열분석 모형을 이용하여 부동산 가격지수를 예측하는 방식으로 이루어진다. 하지만 시계열분석 모형은 선형 모형을 가정하기 때문에 비현실적이고 예측 효율성이 떨어진다는 문제점이 있어 새로운 분석방법 적용의 필요성이 제기되고 있다 (배성완 · 유정석, 2017). 최근 주목받고 있는 머신 러닝(machine learning) 방법은 비선형 추정기법으로 분류(classification)와 회귀(regression)분야에서 활발한 연구와 좋은 성과를 보여주고 있다는 점에서 부동산 가격지수 예측과 관련해서도 활용 가능성이 높을 것으로 기대된다.[1)]

시계열 모형대신 머신 러닝을 이용하여 부동상 가격 지수 예측

시계열분석 모형과 머신 러닝의 예측력을 비교 분석함

경험

머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측

본 연구를 위한 분석 자료로서 종속변수는 부동산 가격지수인 아파트 매매실거래가격지수를 이용하였고, 설명변수는 회사채수익률, 소비자물가지수, 통화량, 광공업지수를 이용하였다. 분석지역은 서울지역, 분석기간은 2006년 1월부터 2017년 8월까지로 설정하였다.

Training data set은 아파트 매매실거래가격지수 ~ 회사채수익률, 소비자물가지수, 통화량, 광공업지수를 이용하였다.

아파트 매매실거래가격지수에 영향을 주는 변수들로 회사채수익률, 소비자물가지수, 통화량, 광공업지수를 사용한다는 의미 → 2006 ~ 2017년까지의 서울 아파트 매매 실거래가격지수가 경험이라고 할 수 있음

지역과 기간을 제한하여 분석을 진행

경험

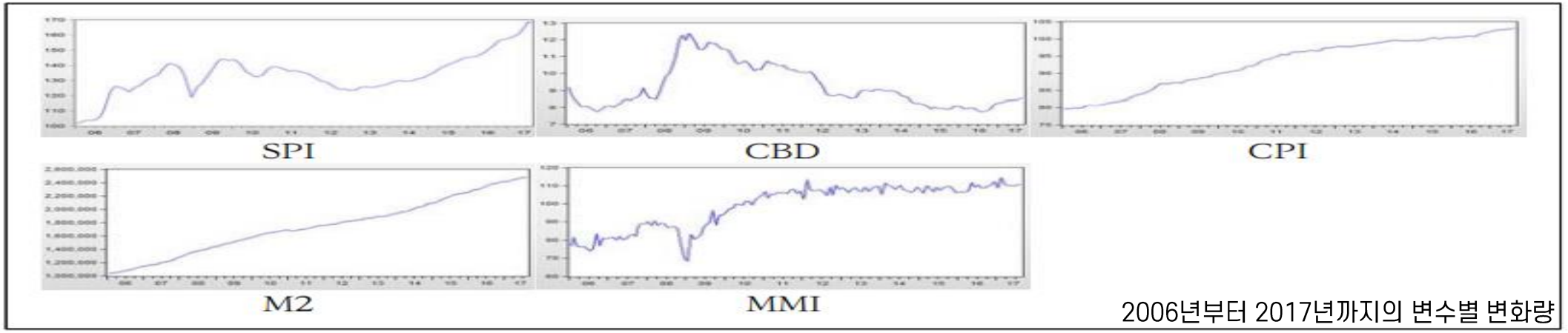
머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측

〈표 1〉 기초통계량

구분		평균	중위수	최대값	최소값	표준편차
SPI	아파트 매매실거래가격지수	133.964	133.100	169.800	100.000	12.973
CBD	회사채수익률	9.232	8.750	12.400	7.720	1.274
CPI	소비자물가지수	93.118	95.574	103.480	79.306	7.288
M2	통화량	1,744,478	1,747,971	2,485,630	1,027,697	410,163
MMI	광공업지수	98.870	103.950	118.000	67.832	12.814

회사채 수익률을 제외한 모든 변수가 분석기간동안 상승하는 추세

아파트 매매실거래가격지수와 광공업지수는 2008년(금융위기)이후 급락하는 모습을 보임



2006년부터 2017년까지의 변수별 변화량

업무

머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측

본 연구는 시계열분석 방법과 머신 러닝 방법의 예측력을 비교하여, 머신 러닝 방법의 실제 활용가능성을 검토하는 것이 목적이다. 시계열분석 모형 중에서는 단변량 시계열분석 모형인 ARIMA모형, 다변량 시계열분석모형인 VAR모형, 베이지언 VAR모형을 이용하였다. 베이지언 VAR모형은 모수에 대한 사전적인 제약방법에 따라 4가지 모형으로 분류된다. 머신 러닝 방법은 SVM, RF, GBRT, DNN, LSTM모형을 이용하였으며 단변량 시계열

다. 머신 러닝 방법 별로 초모수를 변화시키면서 k겹 교차검증에 의해 산출된 평균절대값 오차(mean absolute error, MAE) 및 평균제곱근오차(root mean square error, RMSE)의 평균값이 가장 낮은 모형을 각 방법별 최적 모형으로 결정하였다. 〈그림 3〉은 k겹 교차

머신러닝 모델로 SVM, RF GBRT, DNN, LSTM모형을 이용

SVM
support vector machine

RF
random forest

GBRT
gradient boosting
regression tree

DNN
deep neural networks

LSTM
Long Short Term
Memory networks

업무

머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측

기간1

있다. 이에 따라 기간 1은 '2006년 1월~2016년 8월(128개월)'을 학습(train) 데이터, 안정적인 상승추세를 보여주고 있는 '2016년 9월~2017년 8월(12개월)'을 시험(test) 데이터

기간2

로 설정하였고, 기간 2는 '2006년 1월~2008년 8월(32개월)'을 학습 데이터, 구조적인 변화 또는 시장 충격으로 시장이 급변하는 모습을 보이고 있는 '2008년 9월~2009년 8월(12개월)'을 시험 데이터로 설정하여 시장 상황에 따른 모형별 예측력 차이를 비교·분석하였다.

Train데이터와 test 데이터를 안정적인 상승추세는 기간1로

급변하는 모습을 보이는 기간은 기간2로 배정하여

시장 상황에 따른 모형별 예측력 차이를 비교 분석함

척도

머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측

기간 1

구분		MAE	RMSE	초모수 설정
SVM	단변량(U)	0.019945	0.024170	C=2, γ=0.3, ϵ=0.05
	다변량(M)	0.019083	0.023187	C=6, γ=0.2, ϵ=0.05
RF	단변량(U)	0.010582	0.014256	트리수 = 100
	다변량(M)	0.005912	0.007409	트리수 = 100
GBRT	단변량(U)	0.009445	0.011699	트리수=20
	다변량(M)	0.006705	0.007614	트리수=10
DNN	단변량(U)	0.008536	0.009495	hidden layer node: 20-20-20
	다변량(M)	0.006217	0.007594	hidden layer node: 50-50-50
LSTM	단변량(U)	0.005239	0.007033	노드=20
	다변량(M)	0.011155	0.014922	노드=20

MAE와 RMSE는 낮을수록 모델의 성능이 좋다고 판단 => 오차가 작기 때문

LSTM모델이 안정적인 상황에서 가장 잘 예측한다고 할 수 있음

척도

머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측

기간 2

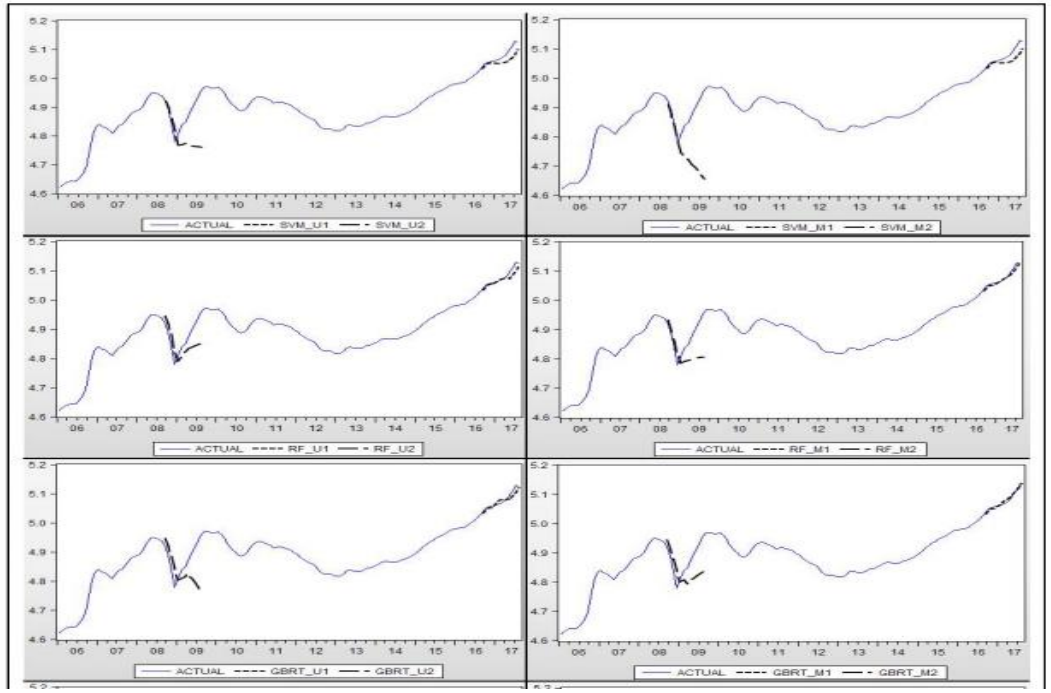
구분		MAE	RMSE	초모수 설정
SVM	단변량(U)	0.084689	0.105881	C=2, γ=0.1, ϵ=0.01
	다변량(M)	0.124583	0.161247	C=2, γ=0.1, ϵ=0.05
RF	단변량(U)	0.049636	0.056236	트리수 = 200
	다변량(M)	0.064522	0.079009	트리수 = 200
GBRT	단변량(U)	0.069382	0.088423	트리수=20
	다변량(M)	0.059743	0.068923	트리수=10
DNN	단변량(U)	0.055915	0.064111	hidden layer node : 20-20-20
	다변량(M)	0.088034	0.109266	hidden layer node : 200-200-200
LSTM	단변량(U)	0.091358	0.097444	노드=20
	다변량(M)	0.038145	0.050033	노드=200

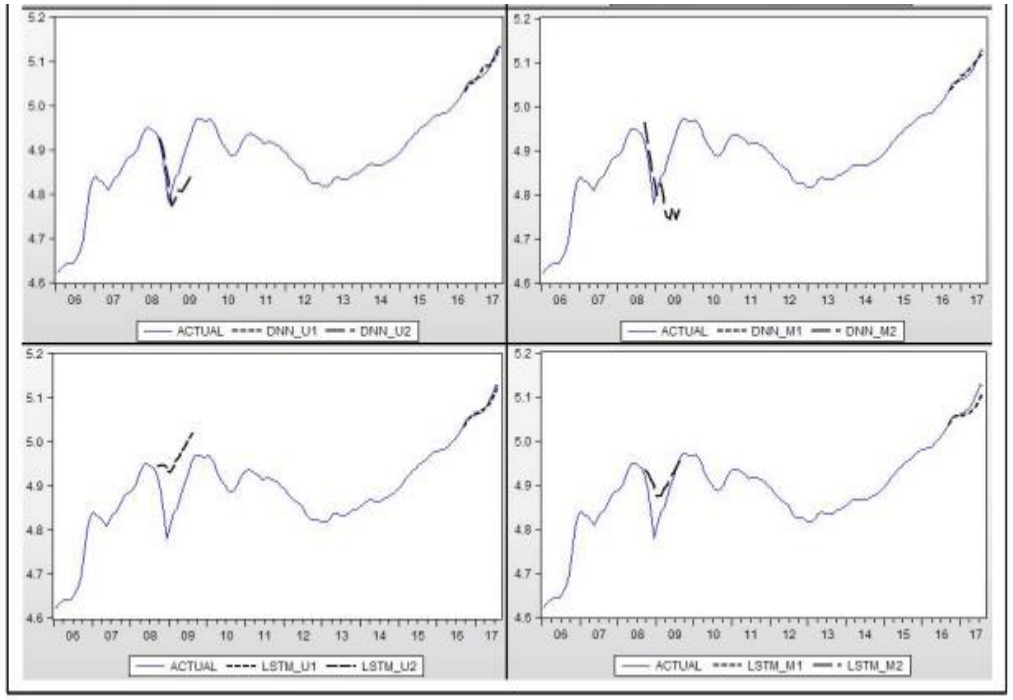
MAE와 RMSE는 낮을수록 모델의 성능이 좋다고 판단 => 오차가 작기 때문

LSTM모델이 안정적인 상황에서 가장 잘 예측한다고 할 수 있음

척도

머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측



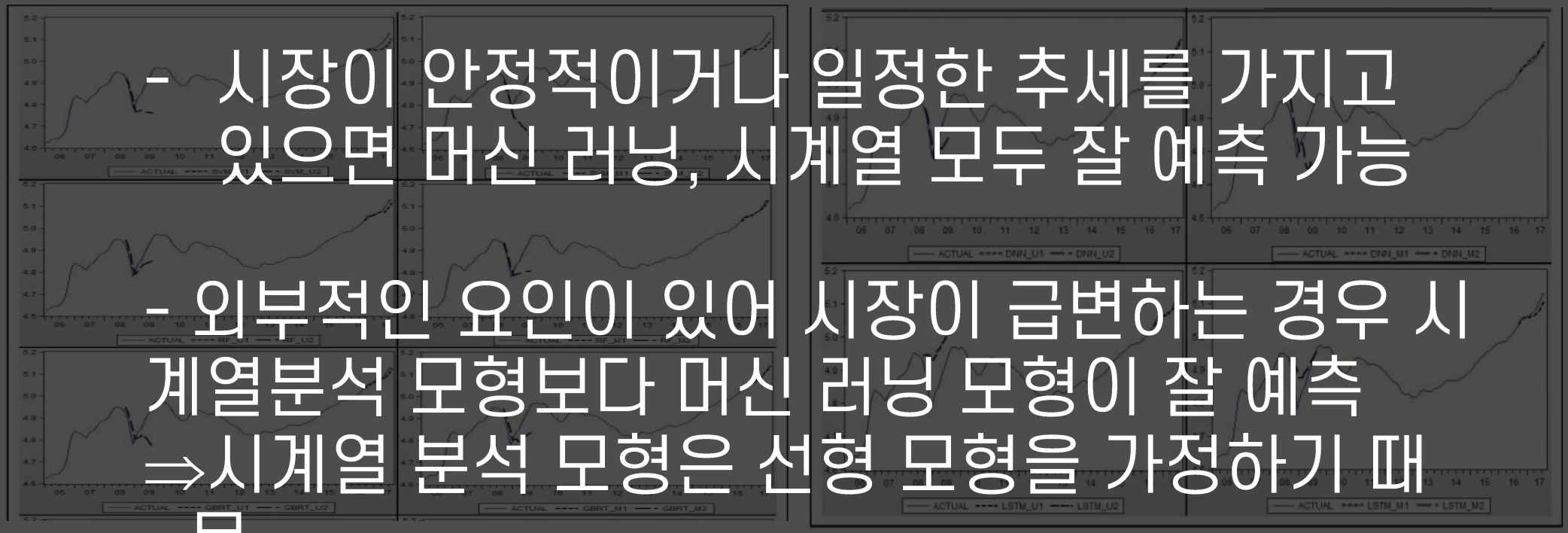


기간 1 = 예측값과 실제 데이터가 거의 일치
기간 2 = 예측값과 실제 데이터간 다소차이, 일부 모형은 상당히 유사
하락추세는 정확히 예측하는 반면 반등 후 상승 하는 부분에서 차이를 보임

척도

결론

머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측



- 시장이 안정적이거나 일정한 추세를 가지고 있으면 머신 러닝, 시계열 모두 잘 예측 가능

- 외부적인 요인이 있어 시장이 급변하는 경우 시계열분석 모형보다 머신 러닝 모형이 잘 예측
⇒시계열 분석 모형은 선형 모형을 가정하기 때문

기간 1 = 예측값과 실제 데이터가 거의 일치
기간 2 = 예측값과 실제 데이터간 다소차이, 일부 모형은 상당히 유사
하락추세는 정확히 예측하는 반면 반등 후 상승 하는 부분에서 차이를 보임

척도

질병 진단 시스템

		실제 결과	
		양성	음성
모델1로 예측한 결과	양성	400	50
	음성	100	450

		실제 결과	
		양성	음성
모델2로 예측한 결과	양성	450	100
	음성	50	400

모델 1
정확도: 85% 민감도: 80% 특이도: 90% 정밀도:89%

모델 2
정확도: 85% 민감도: 90% 특이도: 80% 정밀도:82%

척도

질병 진단 시스템

2개의 모델을 비교하였을 때 어떤 모델이 더 낫다고 이야기할 수 있는지와 그 이유 를 설명하시오.

모델 1과 2의 정확도는 85% 로 같기 때문에, 민감도와 특이도, 정밀도를 통해 판단해 보았다.

민감도란 양성을 예측하는 비율로 모델1보다 모델2가 좋다고 할 수 있다.

특이도란 음성을 예측하는 비율로 모델1이 모델2보다 좋다고 할 수 있다.

두 모델을 비교하여 좋고 나쁨을 판단할 수 없다고 생각한다.

모델1을 선택하였을 경우 음성인데 양성이라고 하여 잘못된 처방을 받을 확률이 높아 위험하고, 모델2의 경우에는 양성인데 음성판정을 받아 처방을 받지 못할 확률이 높다.

두 상황모두 위험한 상황이기 때문에 더 나은 모델이 없다고 생각한다.

결론

느낀점

머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측을 바탕으로 경험, 업무, 척도의 개념을 살펴보았다

- 경험의 경우 데이터를 선택하는 단계이기 때문에 잘못된 선택을 한다면, 뒤의 업무, 척도의 결과가 좋지 못하기 때문에 데이터 선택을 신중히 해야 한다는 생각 하게 되었다.
- 업무의 경우 모델링을 하는 작업인데, 데이터에 다양한 모델을 사용하여 제일 예측을 잘하는 모델을 선택해야 한다는 것을 알게 되었다.
- 척도의 경우 다양한 머신 러닝 모델 평가 방법이 있다는 사실을 알았고, 모델에 맞는 것을 쓰면 된다는 것을 알게 되었다.
- 질병진단시스템을 살펴보면서 더 좋은 모델을 평가 하는게 어려운 상황도 있을 수 있다는 사실을 깨닫고 모델 선택함에 있어 신중해야 한다고 생각하였다.

출처

배성완, & 유정석. (2018). 머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측. *주택연구*, *26*(1), 107-133.