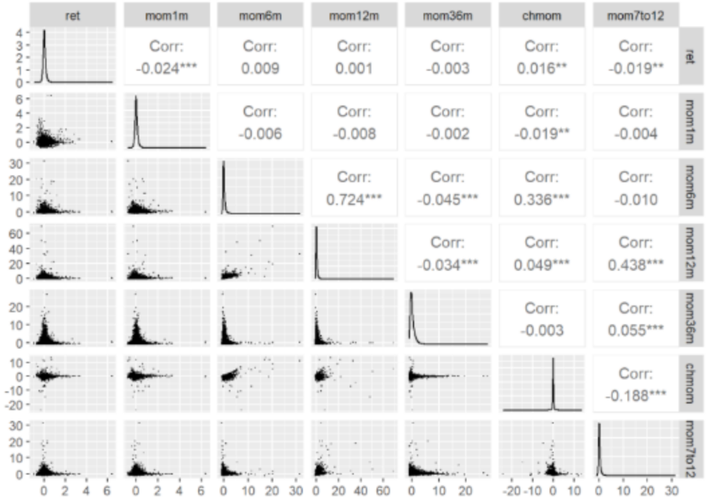
통계분석실습 중간 프로젝트

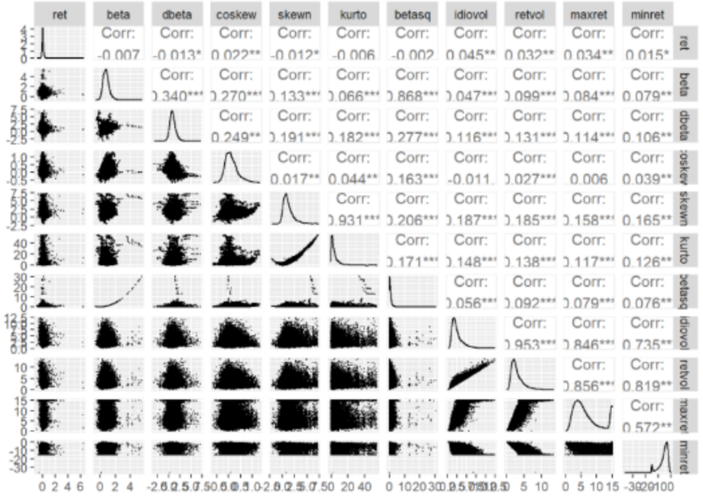
KOSPI 기업의 수익률 예측

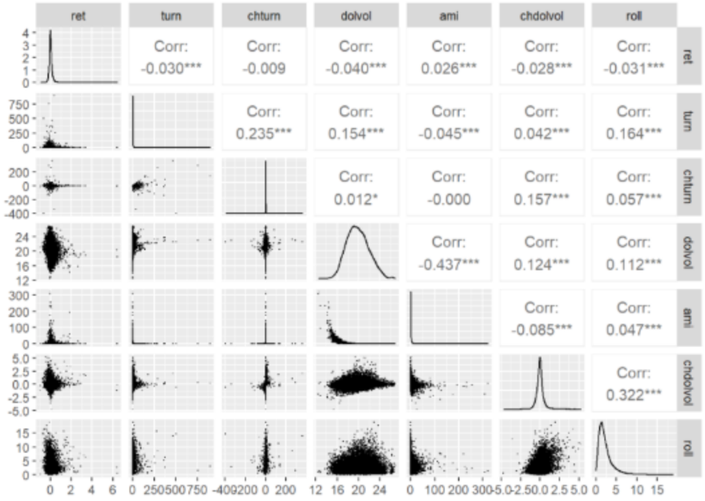
2조 1916129 통계학과 최윤서

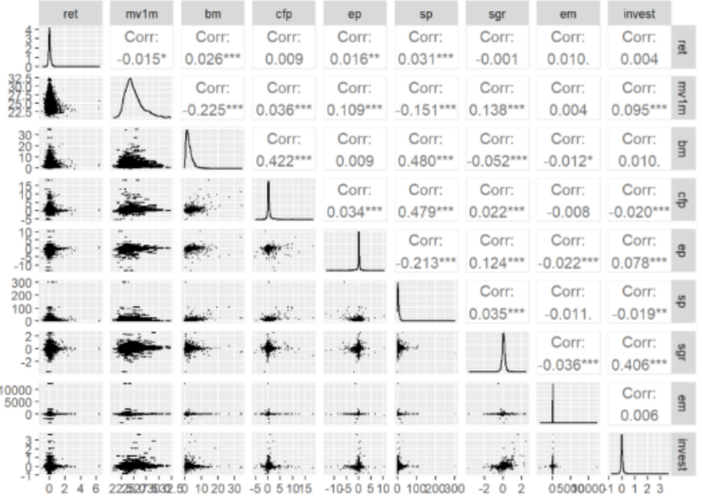
데이터 탐색

첫번째 train-set을 기준으로 데이터 탐색을 진행(2002 ~ 2006)









1. 눈에 띄게 선형관계를 가진 변수가 없어보임.

2. 산점도를 살펴봤을 때도 강력하게 유의미한 관계를 가진 변수를 찾기 어려움.

3. 유의한 상관계수를 가지는 변수가 존재하고, 몇몇 변수들의 산점도에서 약간의 설명력을 확인해볼 수 있지만, 영향력이 매우 작아서 특정 변수만으로 y의 변동을 설명하기 어려워 보임.

주식데이터의 특성 고려

01 시계열 데이터

주식데이터는 시간의 흐름 순으로 작성된 데이터 -> 각 시기별로 가진 특성이 다름 => **해당 시점마다 다른 모델을 선택하여 예측**

자료의 설명변수는 이전 시점의 정보를 담고 있음 => **훈련데이터와 검증데이터를 나눌 때 시점을 기준으로 앞 뒤로 분할**

02 많은 고려 사항

실제 사람이 기업을 평가하고, 주식을 구매할 때는 정말 많은 정보를 고려해서 투자함.

ex) 해당 데이터셋에 존재하는 정보 + 해당 시점의 뉴스 기사 + 기업의 사업 분야 등

-> 즉, 주어진 데이터셋에도 포함되지 않은 많은 정보들이 존재

=> 주어진 데이터셋의 모든 정보를 최대한 사용하고자 함.

분석 진행 개요

01

단일 모델 적합

02

모델 후보 선택

03

시기별 모델 선택

04

최종 모델 선정

회귀 모델 예측 결과 비교

train-set : 5년, test-set: 1년

단일 모형

Training					Test
2002	2003	2004	2005	2006	2007

Training					Test
2003	2004	2005	2006	2007	2008

.
.
.

Training					Test
2013	2014	2015	2016	2017	2018

<단일 모형으로 적합 후 예측한 결과>

	Avg	Std	SR
Linear regression	0.87	6.73	0.45
median linear regression	0.92	7.95	0.4
Random Forest	1.31	7.31	0.62
Xgboost	1.52	7.66	0.69
Ridge	0.91	6.92	0.46
Lasso	0.7	6.65	0.36

회귀 모델 예측 결과 비교

train-set : 5년, test-set: 1년

Random Forest 과적합 문제

year	Training-set mse	Test-set mse
2007	0.00672	0.0355
2008	0.00686	0.0391
2009	0.00708	0.0359
2010	0.00708	0.0249
2011	0.00571	0.0242
2012	0.00575	0.0256
2013	0.00544	0.0163
2014	0.00494	0.0142
2015	0.00419	0.0255
2016	0.00132	0.0168
2017	0.00399	0.0144
2018	0.00347	0.0247

=> 훈련세트와 테스트세트의 mse가 과하게 차이남.

=> 최종 선택 모형에서 제외

모델 선택 (후보 4개)

train-set : 5년, test-set: 1년

mse(mae)를 기준으로 각 시기별로 모델을 선택한 뒤 적합

Training				Valid	Test
2002	2003	2004	2005	2006	2007

Training				Valid	Test
2003	2004	2005	2006	2007	2008

▪
▪
▪

Training				Valid	Test
2013	2014	2015	2016	2017	2018

<모델 후보>

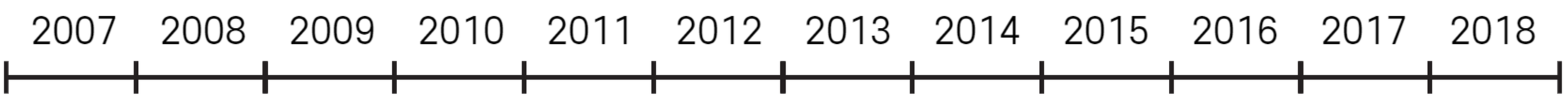
1. linear regression
2. median linear regression
3. ridge
4. boost

<결과값>

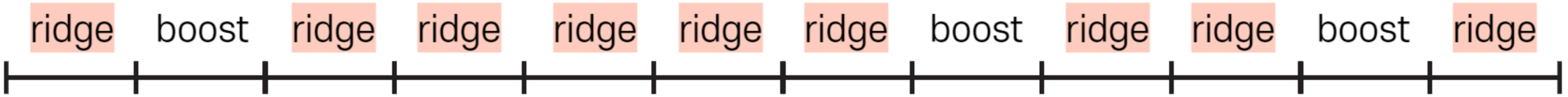
	Avg	Std	SR
모델선택_mse 기준	1	6.77	0.51
모델선택_mae 기준	0.32	7.78	0.14
단일모델_Xgboost	1.52	7.66	0.69

모델 선택 (후보 4개)

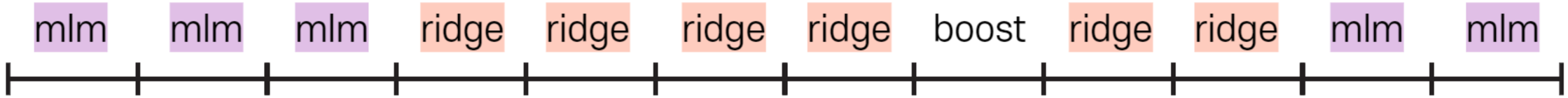
train-set : 5년, test-set: 1년



- mse 기준 시기별 선택 모델 , SR : 0.51



- mae 기준 시기별 선택 모델 , SR : 0.14



* mlm: median linear regression

=> ridge와 median linear regression 모델이 많이 선택됐을 때 성능이 낮아짐.

모델 선택 (후보 4개)

train-set : 5년, test-set: 1년

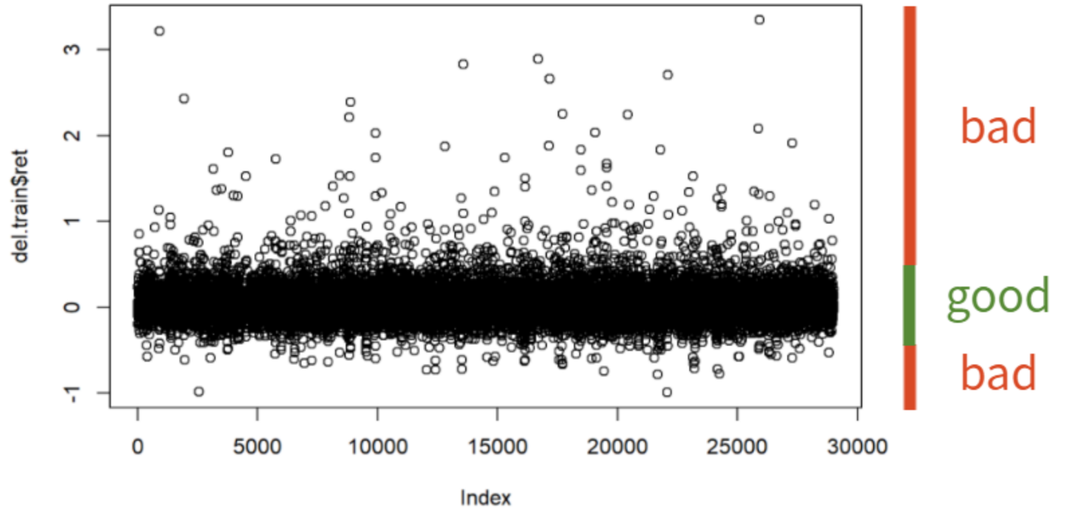
두 가지 모델을 후보에서 제외한 가설

1. Ridge 제외

good bad

Training				Valid	Test
2002	2003	2004	2005	2006	2007

2. median linear regression 제외



모델 선택 (후보 2개)

train-set : 5년, test-set: 1년

mse(mae)를 기준으로 각 시기별로 모델을 선택한 뒤 적합

Training				Valid	Test
2002	2003	2004	2005	2006	2007

Training				Valid	Test
2003	2004	2005	2006	2007	2008

▪
▪
▪

Training				Valid	Test
2013	2014	2015	2016	2017	2018

<모델 후보>

1. linear regression
2. boost

<결과값>

	Avg	Std	SR
모델선택(후보4개)_mse 기준	1	6.77	0.51
모델선택(후보4개)_mae 기준	0.32	7.78	0.14
모델선택(후보2개)_mse 기준	1.02	7.47	0.47
모델선택(후보2개)_mae 기준	1.64	7.7	0.74
단일모델_Xgboost	1.52	7.66	0.69

모델 선택 (후보 2개) - 예측기간 축소

train-set : 5년, test-set: 6개월

mse(mae)를 기준으로 각 시기별로 모델을 선택한 뒤 적합

6개월

Training				Valid	Test
2002	2003	2004	2005	2006	2007

Training				Valid	Test
2002.5	2003.5	2004.5	2005.5	2006.5	2007.5

Training				Valid	Test
2003	2004	2005	2006	2007	2008

.
.
.

Training				Valid	Test
2013.5	2014.5	2015.5	2016.5	2017.5	2018.5

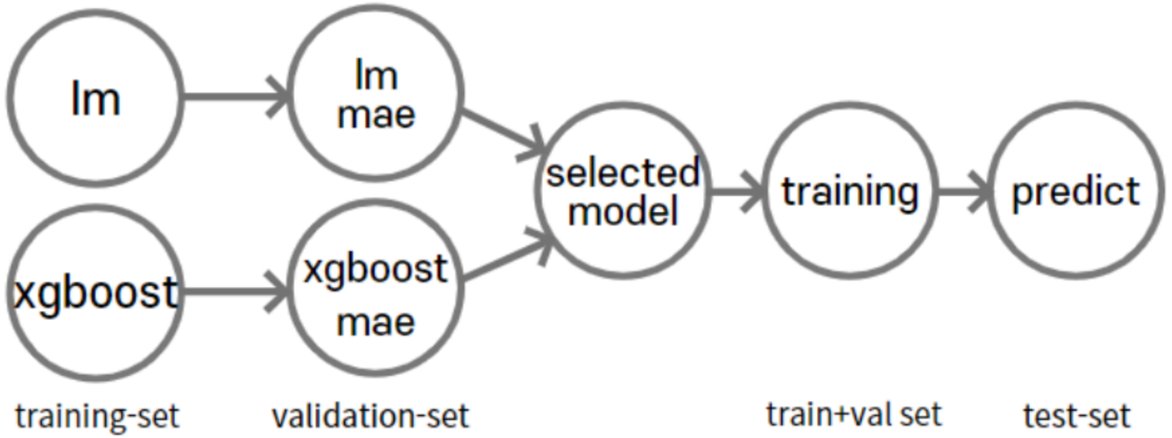
<모델 후보>

1. linear regression
2. boost

<결과값>

	Avg	Std	SR
모델선택(1년)_mse 기준	1.02	7.47	0.47
모델선택(1년)_mae 기준	1.64	7.7	0.74
모델선택(6개월)_mse 기준	1.9	7.37	0.89
모델선택(6개월)_mae 기준	2.29	7.57	1.05

최종 모델



1. 훈련데이터에서 linear regression과 xgboost 모델 적합
2. 1에서 적합한 모델로 검증데이터의 mae값 계산
3. 두 모형 중 더 낮은 검증데이터 mae값을 가지는 모델을 최종 모형으로 선정
4. 최종 선정된 모형을 전체 훈련세트(훈련+검증)에 대해 다시 학습
5. 4에서 학습된 모델로 테스트데이터 예측

6개월

Training				Valid	Test
2002	2003	2004	2005	2006	2007

Training				Valid	Test
2002.5	2003.5	2004.5	2005.5	2006.5	2007.5

Training				Valid	Test
2003	2004	2005	2006	2007	2008

.
.
.

Training				Valid	Test
2013.5	2014.5	2015.5	2016.5	2017.5	2018.5

보완할 점

01 설명변수의 구성

데이터에 대한 이해가 부족하여 모델선택에만 초점을 맞춰 진행한 점이 아쉬움.

가능한 특성 변환(log(x), sqrt(x), x^2 등)을 추가해본 뒤, 주성분 분석, ridge, lasso 방법을 사용해보았다면 좋았을 듯함.

02 후보 모델 개수가 적음: Randomforest 모델 사용

분석 초반 randomforest를 과적합(overfitting)문제로 모델을 선택후보에서 제외시킴.

hyperparameter를 조정하여 과적합 문제를 해결해보고, 모델 선택 후보에 포함했다면 좋았을 듯함.

질의응답