

13주차 회의록

회의일시	2020년 11월 29일 (일요일) 14:00 ~ 15:30	장소	온라인
팀원	박채린, 이해승		
회의안건	<ul style="list-style-type: none"> 전처리 및 분할방법마다 10-fold cv 및 t검정을 진행하여, 최고성능의 데이터 및 모델 선택 생성한 경제 주간 지수의 유효성 입증 방법 		

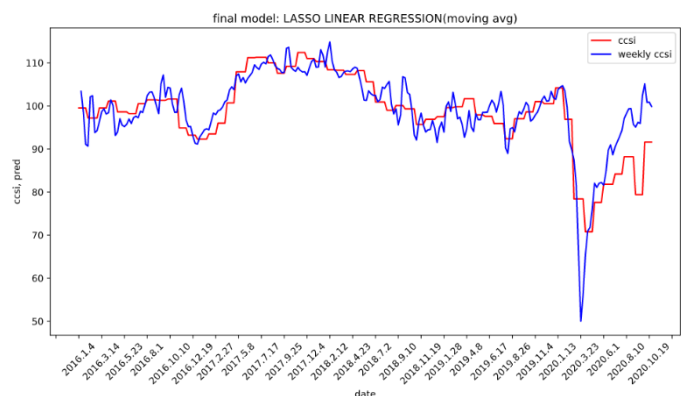
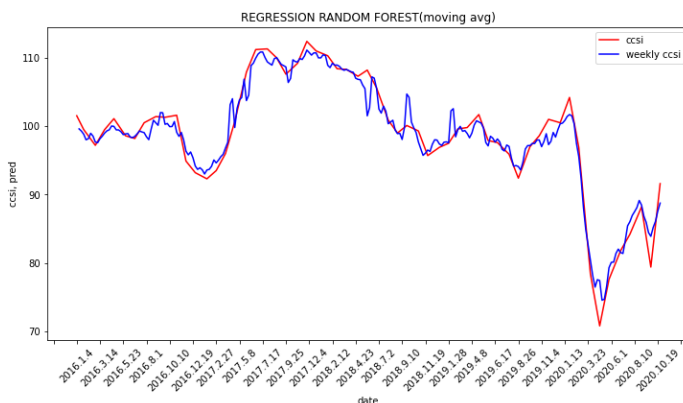
	Shuffle = True	Shuffle = False
Mean	LR Test RMSE: 4.006 Test MAE: 3.122	Lasso(alpha = 0.001) Test RMSE: 8.438 Test MAE: 2.568
Median	LR Test RMSE: 6.203 Test MAE: 3.664	Lasso(alpha = 0.001) Test RMSE: 9.336 Test MAE: 2.605
interpolation	RF(n_estimators=32, max_features=4) Test RMSE: 3.490 Test MAE: 2.483	RF (max_features=3, n_estimators=64) Test RMSE: 13.889 Test MAE: 3.381

전처리 방법과 분할 방법에 따라 총 6가지 방법으로 구분한 후, 각 방법에서 10-fold CV 및 T검정을 진행하여 가장 좋은 성능을 보이는 모델들을 선택했다.

Shuffle = True는 데이터셋을 임의로 섞은 후, train test를 분할한 것이고, False는 2016년~2019년까지를 트레이닝셋으로 2020년을 테스트셋으로 분할한 것이다.

각각 셔플 방법에서 제일 좋은 성능을 보이는 두 모델을 선택했다. (issue: 각 방법에서 사용하는 train, test셋이 다른데, 그냥 test rmse 및 test mae로 비교해도 될까..?)

두 모델의 시각화 결과, interpolation – shuffle 방법이 더 예측력이 좋았다.



➔ 최종 전처리방법

전처리: 내삽

Train, test 분할 방법: Shuffle

➔ 최종 모델: RF(n_estimators=32, max_features=4)

(오늘 할일)

- ✓ 11월 데이터를 predict에서만 활용하기(11월포함한 테이블을 다시 생성해서, step5, 6!) – 채린
- ✓ 주간지표의 유효성 입증 – ccsi 시각화, 전체레코드에 대한 rmse, CCI 순환변동치를 선행하는지 확인.(그래프 시각화) – 혜승[[코드](#)]
- ✓ 지수 이름을 붙여보기...

13 주차(11/23 ~ 29)	<ul style="list-style-type: none">● 전처리 및 분할방법에 따른 best model 총 6개 생성 후, test 결과 시각화 및 모델 평가(RMSE 비교)● 주간 지표의 유효성 입증● 보고서 작성
14 주차(11/30 ~ 12/6)	11/30 교수님 미팅 <ul style="list-style-type: none">● 발표 준비 및 영상 제작
15 주차(12/7 ~ 12/13)	12/7 교수님 미팅 12/11 프로젝트 발표 영상 제출

<interpolation>

		Shuffle = False
Linear regression – random forest	T 통계량	3.595
	기각 여부	기각(mean 1.161)
	선택 모델	Random forest
Linear regression – GAM	T 통계량	-1.02
	기각 여부	채택
	선택 모델	
Linear regression – lasso	T 통계량	1.727
	기각 여부	채택
	선택 모델	
Random forest – GAM	T 통계량	-3.773
	기각 여부	기각(mean -1.432)
	선택 모델	Random forest
Random forest – lasso	T 통계량	-2.820
	기각 여부	기각(mean -0.560)
	선택 모델	Random forest
GAM – lasso	T 통계량	2.454
	기각 여부	기각
	선택 모델	Lasso

$$t_{0.025,9} = 2.262$$

<parameter>

- Linear regression
- Random forest:
12번 모델이 best(max_features=3, n_estimators=64, random_state=23, warm_start=True)
- GAM
- Lasso liner regression:
3번 모델이 best(alpha=0.1, random_state=23)

<rmse>

- Linear regression: 3.9692318060522425
- Random forest: 2.8083498720391513
- GAM: 4.240391592522476
- Lasso liner regression: 3.367514131252989

<test RMSE>

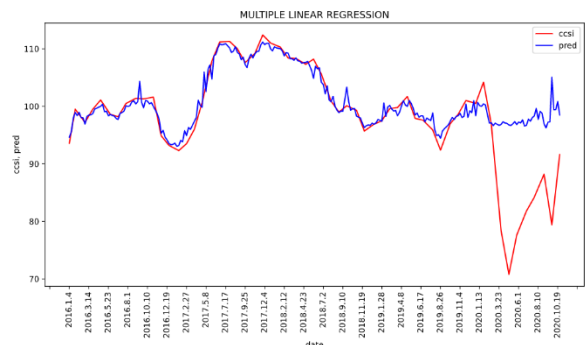
13.889

<test MAE>

3.381

<weekly 단기지표 값에 대한 RMSE>

6.314



<median>

		Shuffle = False
Linear regression – random forest	T 통계량	2.216
	기각 여부	채택
	선택 모델	
Linear regression – GAM	T 통계량	-3.193
	기각 여부	기각(mean -3.491)
	선택 모델	Linear regression
Linear regression – lasso	T 통계량	2.213
	기각 여부	채택
	선택 모델	
Random forest – GAM	T 통계량	-5.023
	기각 여부	기각(mean -4.501)
	선택 모델	Random forest
Random forest – lasso	T 통계량	-0.910
	기각 여부	채택
	선택 모델	
GAM – lasso	T 통계량	4.297
	기각 여부	기각(mean 4.298)
	선택 모델	lasso

$t_{0.025,9} = 2.262$

<parameter>

- Linear regression
- Random forest:
18번 모델이 best(max_features=5, n_estimators=16, random_state=23, warm_start=True)
- GAM
- Lasso linear regression:
5번 모델이 best(alpha=0.001, random_state=23)

<rmse>

- Linear regression: 3.50080569595515
- Random forest: 2.490620830568168
- GAM: 7.420461250157842
- Lasso linear regression: 2.693533915991976

<test RMSE>

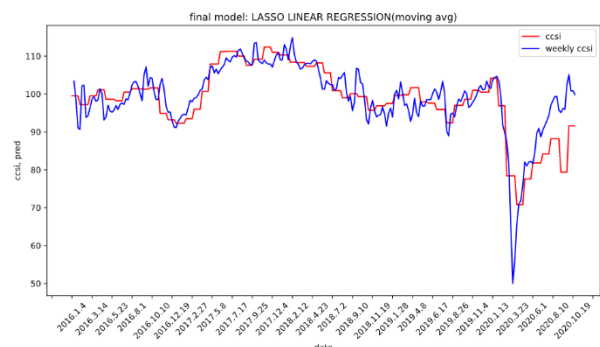
Model1: 13.226 / Model2: 9.347 / Model3: 9.336

<test MAE>

Model1: 3.323 / Model2: 2.606 / Model3: 2.605

➔ Final로 lasso linear regression 선택

<weekly 단가지표 값에 대한 RMSE> 5.867



<mean>

		Shuffle = False
Linear regression – random forest	T 통계량	2.407
	기각 여부	기각(mean: 1.028)
	선택 모델	Random forest
Linear regression – GAM	T 통계량	-3.471
	기각 여부	기각(mean -3.780)
	선택 모델	Linear regression
Linear regression – lasso	T 통계량	1.569
	기각 여부	채택
	선택 모델	
Random forest – GAM	T 통계량	-5.613
	기각 여부	기각(mean -4.808)
	선택 모델	Random forest
Random forest – lasso	T 통계량	-2.041
	기각 여부	채택
	선택 모델	
GAM – lasso	T 통계량	4.453
	기각 여부	기각(mean 4.338)
	선택 모델	lasso

$t_{0.025,9} = 2.262$

<parameter>

- Linear regression
- Random forest:
18번 모델이 best(max_features=5, n_estimators=32, random_state=23, warm_start=True)
- GAM
- Lasso linear regression:
3번 모델이 best(alpha=0.1, random_state=23)

<rmse>

- Linear regression: 3.4248366302639885
- Random forest: 2.3968982908710172
- GAM: 7.204442963268559
- Lasso linear regression: 2.866412504509192

<test RMSE>

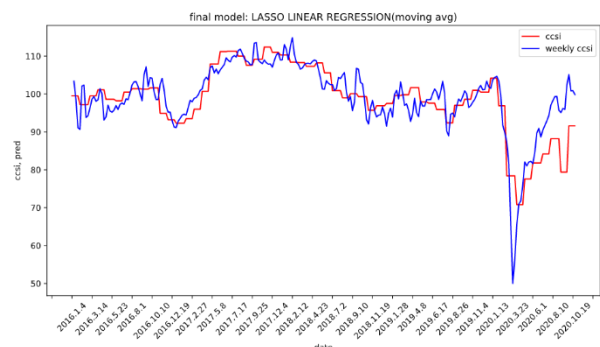
Model1: 13.362 / Model2: 9.257 / Model3: 8.438

<test MAE>

Model1: 3.340 / Model2: 2.678 / Model3: 2.568

➔ Final로 lasso linear regression 선택

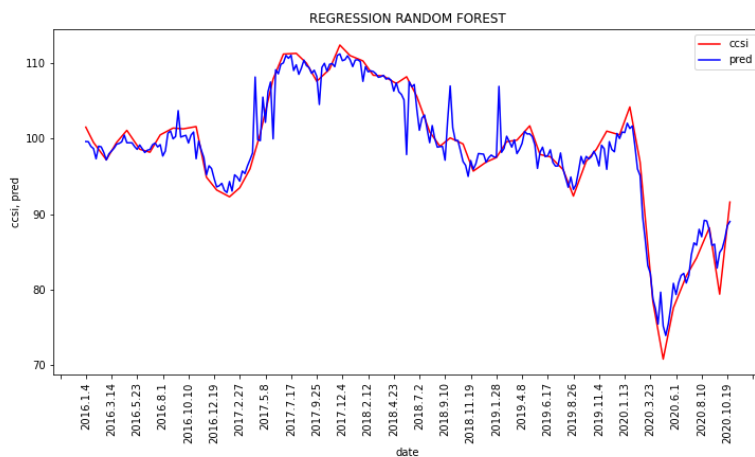
<weekly 단기지표 값에 대한 RMSE> 5.392



< interpolation – shuffle >

		Shuffle = True
Linear regression – random forest	T 통계량	2.145240642741725
	기각 여부	X
	선택 모델	
Linear regression – GAM	T 통계량	2.081938280390646
	기각 여부	X
	선택 모델	
Linear regression – lasso	T 통계량	-0.10784338571792097
	기각 여부	X
	선택 모델	
Random forest – GAM	T 통계량	-1.4891013775742883
	기각 여부	X
	선택 모델	
Random forest – lasso	T 통계량	-6.645985107000552
	기각 여부	O
	선택 모델	random forest model
lasso - GAM	T 통계량	0.902916764264144
	기각 여부	X
	선택 모델	

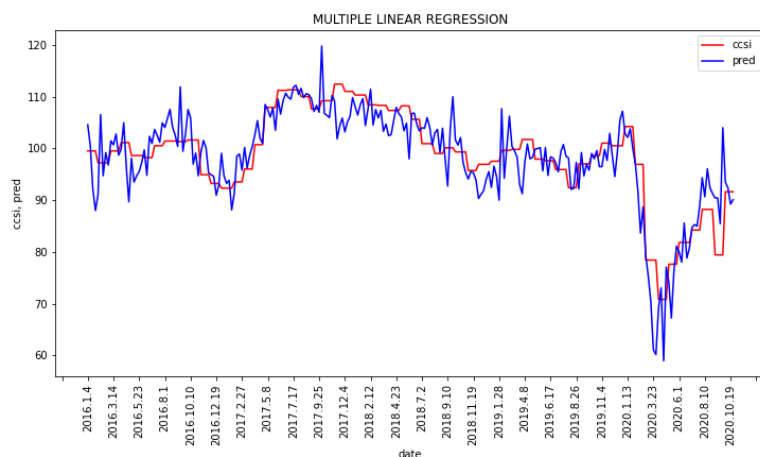
모델 선택 : random forest model



< mean – shuffle >

mean		Shuffle = True
Linear regression – random forest	T 통계량	-0.04883236135535948
	기각 여부	X
	선택 모델	
Linear regression – GAM	T 통계량	-2.874719765496354
	기각 여부	O
	선택 모델	linear regression model
Linear regression – lasso	T 통계량	-0.1068736629495559
	기각 여부	X
	선택 모델	
Random forest – GAM	T 통계량	-3.9092603738698886
	기각 여부	O
	선택 모델	Random Forest model
Random forest – lasso	T 통계량	-0.07449630512451616
	기각 여부	X
	선택 모델	
GAM – lasso	T 통계량	3.827243983117924
	기각 여부	O
	선택 모델	Lasso linear regression model

모델 선택 : Linear regression



<median-shuffle>

median		Shuffle = True
Linear regression – random forest	T 통계량	-0.22801357538794378
	기각 여부	X
	선택 모델	
Linear regression – GAM	T 통계량	-3.1784480286362053
	기각 여부	O
	선택 모델	linear regression model
Linear regression – lasso	T 통계량	-0.26884550425778614
	기각 여부	X
	선택 모델	
Random forest – GAM	T 통계량	-4.063632381568295
	기각 여부	O
	선택 모델	Random Forest model
Random forest – lasso	T 통계량	0.07082139434233481
	기각 여부	X
	선택 모델	
GAM – lasso	T 통계량	3.9482284686579256
	기각 여부	O
	선택 모델	Lasso linear regression model

모델 선택 : Linear regression

