

2019년 KEB 하나은행 인턴 개인과제 발표

통계적 모델링을 통한 금리 예측 및 수익률 곡선 Fitting

: OLS, VIF, t-SNE, ARIMA, LSTM, Nelson-Siegel Svensson

2019.11

트레이딩부 인턴 고은환

01 주제 선정 이유

02 데이터 선택

03 데이터 선별 및 검정

04 모델 선택 및 모델링

05 모델 Setting

06 모델링 결과

07 실제 데이터와 예측치 비교

08 Yield Curve Fitting

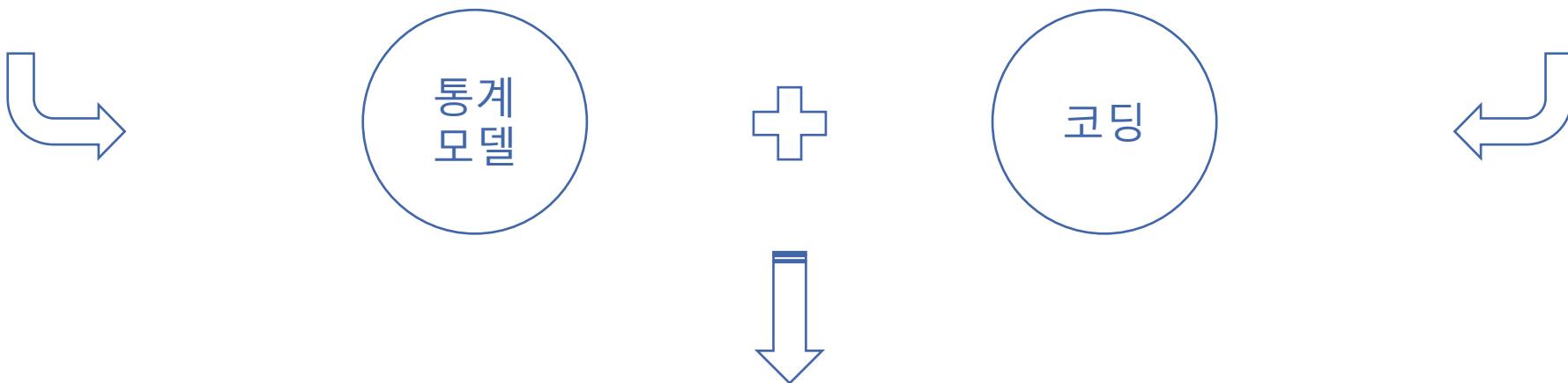
09 마무리

금리에 대한 친근함

1. 은행 예금
2. 기회비용의 기준
3. 대학원 수업
(채권 분석, 이자율 기간구조, Fixed Income Analysis 등)

금리가 갖는 중요성

1. 돈의 가치 척도
2. 국가의 경제 상황을 파악할 수 있는 기준
3. 통화정책의 수단
4. 시장에 미치는 영향



통계적 모델링을 통한 금리 예측 및 수익률 곡선 Fitting

독립 변수

후보: 금리에 영향을 미치는 많은 변수들
 → Available 14 Categories (82개)

유가 및 금 선물
• WTI
• BRENT
• DUBAI
• GOLD

국채금리
• 미국채 1년 금리
• 미국채 5년 금리
• 미국채 10년 금리
• 일본채 1년 금리
• 일본채 5년 금리
• 일본채 10년 금리

실업률
• 4주 기준 실업률

환율
• USD/KRW
• 달러인덱스

국민소득
• 국내총생산GDP
• 국민총소득GNI
• 국내총생산 실질 성장률
• 광공업증가율
• 건설투자증가율
• 설비투자증가율
• 민간증가율
• 정부증가율
• 총수출증가율
• 총수입증가율
• 수출입의 대 GNI 비율

경기산업지수
• 선행종합지수
• 광공업생산지수
• 제조업업황BSI
• 취업자수
• 서비스업생산지수
• 도소매업제외 비율

기준금리
• 한국 기준금리
• 미국 기준금리
• 유럽 기준금리
• 영국 기준금리
• 중국 기준금리
• 일본 기준금리
• 호주 기준금리

국채선물
• 국고채 시장가격지수 _종가
• 국고채 시장가격지수 _12개월 수익률
• 국고채 시장가격지수 _2년 수익률
• 국고채 시장가격지수 _3년 수익률
• 국고채 시장가격지수 _5년 수익률
• 국고채 시장가격지수 _7년 수익률
• 국고채 시장가격지수 _10년수익률
• 국고채 시장가격지수 _듀레이션
• 국고채 시장가격지수 _Convexity
• 국고채 시장가격지수 _YTM

물가
• 물가총지수
• 생산자물가총지수
• 수출물가 총지수
• 수입물가 총지수

한국금리
• 콜금리
• CD91
• 통안증권 91일
• 통화안정 364일
• 1,3,5,10,20년 국채 금리

채권
• 채권거래량_개인
• 채권거래량_외국인
• 채권거래량_기관
• 채권거래량_증권선물
• 채권거래량_합계
• 채권거래대금_개인
• 채권거래대금_외국인
• 채권거래대금_기관
• 채권거래대금_증권선물
• 채권거래대금_합계

국채5년선물매수
• 국채5년선물매수 _개인
• 국채5년선물매수 _외국인
• 국채5년선물매수 _기관
• 국채5년선물매수 _듀레이션
• 국채5년선물매수 _증권선물
• 국채5년선물매수 _합계

증시
• KOSPI
• NASDAQ
• S&P500
• DOW
• NIKKEI225
• SHANGHAI A
• EUROSTOXX50

경제성장률
• 한국 경제 성장률
• 미국 경제 성장률
• 일본 경제 성장률
• 중국 경제 성장률
• 영국 경제 성장률

종속 변수

만기별 국채 금리
 → 1년, 3년, 5년, 10년, 20년

1차 데이터 선별 (82 → 43개)

1. 부분 집합 데이터 제거
2. 데이터의 기간: 2006년 2월 ~ 2019년 11월
→ 수(날짜)가 부족한 데이터 제거

2차 데이터 선별

* 다중공선성(Multicollinearity) 문제

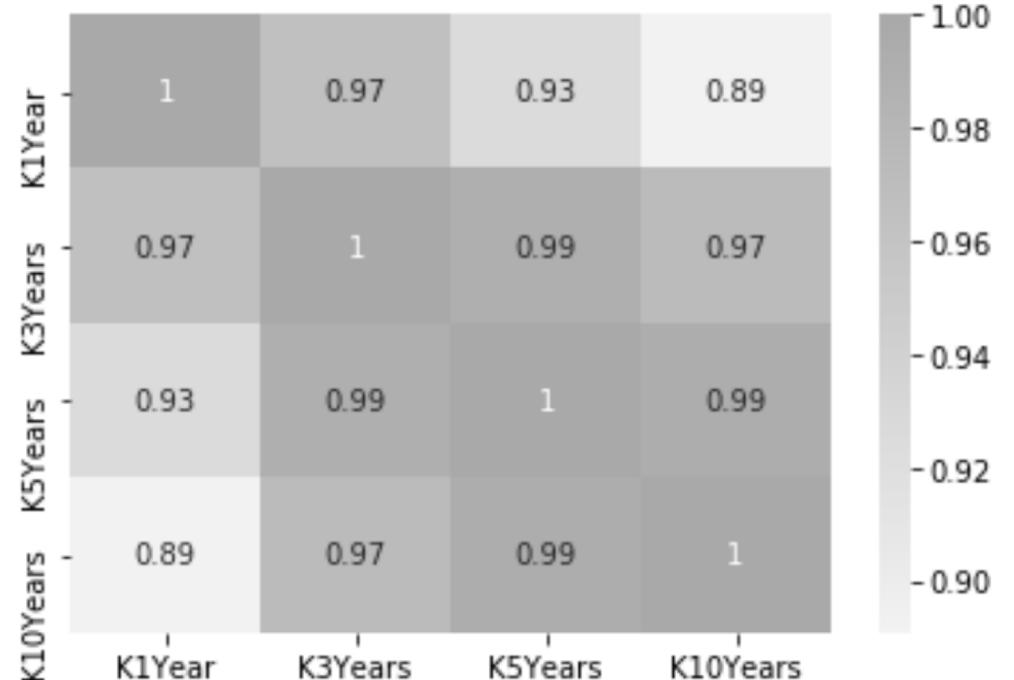
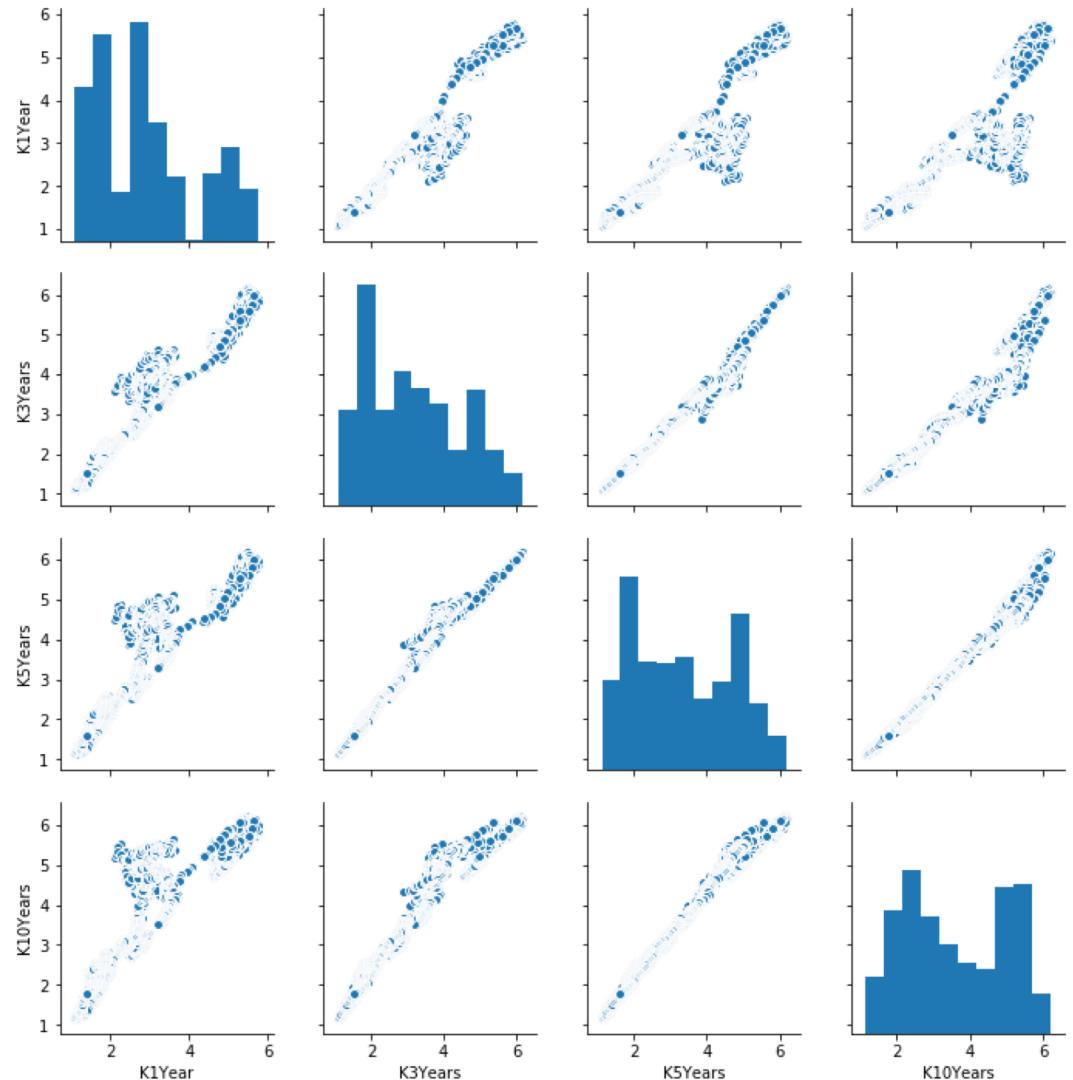
- 독립 변수들간에 서로 관계가 있어 데이터가 꼬이는 문제 (분석 결과 신뢰 불가)
- 독립변수는 서로 상관관계가 있어서는 안되고, 종속변수와 상관관계가 있어야 함

다중공선성 파악 방법

* 다중공선성(Multicollinearity) 문제 해결

- **해결책**
: 사전에 변수들 사이에 상관성이 있는지 통계분석 시도
(가급적 서로 관계가 낮은 변수들을 사용하면 이 문제를 피할 수 있음)
- **Scatter plot Matrix (산점도 그래프)**
→ 산점도 그래프를 통해 독립변수끼리 상관관계가 있는지 파악하는 방법
- **OLS Regression (최소자승법)**
→ 다중공선성이 있으면 독립변수의 공분산행렬의 조건수 (Conditional Number)가 증가
→ p-value가 0.5를 넘을 경우 다중공선성을 의심
- **VIF(Variance Inflation Factors, 분산팽창요인)**
→ 다중 회귀 모델에서 독립 변수간 상관관계가 있는지 측정하는 척도
→ VIF가 10을 넘으면 다중공선성이 있다고 판단, 5를 넘으면 주의 필요

Scatter Plot Matrix



[만기별 국채금리간 상관관계 - Correlation Table]

OLS Regression Results

Dep. Variable:	K1Year	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	4.079e+04
Date:	Thu, 21 Nov 2019	Prob (F-statistic):	0.00
Time:	01:06:31	Log-Likelihood:	5171.2
No. Observations:	3604	AIC:	-1.025e+04
Df Residuals:	3560	BIC:	-9982.
Df Model:	43		
Covariance Type:	nonrobust		
Omnibus:	327.314	Durbin-Watson:	0.093
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1757.700
Skew:	-0.254	Prob(JB):	0.00
Kurtosis:	6.383	Cond. No.	2.69e+08

Warnings:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.69e+08. This might indicate that there are strong multicollinearity or other numerical problems.

	coef	std err	t	P> t	[0.025	0.975]	제조업업황BSI	-0.0004	0.000	-1.224	0.221	-0.001	0.000
intercept	-0.1748	0.164	-1.068	0.286	-0.496	0.146	서비스업생산지수_도소매업제외	0.0208	0.003	6.777	0.000	0.015	0.027
WTI	-0.0004	0.000	-2.183	0.029	-0.001	-3.69e-05	국내총생산GDP	-1.989e-08	6.28e-08	-0.316	0.752	-1.43e-07	1.03e-07
GOLD	-4.113e-05	1.94e-05	-2.123	0.034	-7.91e-05	-3.15e-06	실질GDP	0.0120	0.001	8.218	0.000	0.009	0.015
USDKRW	0.0002	3.75e-05	4.349	0.000	8.97e-05	0.000	한국경제성장률	0.0013	0.002	0.597	0.550	-0.003	0.006
미국채1년	0.0164	0.010	1.646	0.100	-0.003	0.036	미국경제성장률	0.0040	0.001	4.931	0.000	0.002	0.006
미국채5년	0.0054	0.013	0.402	0.687	-0.021	0.032	일본경제성장률	0.0026	0.001	2.021	0.043	7.96e-05	0.005
미국채10년	-0.0221	0.012	-1.916	0.055	-0.045	0.001	중국경제성장률	-0.0033	0.001	-3.512	0.000	-0.005	-0.001
미국기준금리	0.0873	0.008	11.361	0.000	0.072	0.102	영국경제성장률	0.0227	0.003	6.526	0.000	0.016	0.030
유로기준금리	-0.0199	0.007	-2.765	0.006	-0.034	-0.006	국고채시장가격지수듀레이션	-0.0107	0.019	-0.576	0.565	-0.047	0.026
한국기준금리	-0.0843	0.021	-4.102	0.000	-0.125	-0.044	국고채시장가격지수Convexity	-0.0142	0.004	-3.592	0.000	-0.022	-0.006
중국기준금리	0.0125	0.008	1.489	0.137	-0.004	0.029	국고채시장가격지수YTM	0.8073	0.030	26.760	0.000	0.748	0.866
일본기준금리	-0.0306	0.032	-0.957	0.338	-0.093	0.032	KOSPI	0.0001	1.56e-05	8.300	0.000	9.87e-05	0.000
경상수지	2.641e-06	5.44e-07	4.850	0.000	1.57e-06	3.71e-06	NASDAQ	1.946e-05	1.17e-05	1.666	0.096	-3.44e-06	4.24e-05
콜금리	0.0786	0.019	4.126	0.000	0.041	0.116	SnP500	-0.0001	6.2e-05	-1.882	0.060	-0.000	4.86e-06
CD91	0.3447	0.010	33.433	0.000	0.324	0.365	DOW	-4.065e-06	4.74e-06	-0.858	0.391	-1.34e-05	5.22e-06
국고채3년	0.3847	0.027	14.500	0.000	0.333	0.437	NIKKEI225	6.428e-06	1.84e-06	3.493	0.000	2.82e-06	1e-05
국고채5년	-0.7646	0.035	-21.909	0.000	-0.833	-0.696	SHANGHAI	-3.231e-05	3.29e-06	-9.823	0.000	-3.88e-05	-2.59e-05
국고채10년	0.1682	0.038	4.481	0.000	0.095	0.242	EUROSTOXX50	2.606e-05	1.22e-05	2.134	0.033	2.12e-06	5e-05
국고채20년	-0.0314	0.025	-1.245	0.213	-0.081	0.018							
채권외국인거래량	1.049e-08	2.4e-08	0.436	0.663	-3.67e-08	5.76e-08							
채권외국인거래대금	-1.703e-08	2.38e-08	-0.716	0.474	-6.37e-08	2.96e-08							
물가총지수	-0.0069	0.003	-2.590	0.010	-0.012	-0.002							
실업률	0.0038	0.003	1.223	0.222	-0.002	0.010							
선행증합지수	-0.0267	0.003	-9.490	0.000	-0.032	-0.021							
광공업생산지수	0.0142	0.001	15.434	0.000	0.012	0.016							

VIF(Variance Inflation Factors)

1. 다중공선성 문제가 있다는 것이 확인됨
2. VIF가 10이 넘는 변수 제거
3. 최종 데이터 선별 (43 → 11개)

VIF Formula

✓ def. i 번째 독립변수를 다른 독립변수로 선형 회귀한 성능(결정 계수)을 나타낸 것

✓ i 번째 변수의 VIF

$$VIF_i = \frac{\sigma^2}{(n-1)Var[X_i]} \frac{1}{1-R_i^2}$$

✓ 다른 변수에 의존적일 수록 VIF가 커짐

K1Year

VIF Factor	features
0 26343.534252	Intercept
1 14.555174	WTI
2 33.208846	GOLD
3 14.253736	USDKRW
4 249.191546	미국채1년
5 259.510635	미국채5년
6 134.226022	미국채10년
7 173.586436	미국기준금리
8 100.423362	유로기준금리
9 641.907391	한국기준금리
10 68.183775	중국기준금리
11 35.179972	일본기준금리
12 4.659095	경상수지
13 562.569715	콜금리
14 195.549447	CD91
15 1314.837567	국고채3년
16 2356.577348	국고채5년
17 2807.562377	국고채10년
18 1299.326478	국고채20년
19 90.233244	채권외국인거래량
20 90.479554	채권외국인거래대금
21 465.650706	물가총지수
22 2.391419	실업률
23 1363.343818	선행종합지수
24 123.058481	광공업생산지수
25 8.644138	제조업업황BSI
26 1014.318192	서비스업생산지수_도소매업체의
27 339.441489	국내총생산GDP
28 4.878180	국내총생산실질성장률
29 3.770622	한국경제성장률
30 3.779193	미국경제성장률
31 2.265474	일본경제성장률
32 15.845422	중국경제성장률
33 4.861107	영국경제성장률
34 3.966667	국고채시장가격지수듀레이션
35 3.991627	국고채시장가격지수Convexity
36 1710.139537	국고채시장가격지수YTM
37 25.823566	KOSPI
38 973.379369	SnP500
39 654.273769	DOW
40 76.685105	NIKKEI225
41 7.735284	SHANGHAI
42 48.556947	EUROSTOXX50

K5Years

VIF Factor	features
0 26339.446810	Intercept
1 14.535631	WTI
2 31.706616	GOLD
3 14.109692	USDKRW
4 247.211238	미국채1년
5 253.924183	미국채5년
6 131.957096	미국채10년
7 173.343217	미국기준금리
8 100.344960	유로기준금리
9 625.544310	한국기준금리
10 67.931328	중국기준금리
11 35.179273	일본기준금리
12 4.629775	경상수지
13 558.656207	콜금리
14 155.560937	CD91
15 1138.689075	국고채3년
16 2102.928336	국고채10년
17 1296.731072	국고채20년
18 90.203024	채권외국인거래량
19 90.444538	채권외국인거래대금
20 454.356714	물가총지수
21 2.391290	실업률
22 1353.233677	선행종합지수
23 119.961192	광공업생산지수
24 8.643812	제조업업황BSI
25 1000.753745	서비스업생산지수_도소매업체의
26 335.944452	국내총생산GDP
27 4.728197	국내총생산실질성장률
28 3.204784	한국경제성장률
29 3.721497	미국경제성장률
30 2.078645	일본경제성장률
31 14.495006	중국경제성장률
32 4.836735	영국경제성장률
33 3.918210	국고채시장가격지수듀레이션
34 3.958421	국고채시장가격지수Convexity
35 1516.667091	국고채시장가격지수YTM
36 25.646171	KOSPI
37 973.377822	SnP500
38 654.246815	DOW
39 75.761398	NIKKEI225
40 7.730196	SHANGHAI
41 48.541692	EUROSTOXX50

K10Years

VIF Factor	features
0 26343.358950	Intercept
1 14.529867	WTI
2 33.208844	GOLD
3 13.273500	USDKRW
4 249.080661	미국채1년
5 259.122798	미국채5년
6 133.156530	미국채10년
7 171.368251	미국기준금리
8 100.063415	유로기준금리
9 635.628069	한국기준금리
10 68.130073	중국기준금리
11 33.954447	일본기준금리
12 4.649903	경상수지
13 562.548209	콜금리
14 190.087347	CD91
15 1287.485379	국고채3년
16 1765.130251	국고채5년
17 411.514798	국고채20년
18 90.231621	채권외국인거래량
19 90.476600	채권외국인거래대금
20 463.891534	물가총지수
21 2.376736	실업률
22 1335.375263	선행종합지수
23 123.054583	광공업생산지수
24 8.591809	제조업업황BSI
25 1008.900782	서비스업생산지수_도소매업체의
26 335.589281	국내총생산GDP
27 4.860019	국내총생산실질성장률
28 3.713981	한국경제성장률
29 3.718441	미국경제성장률
30 2.263005	일본경제성장률
31 15.127474	중국경제성장률
32 4.858592	영국경제성장률
33 3.962313	국고채시장가격지수듀레이션
34 3.974304	국고채시장가격지수Convexity
35 1688.188675	국고채시장가격지수YTM
36 25.170701	KOSPI
37 971.735385	SnP500
38 654.017378	DOW
39 74.942754	NIKKEI225
40 7.731611	SHANGHAI
41 48.405652	EUROSTOXX50

최종 선별된 독립변수

- 만기별로 국채금리를 종속변수로 놓고 VIF를 돌린 결과, 공통적으로 VIF가 10 미만인 독립변수들이 필터링됨

- 경상수지
- 실업률
- 제조업 업황 BSI
- 실질GDP
- 한국 경제 성장률
- 미국 경제 성장률
- 일본 경제 성장률
- 영국 경제 성장률
- 국고채 시장가격지수 듀레이션
- 국고채 시장가격지수 Convexity
- SHANGHAI A

종속 변수

- 만기별 국채금리
- 1년물 국채금리
 - 3년물 국채금리
 - 5년물 국채금리
 - 10년물 국채금리
 - 20년물 국채금리

그래도 여전히 고차원 데이터

- 차원의 저주 (Curse of Dimensionality)

: 일반적으로 관측 값의 개수가 n , 특징의 개수가 p 일 때,
 $p >> n$ 인 데이터를 고차원 데이터라 부르며, 차원의 저주는
 고차원 데이터 분석에 대한 어려움을 의미

- 특징(feature)이 많으면 모델의 성능이 감소하거나 과적합(overfitting)을 일으켜, 해당 모델을 해석하여 용이한 정보를 얻기가 힘듦

$\mathbb{R}^{\text{多}} \rightarrow \mathbb{R}^{\text{少}}$

차원을 줄이는 방법

- 특징 추출 (Feature Extraction)

: 주어진 특징들을 조합하여 새로운 특징 값을 계산하는 작업
 Ex) LASSO (L1 Regularization)

- 특징 선택 (Feature Selection)

: 전문가 지식(knowledge)이나 데이터 밖의 데이터 (metadata)를 이용하여 일부를 골라내는 작업
 Ex) PCA(주성분분석), t-SNE

- 그 외 머신러닝의 autoencoder, 음수미포함분해의 nnf, 위상수학의 tda 등

t-SNE

- t-SNE (Stochastic Neighbor Embedding)

✓ 비선형 차원 축소 기법

✓ 고차원의 원공간에 존재하는 데이터 x 의 이웃 간 거리를 최대한 보존하는 저차원의 y 를 학습하는 방법론
(stochastic이란 이름이 붙은 이유는 거리 정보를 확률적으로 나타내기 때문)

$$p_{j|i} = \frac{e^{-\frac{|x_i - x_j|^2}{2\sigma_i^2}}}{\sum_k e^{-\frac{|x_i - x_k|^2}{2\sigma_i^2}}}$$

$$q_{j|i} = \frac{e^{-|y_i - y_j|^2}}{\sum_k e^{-|y_i - y_k|^2}}$$

t-SNE 식과 목적

- 첫 번째 식의 p 는 고차원 원공간에 존재하는 i 번째 개체 x_i 가 주어졌을 때 j 번째 이웃인 x_j 가 선택될 확률
- 두 번째 식의 q 는 저차원에 임베딩된 i 번째 개체 y_i 가 주어졌을 때 j 번째 이웃인 y_j 가 선택될 확률
- t-SNE의 목적
 - ✓ 차원축소가 제대로 잘 이뤄졌다면 고차원 공간에서 이웃으로 뽑힐 확률과 저차원 공간에서 선택될 확률이 비슷할 것이므로, p 와 q 의 분포 차이가 최대한 작게끔 하고자 함

∴

- ✓ 고차원 데이터가 가지고 있는 문제를 저차원 데이터로 축소하여 해결 (보통 2차원)
- ✓ 축소가 잘 되었다면 고차원 데이터의 특징은 저차원에서도 유지될 것 → 모델링에 필요한 데이터의 특징을 저차원 데이터를 통해서도 활용할 수 있음
- ✓ 이를 위해 사용되는 것이 t-SNE

모델링에 필요한 데이터 Setting 프로세스

1. Input Data

- ✓ OLS 및 VIF를 통과한 **11개**의 데이터

2. 전체 기간 중 빈 값의 데이터는 전일 데이터로 채움

3. 10년 이상의 데이터이므로 Scaling 필요

- ✓ Standard Scaling, Min-Max Scaling, Log Scaling 등
- ✓ 11개의 데이터 중 **상해A 지수만** Scaling 필요

4. t-SNE 적용

5. 데이터 분류 (In-Sample 및 Out-Sample)

- ✓ IS(Train Data) : OS(Test Data) = 7 : 3
- ✓ IS: 2006년 1월 25일 ~ 2015년 11월 5일
- ✓ OS: 2015년 11월 6일 ~ 2019년 11월 18일

6. 모델에 들어가는 데이터의 단위

- ✓ **window_size = 60** (business day 기준)
- ✓ $t \sim t + 60 \rightarrow t + 61$

7. Overfitting 방지 메커니즘 설정 (for LSTM)

- ✓ Model selection (k-fold Cross Validation)
- ✓ L1 Regularization (Lasso)
- ✓ Early stopping
- ✓ Dropout

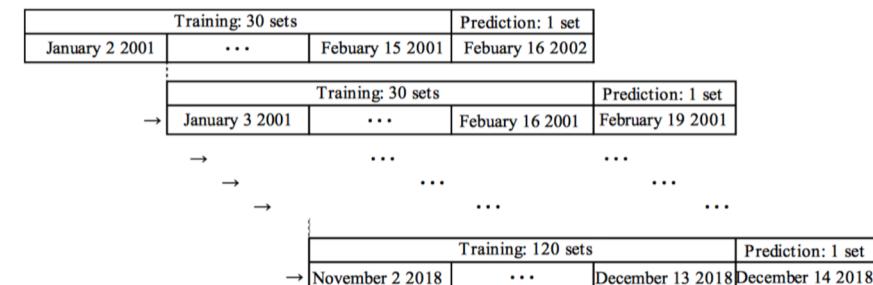


Figure. Our Model Training-prediction set Devided Sliced with Window Size 30

L1 Regularization (LASSO)

- ✓ 여러 Data feature 및 Data transformation을 통해 차원이 높아진 차수들의 계수 Weight(혹은 Theta)를 줄이는 방법
 - ✓ Loss function을 최소화함과 동시에 \mathbf{W} 도 줄이는 해결책
- $$J(\theta) = \frac{1}{2m} \sum_{i=1}^m \left(h_{\theta}(x^{(i)}) - y^{(i)} \right)^2 + \frac{\lambda}{2} \sum_{j=1}^n |\theta_j|$$
- ✓ \mathbf{W} 가 계속해서 업데이트 될 때, \mathbf{W} 의 원소인 어떤 w_i 는 0 이 되도록 한다. (계속해서 특정 상수를 빼기 때문)
 - ✓ 따라서, L1은 영향을 크게 미치는 핵심적인 feature \mathbf{f}_t 들만 반영하도록 한다.

Model Selection (k-fold Cross Validation)

- ✓ 새로운 Data set에 대해 반응하는 모델의 성능을 추정하는 방법
 - 특정 데이터를 Training set과 Test set으로 분할 (7 : 3)
 - Training set을 다시 k개의 그룹(fold)로 나눔
 - k개로 나누어진 Training set에 대해 서로 다른 Validation Fold를 지정하며 아래 과정 k번 반복
 - (k - 1)개의 Train Fold에 대해 학습 진행
 - 나머지 1개의 Validation Fold에 대해 성능 측정
 - 위에서 얻은 각 Hyperparameter의 k개의 결과에 대한 평균을 계산하여 이 평균 값을 대표 Hyperparameter로 지정
 - 마지막으로 이 Hyperparameters를 바탕으로 Test Data에 대해 모델 1회 Test

Dropout

- ✓ 복잡한 Neural Network에서 몇몇 Node들의 관계를 끊어버리는 것
- ✓ 전체적으로 보았을 때, 올바르게 학습할 수 있다는 원리
- ✓ Input layer와 Hidden layer에 있는 Node들을 정해진 Dropout rate 만큼 제거
- ✓ 진행되는 학습 차수마다 랜덤으로 Node들에 적용되며 반복

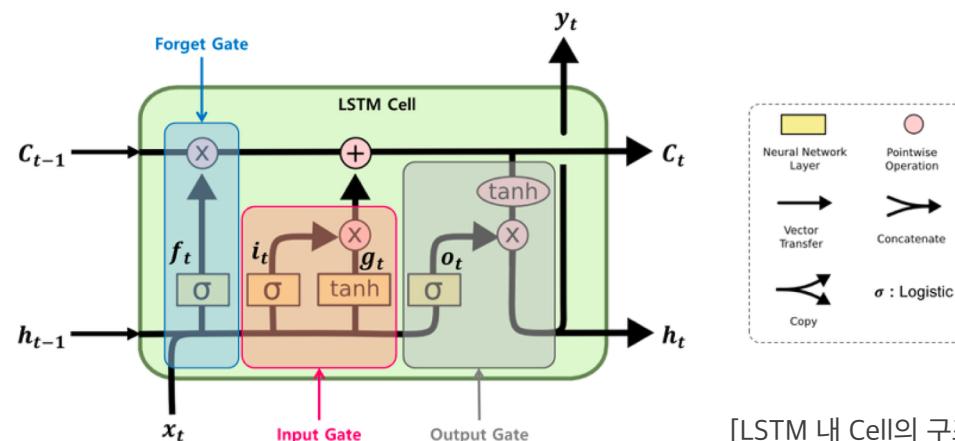
Early Stopping

- ✓ K-fold Cross Validation을 통해 나누어진 3개의 Set 중 Validation set 과정에서 Overfitting이 일어날 때, 정해진 Threshold를 넘어서는 순간 학습을 Stop
- ✓ 그 후, Test set으로 남겨둔 데이터로 Final evaluation 진행

LSTM

- LSTM 활용 목적

- ✓ 시계열은 입력 변수간의 **Sequence 종속성**으로 인해 복잡성이 추가됨 (그래서 시계열 예측 문제는 예측 모델링의 어려운 유형에 속함)
- ✓ 이러한 Sequence 복잡성을 다루기 위해 설계된 강력한 유형의 신경망으로 **순환 신경망(RNN)**이 있음
- ✓ 그러나 시계열의 특성상 전의 정보가 현재의 타임 스텝에 영향을 줄 수 있는데, RNN은 타임 스텝이 길어질 수록 영향이 감소하는 “**장기 의존성 문제 (Long-Term Dependency Problem)**”가 존재
- ✓ 또한 타임 스텝이 길어질 수록 weight가 업데이트 되지 않는 “**그래디언트 소실 문제 (Vanishing Gradient Problem)**” 역시 존재
- ✓ LSTM은 이러한 문제를 해결하고, 매우 큰 아키텍처를 훈련할 수 있기 때문에 자주 사용되는 일종의 반복 신경 네트워크 (저장하고 버릴 정보를 선별함으로써 문제 해결)



다양한 하이퍼 파라미터의 조합

- # of Memory Cell: [32, 64, 128, 256, 512]
- Learning Rate: [0.1, 0.01, 0.001]
- Optimizer: [Adam, Sgd]
- Activation: [tanh, relu, sigmoid]

고정 하이퍼 파라미터

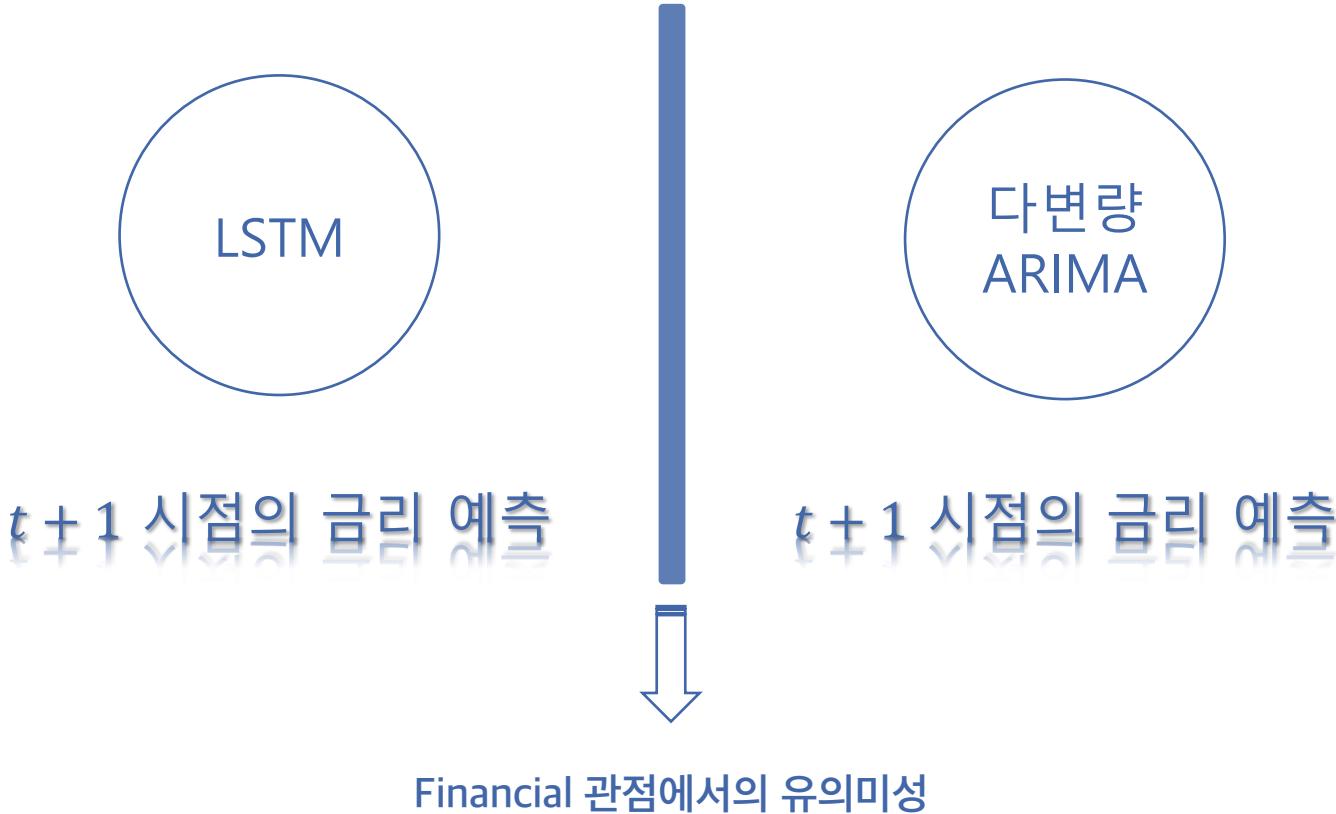
- Epoch: 100
- Loss Function: MSE
- Decay: 0.01
- Dropout: 0.7
- L2 Regularization: 0.001
- Patience of ES: 3

하이퍼 파라미터의 조합 결과 Best ~ Worst

HyperParameter	Train Loss	Valid Loss	Test Loss
128 0.01 Adam tanh	0.001852056	0.001638957	0.000240991
512 0.1 Adam relu	0.001706595	0.00140357	0.000355352
128 0.001 Adam tanh	0.001469253	0.001569606	0.000376253
256 0.1 Adam relu	0.001729884	0.001194295	0.0004574
128 0.1 Adam tanh	0.002188466	0.001895588	0.000481399

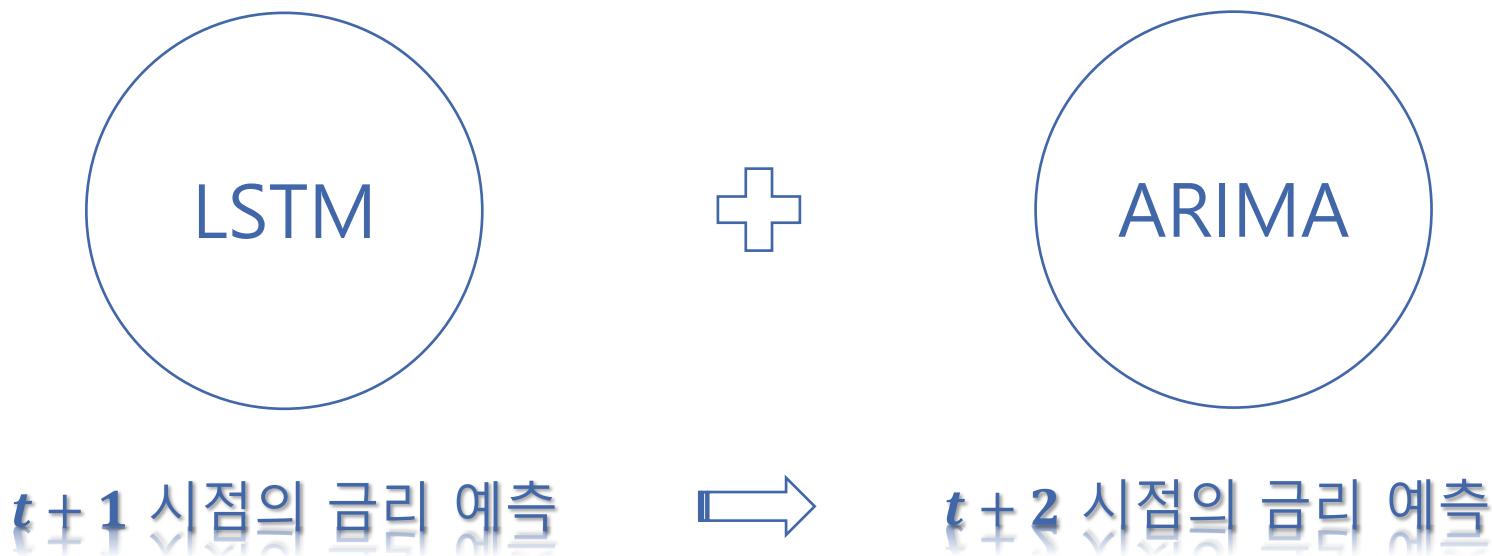
128 0.001 sgd sigmoid	0.054972365	0.067150431	0.075880863
32 0.1 sgd sigmoid	0.0563699	0.066965575	0.076829676
32 0.01 sgd sigmoid	0.059842429	0.070724459	0.079731898
128 0.1 sgd sigmoid	0.05579158	0.069446117	0.080109261
64 0.01 sgd sigmoid	0.057040283	0.070609355	0.082474399

LSTM과 ARIMA의 비교



- LSTM 모델 결과가 기존의 다른 모델들에 의해 설명이 된다면, 그 모델은 **기존 모델의 핵심을 잡아내는데 유효한 모델**이 되는 것
 - 결과가 기존의 모델로 설명이 안된다면 (그리고 결과가 지속적으로 OS에서 유의미하게 나온다면)
그 모델은 **시장의 Anomaly를 잡아낸 새로운 모델**이 될 것

새로운 시도 = LSTM과 ARIMA의 결합



LSTM을 통해 나온 OS기간의 예측 금리를 ARIMA 모델의 Input Data로 사용

ARIMA(p, d, q)

- ARIMA 정의

- ✓ Autoregressive Integrated Moving Average의 약자

- AR(Autoregressive, 자기회귀모형): 이전 관측 값의 오차항이 이후 관측 값에 영향을 주는 모형 (모수: p)
 - I(Integrated, 누적): 차분을 이용하는 시계열모형에 붙이는 표현 (모수: d)
 - MA(Moving Average, 이동평균모형): 이전의 연속적인 오차항이 이후의 관측 값에 영향을 주는 모형 (모수: q)

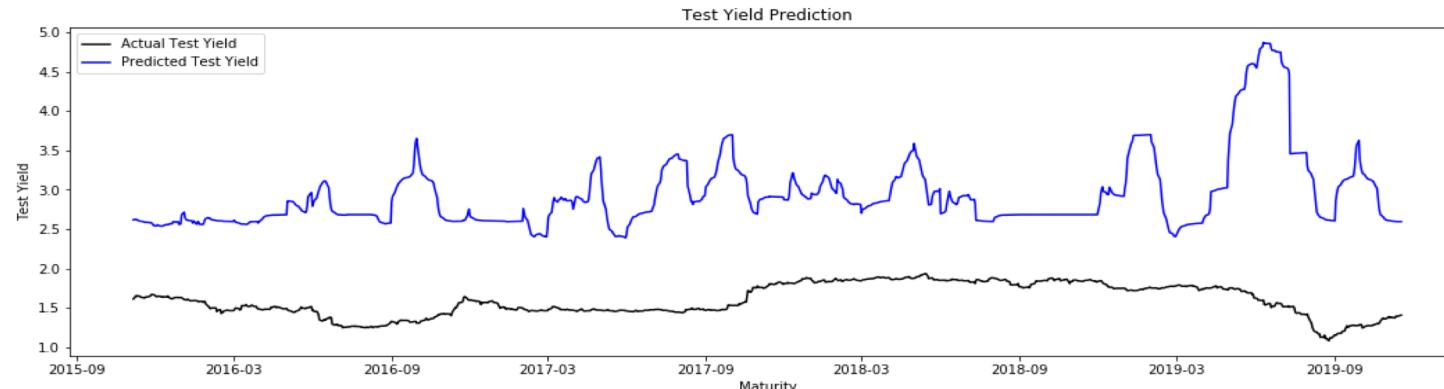
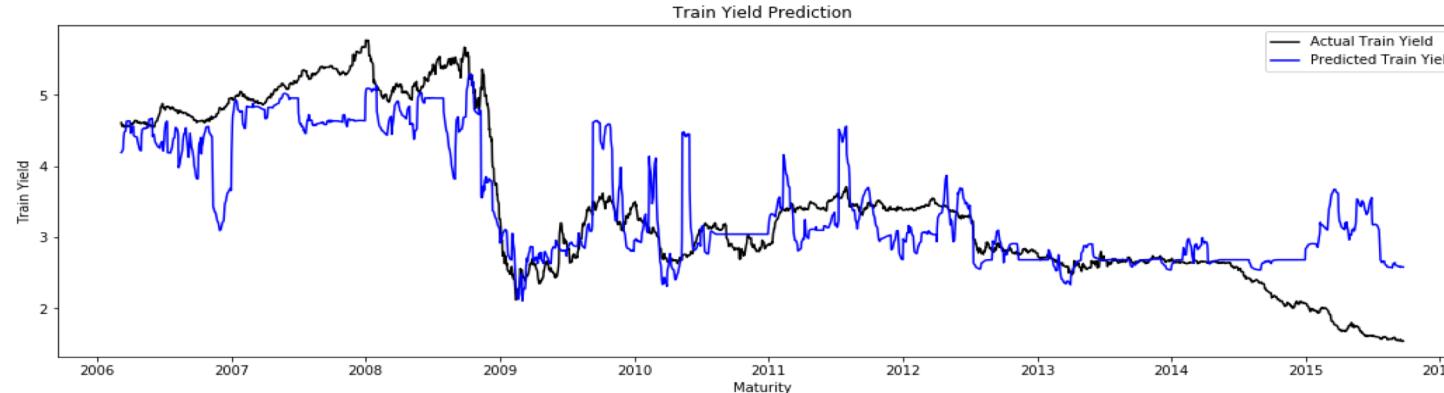
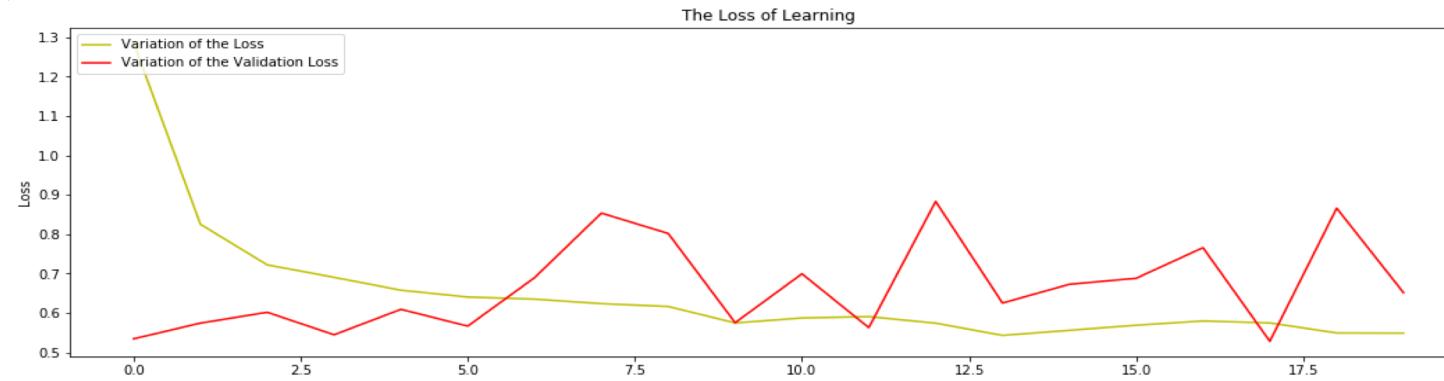
- ✓ 자기회귀와 이동평균을 둘 다 고려하는 모형

- AR, MA, ARMA와 ARIMA의 차이점
→ ARIMA의 경우 AR, MA, ARMA가 설명하지 못하는 **시계열의 비정상성(Non-stationary)**을 설명하기 위해 **관측치간의 차분(Difference)**을 사용
 - $AR(p) = ARIMA(p, 0, 0)$ (where p : Lag of AR, q : Lag of MA, I : the difference count)
 - $MA(q) = ARIMA(0,0,q)$
 - $ARMA(p,q) = ARIMA(p,0,q)$

ARIMA(p, d, q)

- ARIMA의 모수 p, d, q 설정
 - ✓ ACF plot와 PACF plot을 통해 AR 및 MA의 모수를 추정
 - ✓ ACF(Autocorrelation function) : Lag에 따른 관측치들 사이의 관련성을 측정하는 함수
 - ✓ PACF(Partial autocorrelation function) : k 이외의 모든 다른 시점 관측치의 영향력을 배제하고 y_t 와 y_{t-k} 두 관측치의 관련성을 측정하는 함수
 - 시계열 데이터의 특성 → AR → ACF는 천천히 감소하고, PACF는 급격히 감소
 - 시계열 데이터의 특성 → MA → ACF는 급격히 감소하고, PACF는 천천히 감소

	AR(p)	MA(q)
ACF	점차적으로 감소	시차 q 이후에 0
PACF	시차 p 이후에 0	점차적으로 감소



[Figure 1. Loss 값]

노랑: 모델의 IS 데이터 학습 정확도

빨강: 모델의 IS k-fold CV 데이터 학습 정확도

[Figure 2. IS기간 예측치]

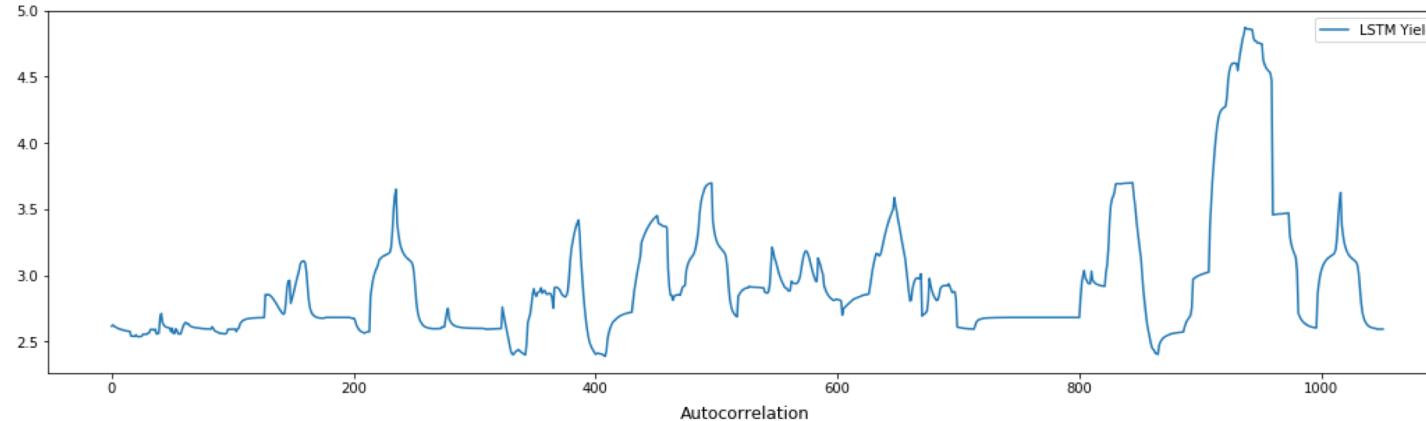
파랑: 모델의 IS기간 국채 1년물 금리 예측치

검정: IS기간 국채 1년물 실제 금리

[Figure 3. OS기간 예측치]

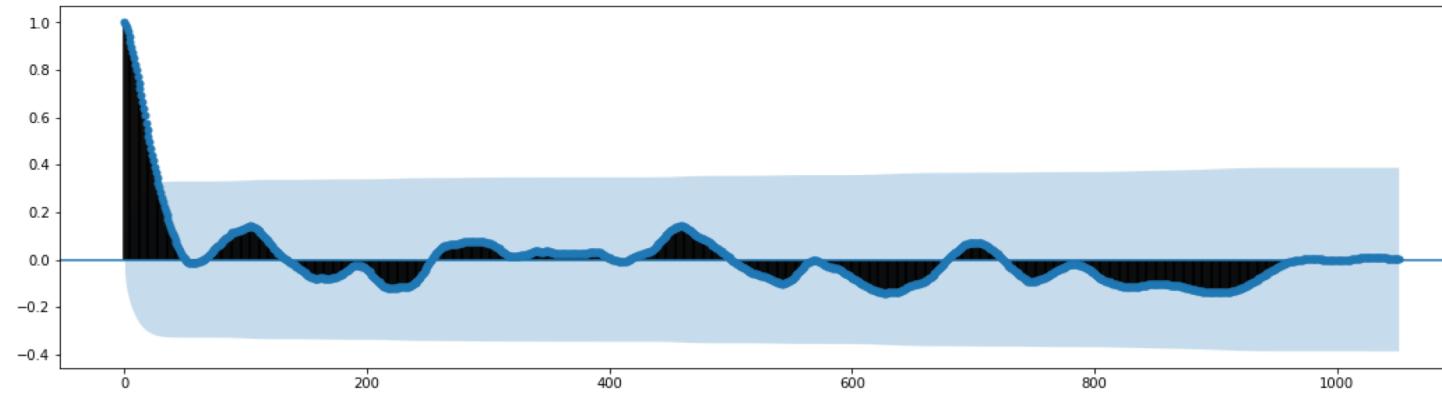
파랑: 모델의 OS기간 국채 1년물 금리 예측치

검정: OS기간 국채 1년물 실제 금리



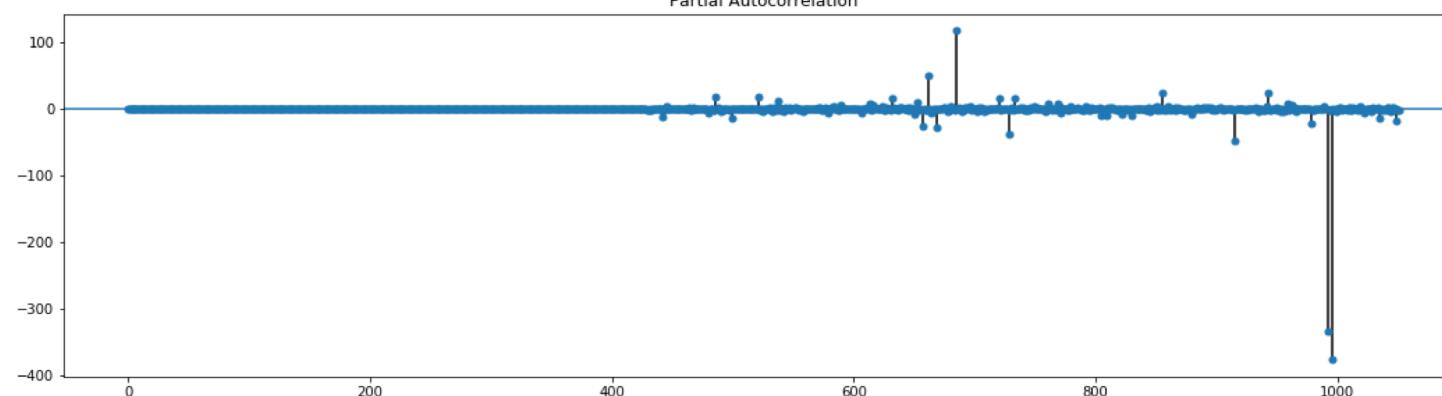
[Figure 4. OS기간 예측치]

파랑: LSTM 모델의 OS기간 국채 1년물 금리 예측치



[Figure 5. LSTM 예측치에 대한 ACF plot]

처음 60일 정도의 Time Lag에서 급격히 감소하여 0을 넘어감

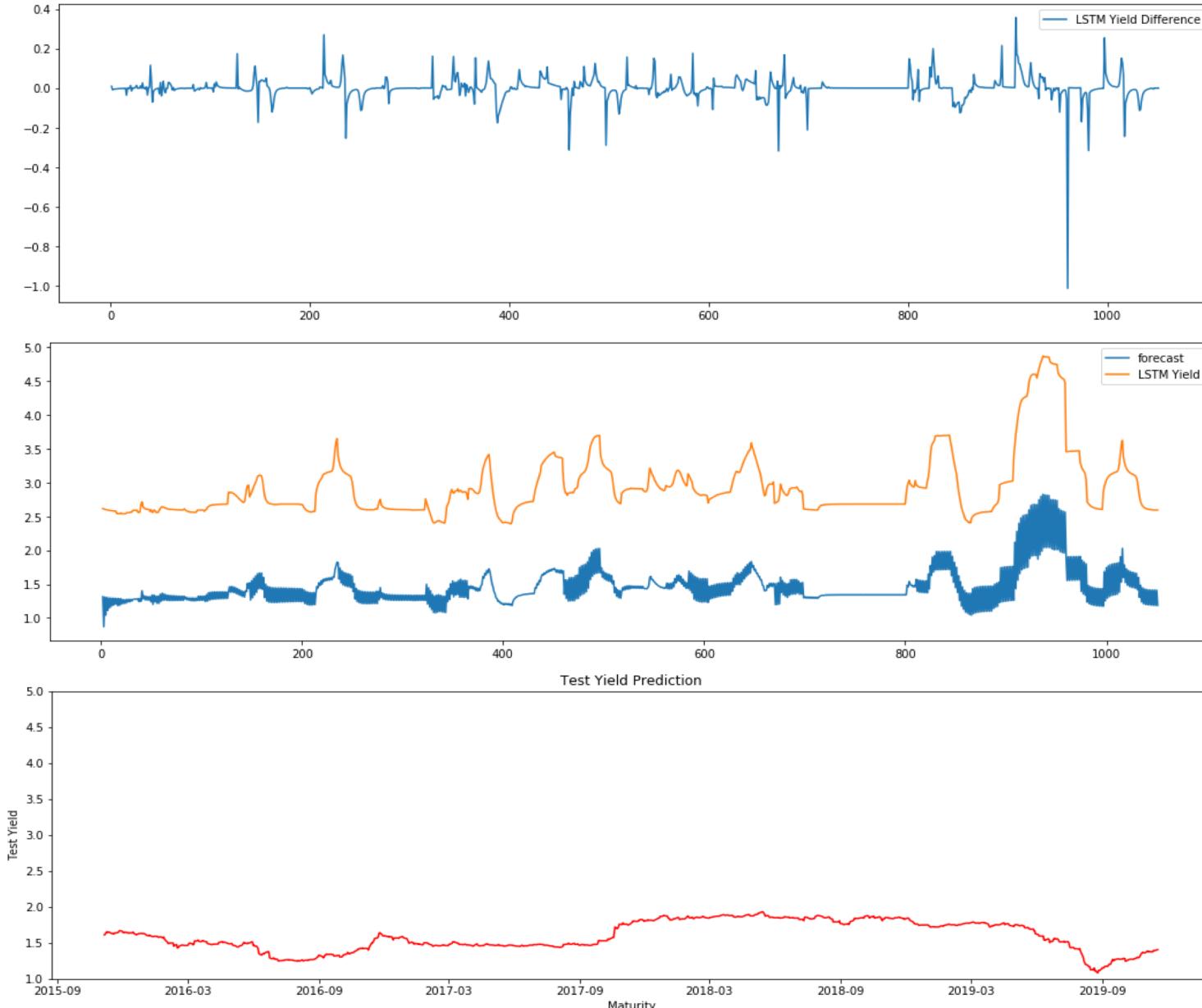


[Figure 6. LSTM 예측치에 대한 PACF plot]

점진적 변화 후 마지막 기간에 급감



MA의 특성 ($p:0, q=1$)



[Figure 7. OS기간 LSTM 예측치 차분]

파랑: OS기간 국채 1년물 금리 예측치 차분값
→ 제외

[Figure 8. ARIMA 모델의 예측치]

- ARIMA 모델은 LSTM의 예측치(y_{t+1})를 Input Data로 사용하여 예측하므로, LSTM에 들어간 Input Data (순수 OS data)의 시점 (t)로부터 2 Time Steps 뒤의 금리 Y_{t+2} (**파랑**)를 예측함
- 순수 OS Data의 마지막 t 는 19년 11월 18일, ARIMA 모델의 마지막 $t + 2$ 는 19년 11월 20일

→ 11월 20일 국채 1년물 금리: 1.380

11월 20일 ARIMA and LSTM의 예측치: 1.411

[Figure 9. OS기간 예측치]

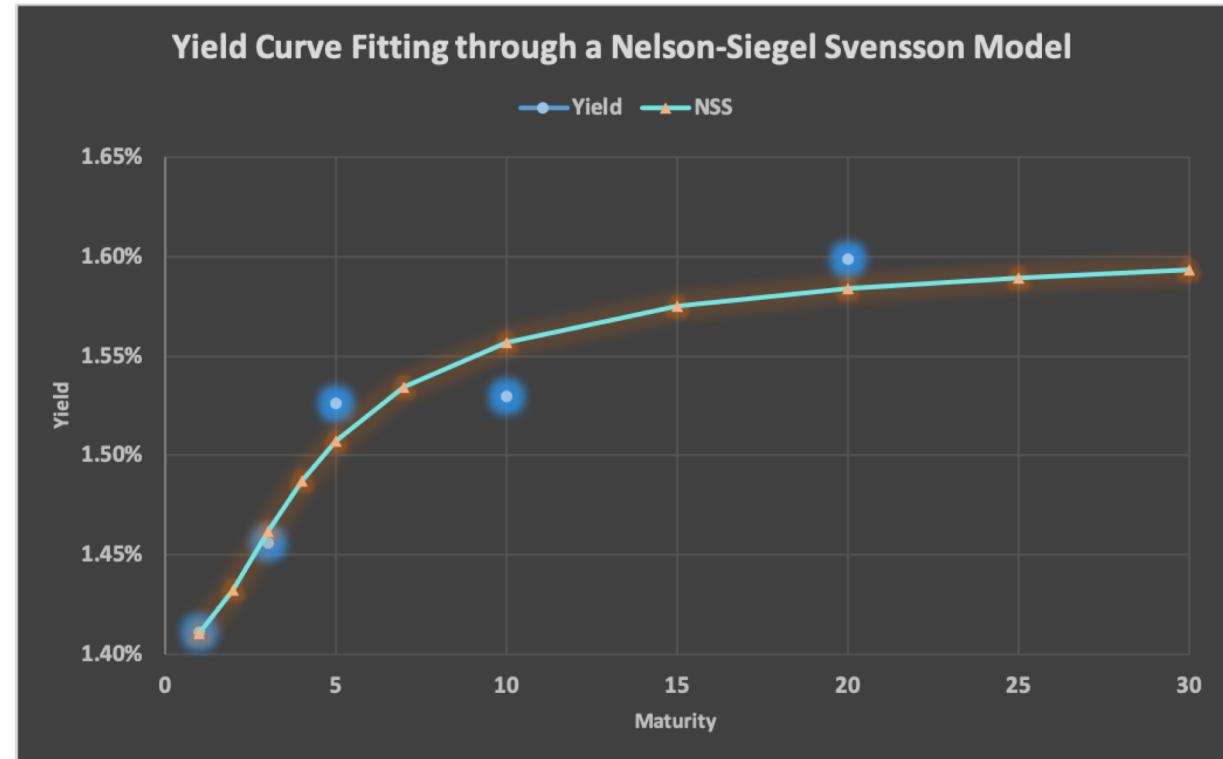
빨강: OS기간 국채 1년물 실제 금리

Maturity	19년 11월 20일의 Real Yield	19년 11월 20일의 Forecast Yield
1Year	1.380	1.411
3Years	1.450	1.456
5Years	1.519	1.526
10Years	1.671	1.530
20Years	1.643	1.599

Yield Curve Fitting

- 11월 18일에 Fitting 한 11월 20일의 Yield Curve
- 4 Factors Nelson-Siegel Svensson (NSS) Model
- NSS Yield =

$$\beta_1 + \beta_2 \frac{1-e^{-\frac{Y_1}{\lambda_1}}}{\frac{Y_1}{\lambda_1}} + \beta_3 \left(\frac{1-e^{-\frac{Y_1}{\lambda_1}}}{\frac{Y_1}{\lambda_1}} - e^{\frac{Y_1}{\lambda_1}} \right) + \beta_4 \left(\frac{1-e^{-\frac{Y_1}{\lambda_2}}}{\frac{Y_1}{\lambda_2}} - e^{\frac{Y_1}{\lambda_2}} \right)$$



- Bear Steepening
 - ✓ 장기채 매수 & Floater 매수 → Duration을 희석시킴(낮춤)으로써, 설령 금리가 더 상승하더라도 그로 인한 손실을 막기 위한 것
(듀레이션이 낮을 수록 이자율 상승에 대한 손실 ↓, 일종의 Immunization)
- Rolling Effect etc.

느낀 점 & 보완할 점



감사합니다