딥러닝 시계열 모델 활용 인공지능 연구

다변량 시계열 데이터를 이용한 암호화폐 가격 예측

(Cryptocurrency price prediction using multivariate time series data)

KMU-KIISS AI실무능력인증과정 - 이기현

목차

1. 주제 선정 배경 - 연구동기 및 중요성

딥러닝 시계열 모델 연구의 활성화

미래를 보다 더 정확하게 예측하기 위해 스스로 분석하고 기존에 발견하지 못한 패턴을 찾아낼 수 있는 딥러닝 시계열 모델 연구 활성화

다변량 시계열 모델의 효과 분석

시계열 예측 모형에 여러가지 다양한 입력 변수들을 통해 딥러닝 모델에 다변량 변수가 미치는 영향도를 파악하여 시계열 모델 및 다변량 효과 분석

여러 산업에 적용 가능한 예측 모델 연구

시계열 예측 모델를 통해
유통업, 헬스 케어, 금융 분야등에서 가격예측, 수요예측, 재고관리 최적화등에 적용 가능한 예측 모델 연구

1.주제 선정 배경 - 데이터 셋 설명

- **수집데이터 : 과거 5년간의 가상화폐 지표, 경제 지표, 소비자 지표와 네이버 뉴스 기사 수집**

→ 가상화폐(이더리움)의 거래가, 거래량, 변동가, 이동평균 가격

→ 세계 주요 6개국 통화인 일본 엔, 유로, 영국 파운드, 스위스 프랑, 캐나다 달러, 스웨덴 크로네에 대한 미 달러 환율 및 달러/원화 환율

→ 검색 트래픽 데이터인 구글 트랜드 지수

→ 그래픽 카드의 가격을 결정짓는 핵심 부품인 GPU, RAM 을 생산하는 기업의 주가(엔비디아(NVIDIA), AMD, 삼성전자, 하이닉스)

→ 두바이유, 브랜트유, WTI(서부 텍사스유) 등의 배럴당 가격

→ 금을 비롯한 원자재 등의 실물 자산 대체재 가격

→ 옥수수 커피 코코아 등 곡물 시장 가격

→ 심리적 요인를 반영한 공포 탐욕 지수

→ 한국과 미국 공휴일

1.주제 선정 배경 - 데이터 셋 설명

수집데이터

[네이버 뉴스 크롤링]
LDA(Latent Dirichlet Allocation) 토픽 모델링으로 주제 분류

네이버뉴스크롤링_V1.0.ipynb

파일 수정 보기 삽입 런타임 도구 도움말 모든 변경사항이 저장됨

+ 코드 + 텍스트

```
[ ]  df_topic_news = df_topic_sents_keywords.reset_index()
     df_topic_news.columns = ['Document_No', 'Dominant_Topic', 'Topic_Perc_Contrib', 'Keywords', 'conte

     df_topic_news['Dominant_Topic'] = df_topic_news['Dominant_Topic'] +1
     df_topic_news.Dominant_Topic = df_topic_news.Dominant_Topic.astype(str)
     df_topic_news['Dominant_Topic'] = df_topic_news['Dominant_Topic'].str.split('.').str[0]

     <class 'pandas.core.frame.DataFrame'>

[ ]  df_topic_news
```

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	
0	0	11	0.9356	테라, 루나, 폭락, 클레이튼, 사태, 스테이블코, 권, 지수, 바이낸스, 스테이블	편집자주암호 산은 금융의 같은 거품일
1	1	11	0.7880	테라, 루나, 폭락, 클레이튼, 사태, 스테이블코, 권, 지수, 바이낸스, 스테이블	파이낸셜뉴스 과 동조현상 자산 시장이
2	2	11	0.5273	테라, 루나, 폭락, 클레이튼, 사태, 스테이블코, 권, 지수, 바이낸스, 스테이블	출처 코인니스 으로 온체인
				테라, 루나, 폭락, 클레이	

[뉴스 기사 핵심 키워드 추출]
형태소 분석기와 한국어 키버트(Korean KeyBERT) 이용

	date	contents
0	2022-05-28	편집자주암호화폐와 가상자산은 금융의 미래일까, 도박 같은 거품일까. 블록체인, 비트...
1	2022-05-27	[파이낸셜뉴스] 한동안 나스닥과 동조현상을 보이던 가상자산 시장이 최근에는 나스닥 ...
2	2022-05-26	출처 : 코인니스 데이터 "처음으로 ERC-721 온체인 거래량, ERC-20 추월...
3	2022-05-25	2030세대 거래액 비중 전체의 62.4%여성 거래액 비중 40.2%로 전년比 23...
4	2022-05-24	[파이낸셜뉴스] 프랑스 명품 브랜드 발렌시아가(balenciaga)가 비트코인(BT...

```
    doc_embedding = model.encode([data_text])
    candidate_embeddings = model.encode(candidates)

    kword = max_sum_sim(doc_embedding, candidate_embeddings, candidates, to
    print(f'{i} 번째 단어 : {kword}')
    keyword.append(kword)

# DataFrame 저장
keyword_df = pd.DataFrame(keyword)
keyword_df.to_csv ("/content/drive/MyDrive/Colab Notebooks/실무인증/Data/ke

901 번째 단어 : ['파트너십 디렉터 삼성']
902 번째 단어 : ['대표 스케일 네트워크']
903 번째 단어 : ['블록체인 실비오 칼리']
904 번째 단어 : ['아칸소 교수 우버']
```

1.주제 선정 배경 - 데이터 셋 설명

단계적 선택법(AIC)을 이용한
최적의 회귀 방정식의 설명 변수 선택

trade_price ~ oil_gsl + oil_hgsl + oil_lo + oil_du + oil_brt + oil_cl + cmdt_c + cmdt_gc + c

OLS Regression Results

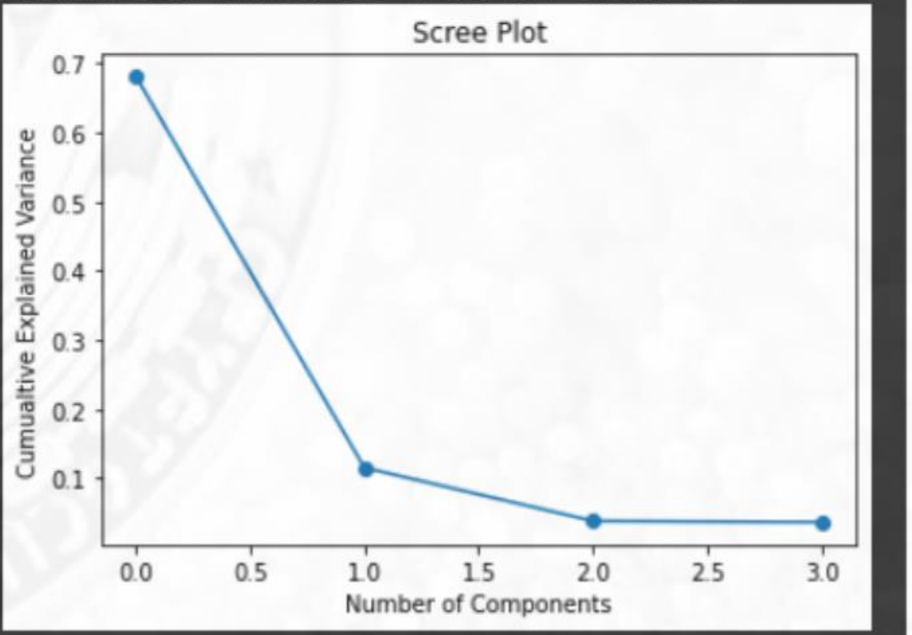
Dep. Variable:	trade_price	R-squared:	1.000
Model:	OLS	Adj. R-squared:	1.000
Method:	Least Squares	F-statistic:	4.909e+07
Date:	Wed, 08 Jun 2022	Prob (F-statistic):	0.00
Time:	09:54:10	Log-Likelihood:	-14329.
No. Observations:	1707	AIC:	2.880e+04
Df Residuals:	1638	BIC:	2.917e+04
Df Model:	68		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.834e+04	1.99e+04	1.422	0.155	-1.07e+04	6.74e+04
oil_gsl	-3.2146	4.363	-0.737	0.461	-11.772	5.343
oil_hgsl	3.1218	5.040	0.619	0.536	-6.764	13.007
oil_lo	0.7015	3.078	0.228	0.820	-5.337	6.740
oil_du	12.3498	19.856	0.622	0.534	-26.595	51.295
oil_brt	-19.4256	27.055	-0.718	0.473	-72.492	33.641
oil_cl	1.7133	15.580	0.110	0.912	-28.846	32.272
cmdt_c	-0.0023	0.057	-0.040	0.968	-0.114	0.109
cmdt_gc	-0.9696	2.323	-0.417	0.676	-5.526	3.586
cmdt_pl	0.6613	1.108	0.597	0.551	-1.511	2.834
cmdt_si	-23.3516	48.481	-0.482	0.630	-118.443	71.740
cmdt_pa	0.3537	0.304	1.163	0.245	-0.243	0.950
cmdt_go	-2.0559	2.649	-0.776	0.438	-7.251	3.139
cmdt_ho	1011.7676	878.432	1.152	0.250	-711.201	2734.736
cmdt_ng	262.4252	105.298	2.492	0.013	55.893	468.958

PCA (principal component analysis)
차원 축소, 주성분 분석법

고유값 :
[39.6151075 16.28266462 9.36325673 9.13786005]

분산설명력 :
[0.68108899 0.11506232 0.03804837 0.03623858]



2. 연구목적 - 문제 정의 및 가정

연구 목적

가상화폐 가격을 예측하기 위해 시계열 예측 모형들을 이용하여 다양한 입력 변수들이 예측 성능에 미치는 영향도 파악

연구 방법

Baseline, DeepAR, TFT 실험과정으로 찾은 최적의 하이퍼 파라미터의 조합으로 시계열 모델을 구축

→ Baseline RNN 기반의 LSTM 모델

→ DeepAR(시계열 데이터 처리에 우수한 확률적 예측 모형)

→ TFT(Temporal Fusion Transformers , Attention 기반 구조 모형)

성능 지표

백분율 오류를 기반으로 한 정확도 측정 방법인 SMAPE(Symmetric mean absolute percentage error) 사용

성능 향상 방법

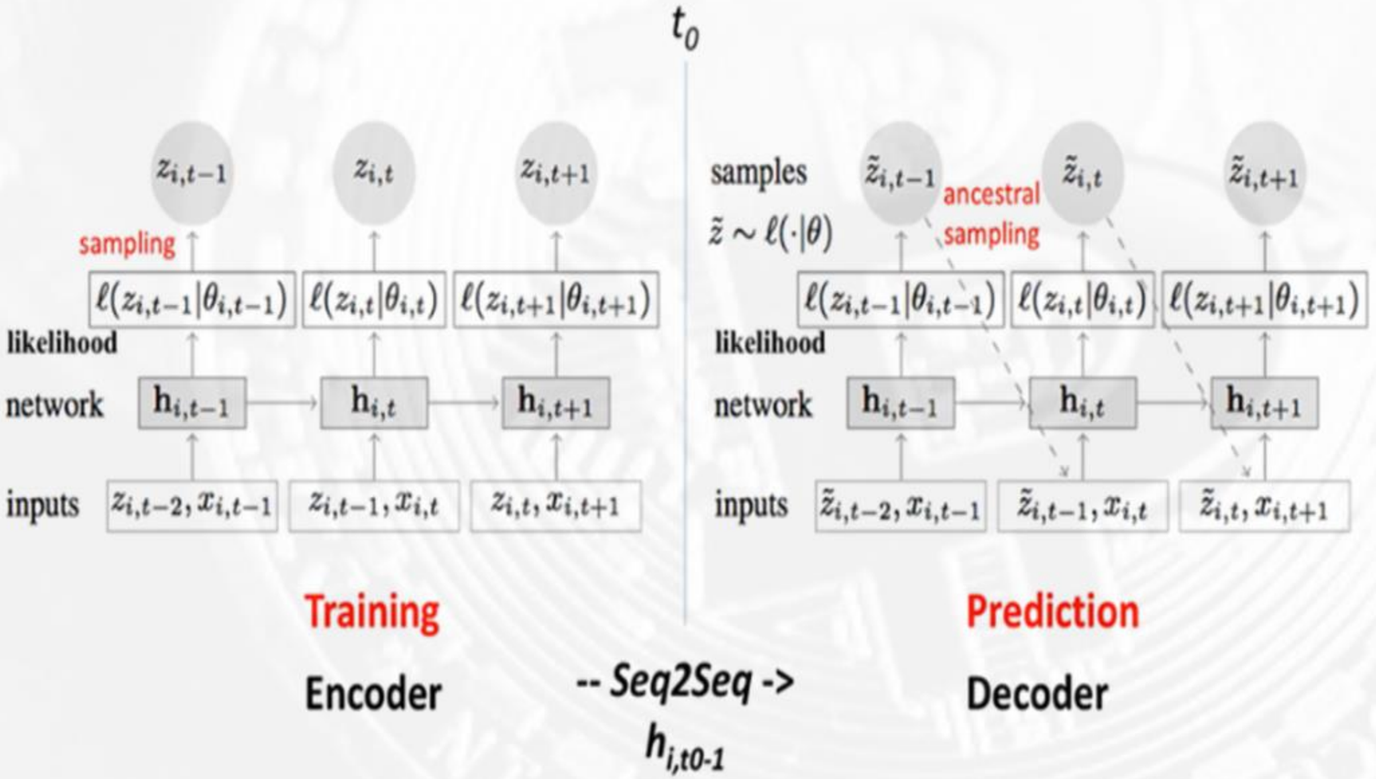
암호화폐와 관련된 뉴스 데이터를 이용하여 텍스트 마이닝 기법을 일부 적용하여 비정형 데이터를 공변량으로 활용

분석에 가치가 있는 항목들에 대한 정보를 수집한 후 적절한 통계학적 절차를 통해 필수적인 공변량을 선택하고

딥러닝 시계열 모델에 반영하여 암호 화폐 예측을 위한 공변량의 효과를 파악

3. 제안 방법론 - 모델 파이프 라인(DeepAR)

Auto-regressive recurrent network model을 기반으로 하는 확률 모형으로
미래의 값이 아니라 미래 확률 분포를 추정하며 여러 공변량을 활용해 학습이 가능

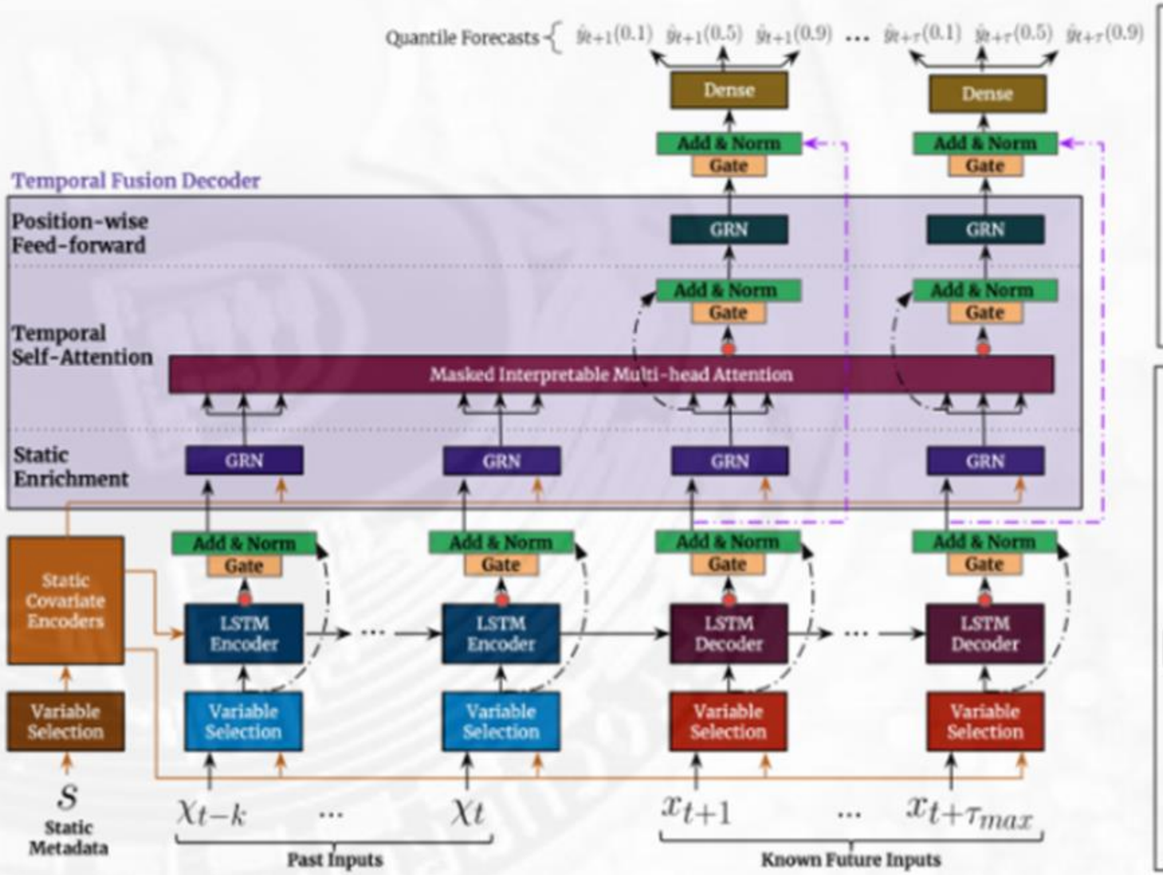


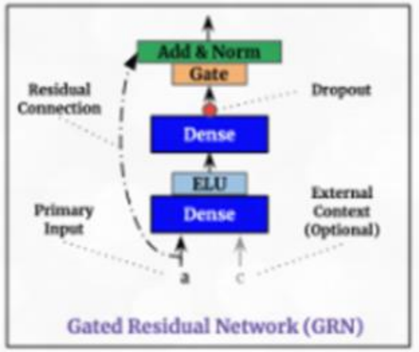
- Probabilistic forecasting
- Negative binomial likelihood
- 교사 강요(Teacher Forcing) 방법
- Covariate(공변량)과 함께 학습
- Cold-start 문제 해결(유사한 제품의 수요 데이터를 활용)

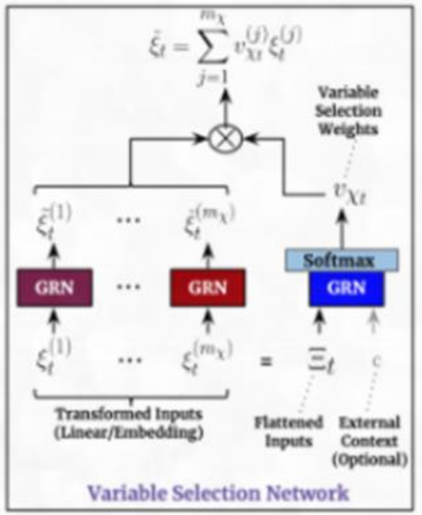
3. 제안 방법론 - 모델 파이프 라인(TFT)

미래에는 알 수 없는 관측 변수(Observed Inputs)들과 함께 알고 있는 변수(Time Varying Known Input),
시간에 따라 변하지 않는 변수인 Static Covariates을 입력으로 활용하여 Multi-Horizon Forecasting을 하는 Attention 기반 구조로 구성

- Gating Mechanisms
- Variable Selection Networks
- Static Covariate Encoders
- Temporal Processing
  (해석 가능한 multi-head attention 기법)
- Prediction Intervals(quantile 예측)



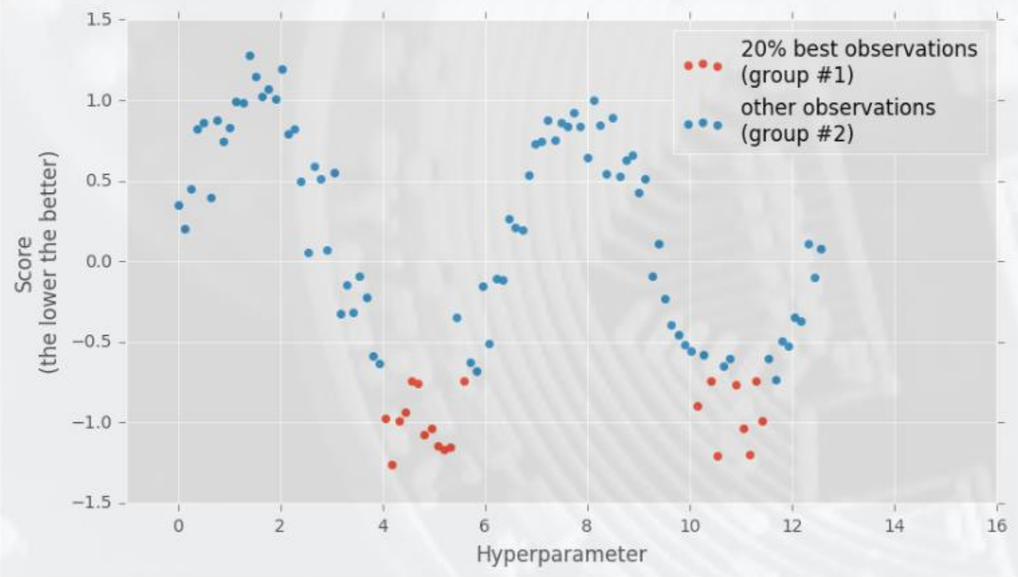


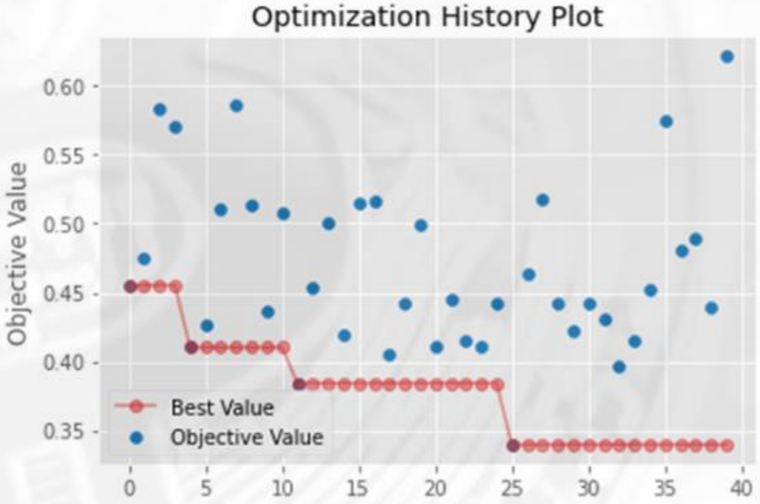


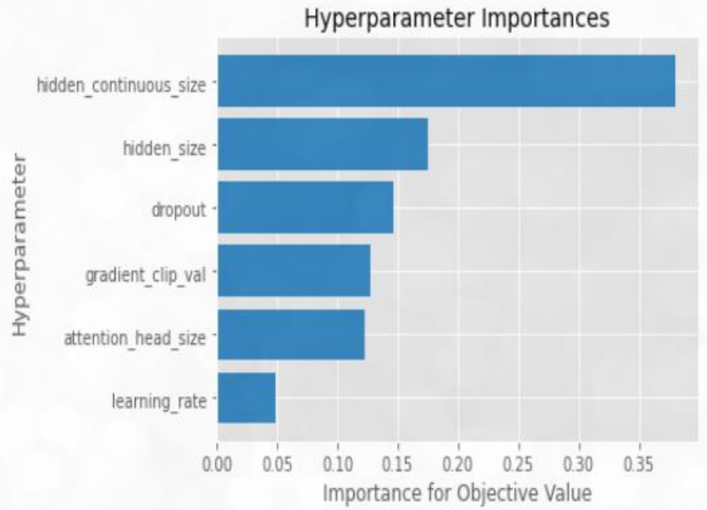
4. 실험 및 결과 - 실험 시나리오 설계

PCA 차원 축소, 회귀 방정식으로 선택된 설명변수들 그리고 전체 데이터 셋에 대해
각각 Baseline인 LSTM, DeepAR, TFT 모델에 적용하여 평가지표인 SMAPE값을 비교, 모델의 성능을 측정
최적화된 하이퍼 파라미터를 구하기 위해 TPE(Tree-Structured Parzen Estimator) Sampler를 이용(optuna)

TPE(Tree-Structured Parzen Estimator)



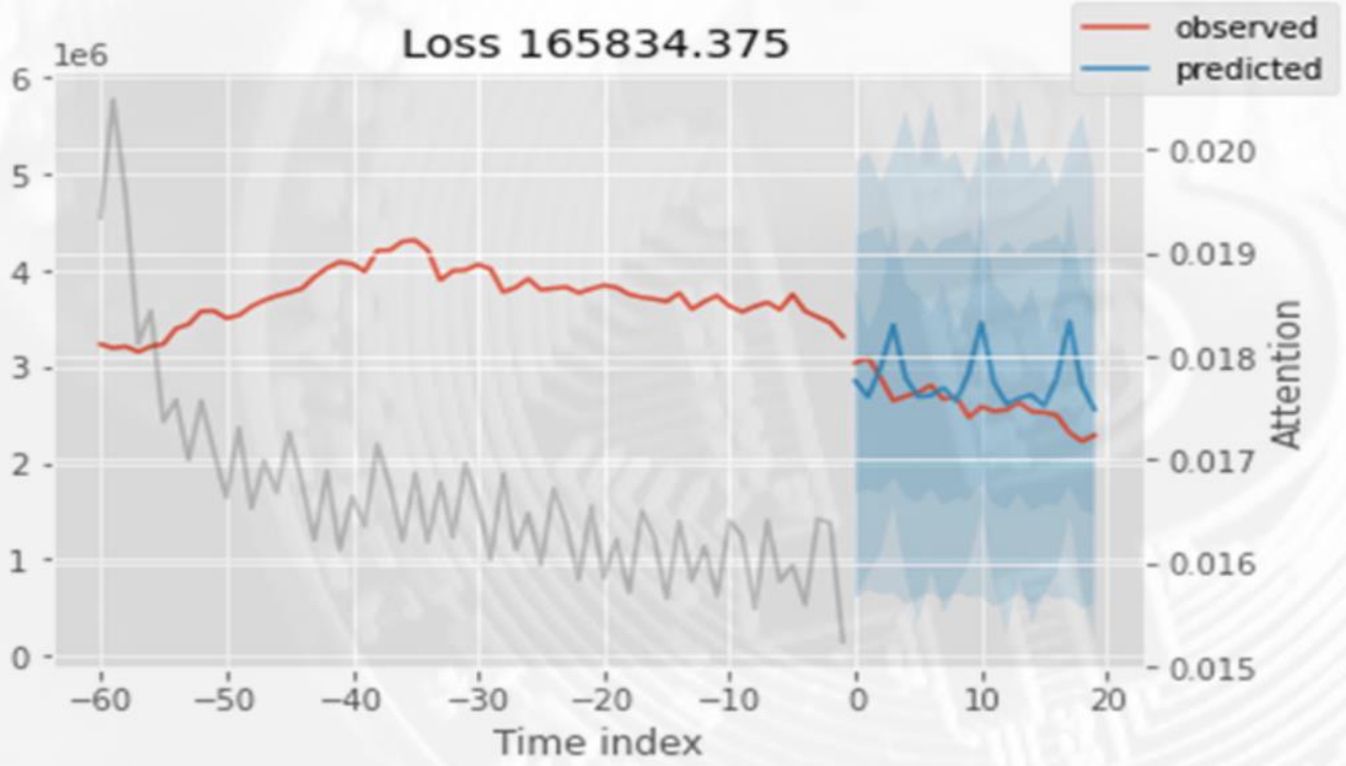





```
Number of finished trials:  25
Best trial:
  Value:  0.09957671910524368
  Params:
    learning_rate: 0.001472046764573504
    hidden_size: 64
    hidden_continuous_size: 32
    attention_head_size: 6
    gradient_clip_val: 0.0003245250702503594
    dropout: 0.2
```

5. 결론 - 모델 성능 결과

Temporal Fusion Transformation

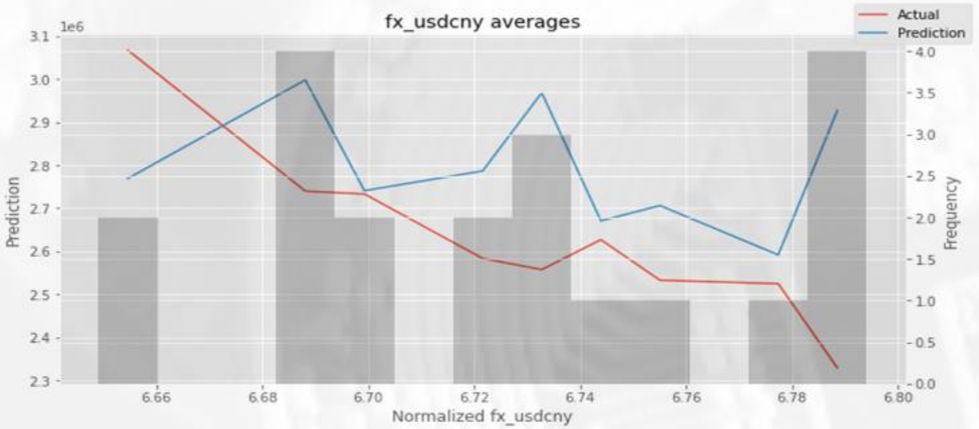


Information on dataset and optimal Series model configuration		Base	Stepwise	PCA
DeepAR	SMAPE	0.156	0.286	0.167
	Network parameters			
	RNN_layers	5	5	2
	Hidden_size	256	128	256
	Training parameters			
	Minibatch size	128	128	128
	Learning_rate	0.037	0.005	0.047
	Gradient_clip_val	0.002	3.00E-04	0.003
TFT	SMAPE	0.113	0.354	0.238
	Network parameters			
	Dropout rate	0.2	0.2	0
	Attention_head_size	6	5	7
	Hidden_size	64	128	32
	Training parameters			
	Minibatch size	128	128	128
	Learning_rate	1.40E-03	1.00E-04	0.8479
	Gradient_clip_val	3.00E-04	0.371	7.00E-04
LSTM	SMAPE	23.6		
	Network parameters			
	RNN_layers	4		
	Hidden_size	16		
	Training parameters			
	Minibatch size	128		
	Learning_rate	0.0149		

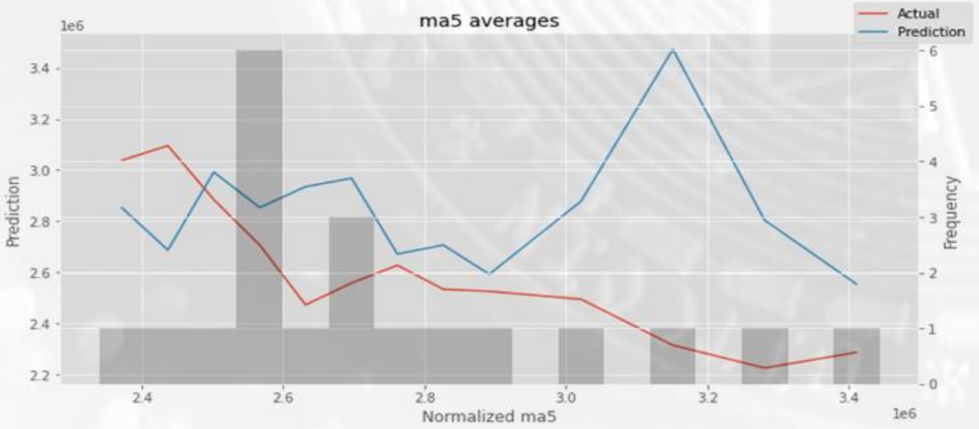
5. 결론 - 모델 성능 결과

TFT 변수들의 예측 값

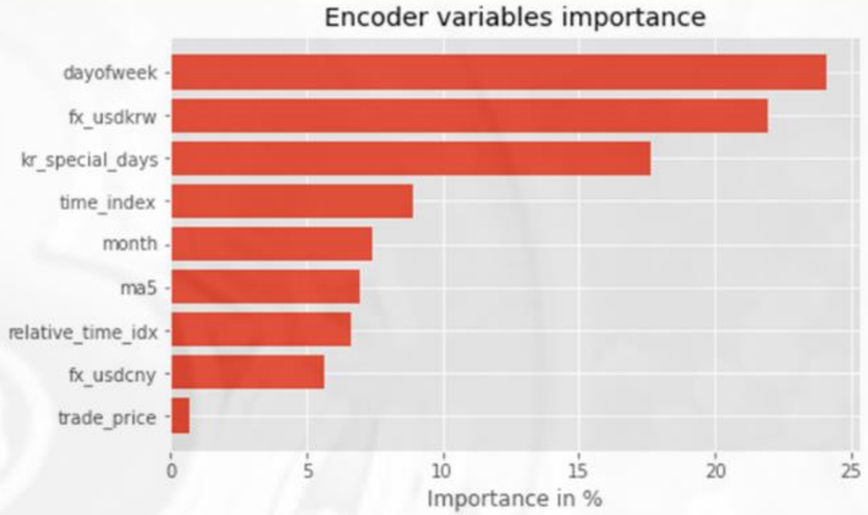
< 달러/위안 환율 >

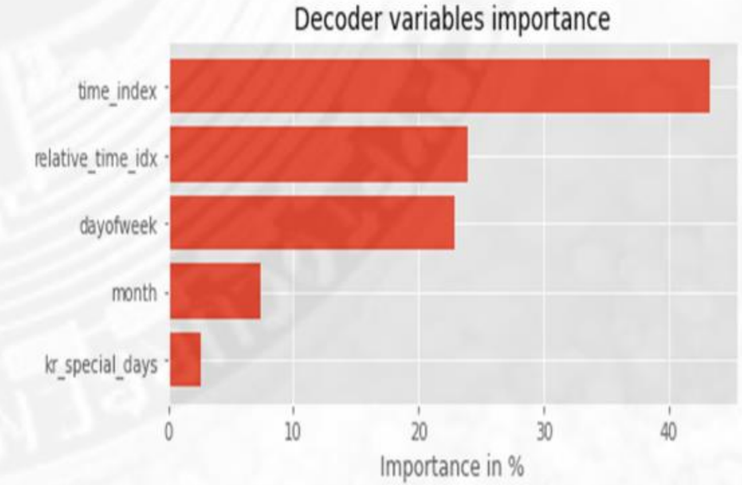


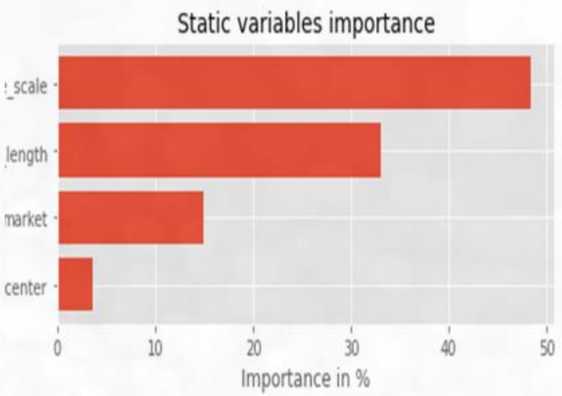
< 이동평균(5일) >



TFT Feature importance

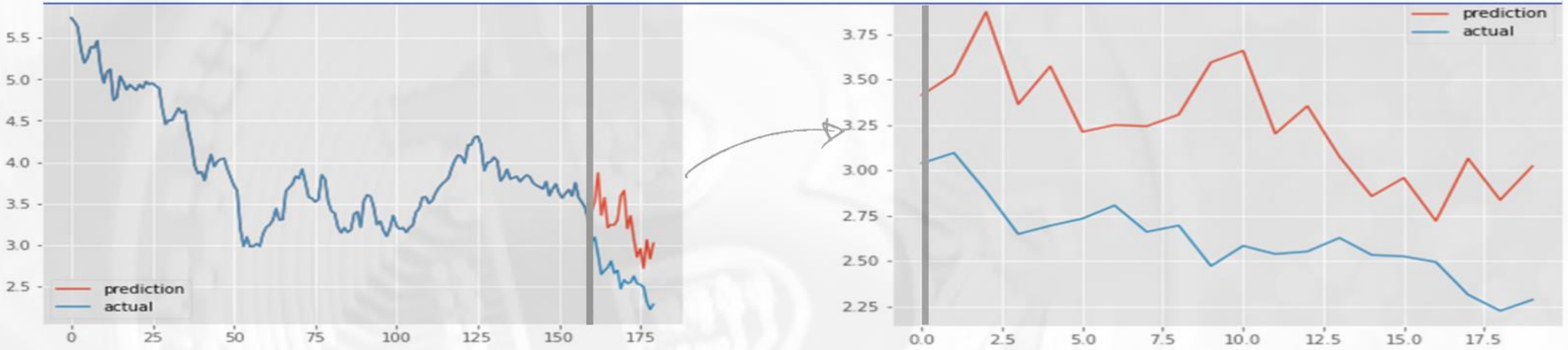




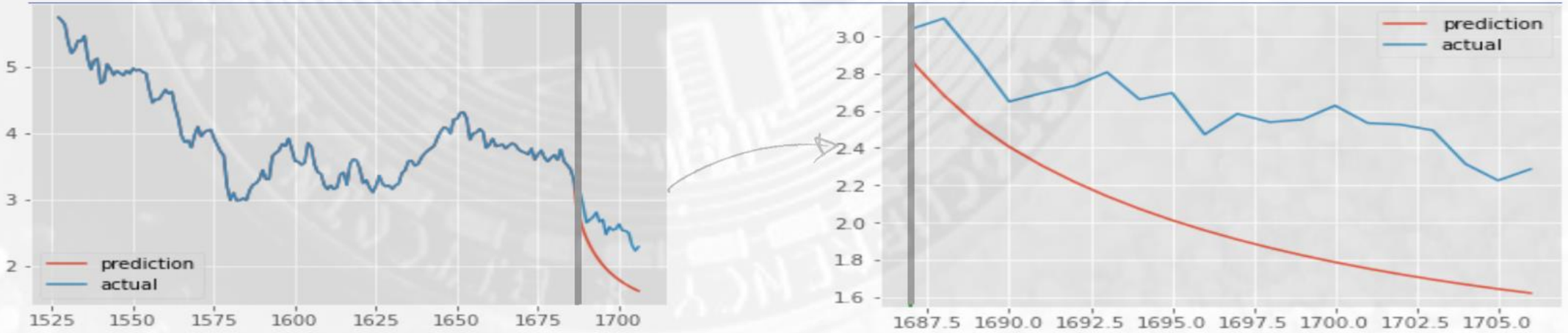


5. 결론 - 모델 성능 결과

DeepAR



LSTM



5. 결론

PCA 차원축소, 변수 선택법을 사용하는것보다
다양한 Input 데이터를 직접 사용하는 경우 DeepAR, TFT 둘다 성능이 비교적 좋았음

Multi-Horizon Forecasting은 DeepAR, TFT가 Baseline인 LSTM보다 우수한 성능을 보여줌

TFT의 Feature importance를 통해 변수들의 중요도를 파악한 결과
미래 시점의 값을 현재 시점에서도 알 수 있는 변수(time_varying_known_categoricals)에 등록한
공휴일, 요일, 월 등이 데이터에 영향을 줌

향후 연구 방향

Temporal Fusion Transfomers에서 입력되는 time_varying_unknown_reals 값이 feature에 따라 성능 편차가 크므로
EDA 단계에서 모델에 적합한 feature들을 찾아내는 것이 필요함
Scale이 큰 값이 들어오는 경우 predict에 대한 오차가 점점 커지는 현상 발생

참고 논문

Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting
DeepAR: Probabilistic forecasting with autoregressive recurrent networks
Attention Is All You Need

QnA

질문이 있다면 말씀해주세요.

https://github.com/ai-practice-course/ki-test