JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 25, No. 7: 884~889, Jul. 2021

시계열 데이터를 활용한 코로나19 동향 예측

김재호[1] · 김장영[2*]

Covid19 trends predictions using time series data

Jae-Ho Kim[1] · Jang-Young Kim[2*]

[1]Graduate Student, Department of Computer Science, The University of Suwon, Hwaseong, 18323 Korea
[2*]Associate Professor, Department of Computer Science, The University of Suwon, Hwaseong, 18323 Korea

요 약

국내 코로나19의 감염자 수가 백신과 사회적 거리 두기, 백신 등 여러 가지 노력 덕분에 차츰 줄어드는 듯 보였으나 2020년 2월 20일 특정한 사건 이후 감염자 수가 증가한 것처럼, 2020년 12월부터 또다시 급격히 감염자 수가 증가하는 추세이며 꾸준히 일일 500명가량의 감염자 수가 이어지고 있다. 따라서 Kaggle의 데이터셋을 이용해서 Prophet 알고리즘을 통해 미래 코로나19를 예측하고 사이킷런을 통해 결정계수, 평균 절대 오차, 평균 백분율 오차, 평균 제곱 차, 평균 제곱근 편차를 통해 이 예측에 대한 설명력을 더한다. 또한 코로나19가 급격히 특정한 사건이 없었을 경우 국내 감염자 수를 예측해 앞으로 우리가 미래의 질병에 대해서 방역과 방역 수칙 실천의 중요함을 강조한다.

ABSTRACT

The number of people infected with Covid-19 in Korea seemed to be gradually decreasing thanks to various efforts such as social distancing and vaccines. However, just as the number of infected people increased after a particular incident on February 20, 2020, the number of infected people has been increasing rapidly since December 2020 by approximately 500 per day. Therefore, the future Covid-19 is predicted through the Prophet algorithm using Kaggle's dataset, and the explanatory power for this prediction is added through the coefficient of determination, mean absolute error, mean percent error, mean square difference, and mean square deviation through Scikit-learn. Moreover, in the absence of a specific incident rapidly increasing the cases of Covid-19, the proposed method predicts the number of infected people in Korea and emphasizes the importance of implementing epidemic prevention and quarantine rules for future diseases.

키워드 : 코로나19, Prophet 알고리즘, 결정계수, 평균 절대 오차, 평균 백분율 오차

Keywords : Covid19, Prophet algorithmn, Coefficient of determination, Mean absolute error(MAE), Mean percentage error(MPE)

Received 17 May 2021, **Revised** 24 May 2021, **Accepted** 5 June 2021

*** Corresponding Author** Jang-Young Kim(E-mail:jykim77@suwon.ac.kr, Tel:+82-31-229-8345)
Associate Professor, Department of Computer Science, The University of Suwon, Hwaseong, 18323 Korea

Open Access http://doi.org/10.6109/jkiice.2021.25.7.884 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

1.1. 개발동기

세계적으로 코로나19가 급증하고 처음으로 코로나19가 발현된지 1년 반이란 시간이 지났음에도 불구하고, 계속해서 마스크를 끼고 다닌다. 백신이 나왔음에도 효과는 미미하고, 심지어 부작용이나 접종을 받았음에도 불구하고 감염되는 사례 등 여러 이유로 코로나19 백신접종을 꺼리고 있는 나라도 있다. 또한 코로나19의 변이종으로 인해서, 다른 백신이 개발되야 한다는 목소리도 나오고 있다. 그에 따라 실시간으로 코로나에 대해서 분석하는 것은 중요한 분야라고 할 수 있다. 본 논문은 시계열 데이터를 이용해서 실시간 및 시각적으로 코로나 19를 예측한다. 코로나19 현상이 계속해서 지속된다면, 앞으로의 코로나19의 추세를 예측하고, 현시점에서 안일하게 생각하는 코로나 19를 다시 강조한다.

1.2. 코로나19 정보

SARS-CoV-2의 병원체로 전파 경로는 사람과 사람 사이 전파는 인플루엔자와 비슷하게 호흡기 분비물을 통해서 발생한다. 사람의 재채기, 기침이나 말을 할 때, 음식물을 섭취할 때 나오는 호흡기 분비물에 포함된 바이러스가 나오고 타인의 점막에 닿으면 감염될 수 있다. 또한 감염은 사람이 오염된 표면, 물건을 만지거나 눈, 코, 입을 비비거나 만질 때 발생할 수 있다. 호흡기 분비물은 일반적으로 2미터 이상을 가지 않는다. 잠복기는 1~14일이고(평균 5~7일) 코로나19 진단 기준은 임상과 양상에 관계없이 검사기준에 따라 코로나19 감염에 대하여 양성이 뜬 자이며, 검사 기준은 코로나19 유전자(PCR) 검출, 바이러스 분리된 경우이다. 주요 증상 및 징후에는 임상 증상은 무증자부터, 경증, 중등증, 중증까지 다양하다. 주요 증상은 발열, 기침, 인후통, 후각 및 미각 소실, 어지러움, 오한, 두통, 근육통, 호흡곤란, 식욕감퇴, 피로, 가래, 객혈, 결막염, 흉통, 소화기증상, 콧물이나 코막힘, 피부염 증상 등이 다양하게 나타난다. 치료법으로는 특이치료제는 존재하지 않으며 증상에 따른 진해제, 해열제, 수액공급 등 대증적인 치료를 한다. 호흡 곤란 시 체외 막 산소공급, 기계 호흡 등을 시행하며, 산소 치료가 필요한 환자에서 렘데시비르나 칼레트라의 효과가 일부 확인되었다. 따라서, 우리나라를 포함한 여러 다른 나라에서 약품에 대한 긴급 승인이 되었거나, 승인 준비 중이며, 특이적인 항바이러스제 없다.[1]전 세계 치사율은 지역적, 인구 집단연령 구조적, 감염 상태 및 기타 요인 등에 따라 0.1~25%로 다양하다. 전 세계 치사율은 0.00%~1.63%(중위값 0.27%)이다.[2]

1.3. 국내 코로나 동향

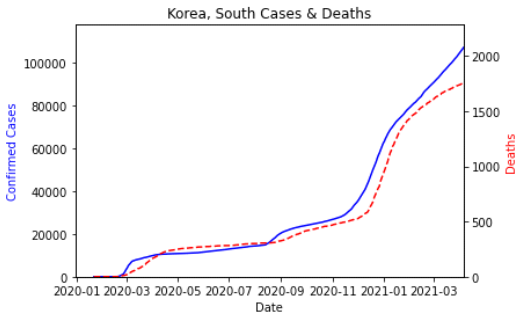

Fig.1 Corona Trends in Korea

그림1은 국내 코로나 감염자 수와 사망자 수를 보여준다. 코로나 감염자수가 점차 줄어들며 방역에 성공하는 듯 보였으나, 2020년 12월 가량부터 현재까지 코로나 감염자 수와 사망자 수가 급격히 증가하고 있다. 따라서 전국적으로 사회적 거리두기가 시행중이다. 그에 따라 방역수칙으로 인해서 다중이용시설이나 일상과 사회, 경제적 활동에 많은 제약이 따르고 있다.[3]

Ⅱ. 기존연구

2.1. SEIR모형을 이용한 전염병 모형 예측 연구

질병 확산 모형은 질병의 확산 과정을 모형화 함으로써 질병이 발생하고 퍼지는 시간 내에서 통제하기 위하여 활용하고자 하는 모형이다. 본 연구에서는 질병 확산 모형의 가장 대표적인 SIR 모형에 기본적인 확장 접근을 하여 접촉군이라는 단계를 추가한 SEIR 모형을 이용하여 모형 구축을 했다. 이 모형은 감염 대상군의 사람들이 질병에 노출 된 잠복기를 거쳐 일정 시간이 경과한 후 감염되어 감염군으로 이동한 후 다시 회복군으로 이동하는 모형이다. 이와 같이 질병에 감염된 후 감염력이 생기는 잠복기가 있는 경우에 연구에 활용될 수 있다. 본 연구에서는 2015년 국내에서 발생한 메르스 코로나 바이러스에 의한 호흡기 감염증 자료를 수집했다. 질병

의 확산 과정이 결정적이 아닌 확률적인 흐름을 따른다고 가정하여 포아송 확률과정을 따른다고 보고 확률적 화학반응 모형을 이용하여 모형을 구축했다. 모형을 구현하기 위해 SEIR 모형의 세 모수인 질병에 노출된 정도를 나타내는 접촉률, 질병의 감염 정도를 나타내는 감염률, 질병의 회복정도를 나타내는 회복률을 추정하고, SEIR 모형에 적합하고 전염병 확산에 대한 예측을 수행하였다. 또한 접촉군이 정확하게 관찰되지 않을 부분을 보완하기 위하여 접촉군을 생성하는 과정을 전체 모형 구축 했다.[4]

2.2. 코로나바이러스와 미래 질병 대응을 위한 과제

중국에서 보고되기 시작한 코로나 바이러스 감염증19(이하, 코로나19) 확진자가 지난 2020년 1월 20일부터 우리나라에서도 지속적으로 보고되고, 점차 여러 국가로 확산되고 있는데, 국내외 전문가들은 앞으로의 상황에 대해 각기 다르게 전망하고 있다. 코로나19는 지금까지 경험하지 못한 새로운 감염병으로, 정의하고 예측하고 대응하는 것은 어려운 과제다. 국내의 다양한 변화 속에서, 특히 초저출산 및 인구고령화 가속, 신종 감염병 및 재출현 감염병의 국내 유입과 유행, 기후변화·미세먼지 등 환경보건 부문의 건강 피해 증가, 4차 산업혁명에 따른 보건의료 분야 대응의 변화에 대한 관심이 필요하다. 이를 진단하고 미래에 다가올 위협에 대비할 필요가 있다. 코로나19 사태가 보건 당국뿐 아니라 경제, 외교, 교육, 환경 등 다양한 부문과의 연계·협력을 요하는 것과 같이 미래 질병 대응에는 보건정책뿐 아니라 다부처 협력과 융·복합 전략이 필요하다. 공중보건정책의 비전과 미래 질병 어젠가에 대비할 수 있도록 보건 당국의 역할을 지지하고 역량을 강화할 수 있는 기반을 마련해야 한다.[5]

Ⅲ. 분석알고리즘

일반 데이터는 순서와 상관없이 표의 값을 중요시해 값을 통계를 내지만, 시계열 데이터는 인덱스에 시간이 있고, 데이터들이 시간을 가지고 있어 시간별로 데이터가 나열돼있다. 시계열 데이터는 시간적으로 어떻게 변화하는지가 중요하다. 즉, 변화를 포착한다. 따라서 코로나 19를 Prophet을 이용해 시계열 데이터를 분석, 예측하고 그 예측을 평가한다. 평가 방법은 사이킷런을 통한 결정계수와 MAE, MSE, RMSE, MPE을 이용한다.

3.1. Prophet 알고리즘 [6], [7]

Prophet 모델 주요 구성요소로는 Trend, Seasonality, Holiday이다. 이 세 가지를 통해 다음과 같은 공식으로 나타낼 수 있다.

$$y(t) = g(t) + s(t) + h(t) + \epsilon i \quad (1)$$

- g(t) : Trend를 구성하는 g(t) 함수는 주기적이지 않은 변화인 트렌드를 나타낸다. 이 경향은 부분적으로 선형 또는 로지스틱 곡선으로 이루어지며, 비주기적 변화를 반영하는 추세함수이다. 시계열 데이터가 시간에 따라 증가하거나 감소 또는 일정 범위 값을 유지할 때, 그에 대한 경향 보여준다.

$$g(t)=\frac{C(t)}{1+\exp(-(k+a(t)^T\delta)(t-(m+a(t)T\gamma)))} \quad (2)$$

C - Carrying Capacity, k - growth rate

m - Offset parameter

- s(t) : weekly/yearly 등 주기적으로 나타나는 패턴들을 포함하는 것을 말한다. 주기적 변화를 반영하는 함수이다(ex)주간/연간). 즉, 반복되는 효과를 주며, 일정한 빈도로 주기적으로 반복되는 패턴이다. 특정한 달/요일에 따라 기댓값이 달라진다.

$$S(t) = \sum_{n=1}^{N}\left(a_n\cos\left(\frac{2\pi nt}{P}\right) + b_n\sin\left(\frac{2\pi nt}{P}\right)\right) \quad (3)$$

P - Period unit

- h(t) : 휴일과 같이 불규칙한 이벤트들을 나타낸다. 만약에 특정 기간에 값이 비정상적으로 증가하거나, 감소했다면 Holiday로(불규칙적 이벤트) 정의하여 모델을 반영할 수 있다. 여기서 이벤트의 효과를 독립으로 가정하고 이벤트 앞뒤로 범위를 지정해 해당 이벤트의 영향력의 범위를 설정해야 한다.

$$Z(t) = \left[\left(t\in D_1\right),...,1\left(t\in D_L\right)\right] \quad (4)$$

h(t) = Z(t)k

- t - Specific time , D - Holiday list
- ϵi : 정규분포라고 가정한 오차를 말한다.

3.2. 결정계수(Coefficient of determination)

결정 계수는 상관 계수와는 대조적으로 변수 간 서로 영향을 주는 정도, 인과 관계 정도를 나타낸 수치이다. 결정계수는 추정한 모형이 주어진 값이나 자료에 얼마나 적합한 정도의 값이다. 결정계수의 값은 0에서 1사이의 값이며, 종속변수와 독립변수 사이에 상관관계가 높을수록 값이 1에 가까워진다. 만약 타깃의 평균정도를 예측하는 수준이라면 결정계수는 분자와 분모가 비슷해져 0에 가까워지고, 예측이 타깃에 가까워지면 분자가 0에 가까워지기 때문에 값이 1에 가까워진다. 즉, 결정계수가 0에 가까울수록 유용성이 낮고, 결정계수의 값이 1에 가까울수록 유용성이 높다고 할 수 있다.[8]

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \quad (5)$$

- SST = Total Sum of Squarews : 관측값에서 관측값의 평균을 뺀 결과의 총합이다.
- SSE = Explained Sum of Squares : 추정값에서 관측값의 평균을 뺀 결과의 총합이다.
- SSR = Residual Sum of Squares : 관측값에서 추정값을 뺀 값 즉, 잔차의 총합이다.

3.3. MAE(Mean Absolute Error, 평균 절대 오차)

모델의 예측값과 실제값의 차를 모두 합한 후 절댓값을 취하고 평균을 낸다. 직관적으로 알 수 있는 지표지만 절댓값을 가지기 때문에 값이 under performance인지 over performance인지 알 수 없다. 이 값은 평균 백분율 오차를 통해 알 수 있다.[9]

- under performance: 모델이 실제보다 낮은 값으로 예측
- over performance: 모델이 실제보다 높은 값으로 예측

3.4. MPE(Mean Percentage Error, 평균 백분율 오차)

모델이 underperformance 인지 overperformance 인지 판단할 수 있다. MPE값이 양수이면 under, 음수이면 over이다.[9]

3.5. MSE(Mean Squared Error, 평균 제곱 차)

실제 값과 예측값을 빼고 제곱해 평균화한다. 예측값과 실제값 차이의 면적의 합이다. 지표 자체가 직관적이고 단순하다. 값이 작을수록 에러율이 적다는 것이다. 특이 값이 존재하면 수치가 많이 늘어난다. 제곱을 하기 때문에 음수는 존재하지 않으며 MAE와는 다르게 수치가 모델의 예측값과 실제값 차의 면적의 합이다. 따라서, 특이 값이 존재할수록 값이 많이 늘어난다. [9]

3.6. RMSE(Root Mean Squared Error, 평균 제곱근 편차)

예측한 값과 실제 환경에서의 값의 차를 다룰 때 주로 사용한다. 정밀도를 나타내며, 각각의 차이 값을 잔차(residual)라고 한다. 평균 제곱근 편차는 잔차들을 하나의 값으로 종합할 때 사용된다. MSE은 오류의 제곱을 취하므로 실제 오류 평균보다 훨씬 커지는 특성이 있다. 따라서, MSE에 루트를 취한 RMSE 값을 사용한다. 에러값에 제곱을 취하기 때문에 에러값이 클수록 가중치가 크게 반영된다. 값이 작을수록 정밀도가 높다고 측정한다. 이상치에 민감하며, 해석이 쉬워진다.[9]

Ⅳ. 실험결과 및 분석

4.1. 코로나19의 시계열분석 시각화

Prophet 모델 학습을 생성하고 결과를 시각화한다. test size만 입력받아 데이터프레임을 트레이닝용 데이터프레임과 테스트용 데이터프레임으로 나눠준다. 전체 테스트 사이즈가 0.3이라 하면, 0.7만큼을 트레이닝에 쓰고 나머지를 테스트에 사용한다. (df.shape[0])×(1-test_size) 따라서 전체 중에 테스트 사이즈를 1에서 뺀다. 본 데이터는, 397일 동안 training, 44일 동안 test 했다.[3]

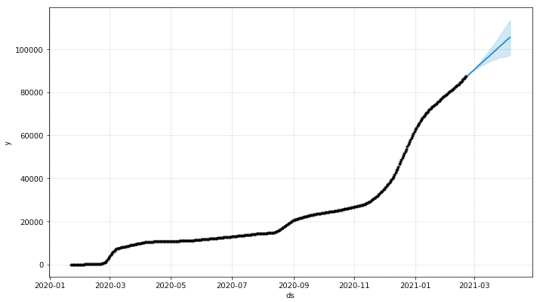

Fig.2 Prophet Model Graph

그림2에서 검정색 선이 training에 사용된 실제 과거 데이터이고, 뒤에 파란색은 prediction에 사용된 test 데이터이다. 뒷부분에 떨어진 부분은 급격한 데이터를 따르지 못한다. 후반부 변화를 잘 따르고, 추정한다. 가운

데 선이 실제 prediction이고 위아래는 신뢰도 95% 구간을 그려준다. 또한, 전체적으로 증가하는 트렌드이다.

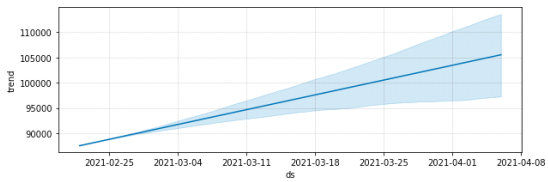

Fig.3 Covid19 trend

그림3은 전체적으로 코로나가 증가하는 트랜드임을 보여준다. 파란색 범위는 95%의 신뢰구간을 보여준다.

4.2. 실제 데이터와 모델의 비교

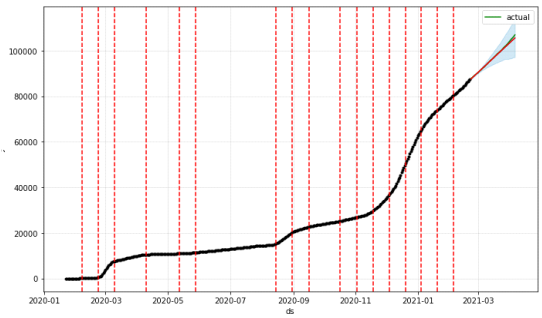

Fig.4 Compare actual and predicted values

그림 4는 빨간선이 코로나19의 예측값, 초록선은 실제값이다. 신뢰구간은 95%를 보여준다.

Table.1 Compare actual and predicted values

Date	Actual Value	Prediction Value	Error Value
2021-02-22	87681	87,508	173
2021-02-23	88,120	87,938	182
2021-02-24	88,516	883,75	141
2021-02-25	88,922	88,814	108
2021-02-26	89,321	89,247	74
2021-02-27	89,676	89,674	2
2021-02-28	90,031	90,062	-31
2021-03-01	90,372	90,438	-66
2021-03-02	90,816	90,868	-52

표1은 실제값과 예측값의 차이를 보여준다.

4.3. 모델 학습 결과 평가

test_df와 pred를 비교하여 r2_score 모델로 결정계수를 평가한다. 결정계수 값이 1에 가까울수록 잘 예측된 것 0에 가까울수록 예측되지 않았다는 것을 의미한다. 또한 평균 절댓값 오차는 모델의 예측값과 실제값의 차를 모두 합한 후 절댓값을 취하고 평균을 내기 때문에 에러의 크기를 그대로 반영한다. 때문에 직관적으로 알 수 있는 지표이며, MPE는 MAE에서 알 수 없는 under performance인지, over performance 인지 알 수 있다. MSE는 예측값과 실제값의 차이를 제곱해 평균화한다. RMSE는 통해 모델의 정밀도를 알 수 있다.

Table.2 Result evaluation value

Coefficient of determination	0.9144467335874181
MAE	1340.9974356183473
MPE	-19.39964443585403
MSE	5886837.758508826
RMSE	2426.280642981934

표2에선 평가 값을 보여준다. 결정계수는 0.9144467335874181이며 인과관계의 정도 즉, 예측 정도를 말한다. 1에 가까울수록 예측이 잘되었다는 것을 의미하기 때문에 위 모델은 예측이 잘되었음을 뜻한다. MAE는 모델의 예측값과 실제값의 차를 모두 더한다. 직관적으로 알 수 있는 지표이며 평균적으로 하루당 1340~1341명 차이가 나는 것을 보여준다. MPE는 음수이므로 실제값보다 예측값이 더 낮다는 것을 의미한다. MSE값은 예측값과 실제값 차이의 면적 값이다. RMSE 값은 $\sqrt{MSE}$값을 취한다.

4.4. 특정한 사건이 없었을 경우 코로나19 예측

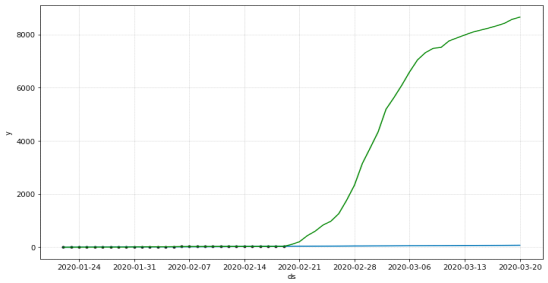

Fig.5 Prediction of Corona 19 without specific events

그림5는 확진자 수가 급증하기 전 데이터를 학습 후,

해당 시점으로부터 30일간의 데이터 예측하여 실제 값과 비교한다. 초록색 선은 2020년 2월 20일경 특정 종교 집단에서 집단 감염이 발생하고 그 이후 코로나 감염자 수가 급증했음을 보여준다. 만약 이러한 사건이 없었을 경우 코로나19 감염자 수 예측은 파란색이고, 거의 증가하지 않음을 알 수 있다.

Ⅴ. 결론 및 향후연구

본 논문은 현재 심각하게 대두되고 있는 코로나19를 시계열 데이터분석을 하고 예측하고 시각적으로 보여준다. FBProphet 알고리즘을 통해서 데이터를 train, test 하고 그 값을 직접적으로 비교한다. 이 모델의 결정계수와 평균 절대 오차(MAE), 평균 백분율 오차(MPE), 평균 제곱 차(MSE), 평균 제곱근 편차(RMSE)로 평가한다.

2020년 2월 사건처럼 심각하게 급증하는 구간들이 있는데 그에 대한 심각성을 강조한다. 백신이 나왔지만, 백신 접종자들의 집단감염 사례, 변종 바이러스 등장에 따라 방역 수칙을 철저히 지켜야 할 필요가 있다.

향후 연구로 특정한 집단감염 사례를 포함한 데이터와 포함하지 않은 데이터를 비교해서 분석한다면 조금 더 완벽한 데이터를 얻을 수 있다. 우리가 쉽게 코로나19를 예측할 수 없는 것은 갑자기 급증하는 집단감염 사례나, 코로나 변이 바이러스 때문이다. 이것을 잘 이용한다면, 더욱 정확한 예측을 할 수 있다.

REFERENCES

[1] Coronavirus Infectious Disease-19(Covid-19) [Internet]. Available: http://ncov.mohw.go.kr/.

[2] Infection fatality rate of COVID-19 inferred from seroprevalence data (Bulletin of the World Health Organization. 2021), [Internet]. Available: https://www.who.int/bulletin/volumes/99/1/20-265892/en/.

[3] Covid19 Cases, Deaths Data set [Internet]. Available: https://www.kaggle.com/antgoldbloom/covid19-data-from-john-hopkins-university.

[4] M. J. Do, J. T. Kim, and B. S. Choe, “A study of epidemic model using SEIR model,” *Journal of the Korean Data and Information Sciencde Society*, vol. 28, no. 2, pp. 297-397, Mar. 2017.

[5] S. M. Chae, “Coronavirus Infectious Disease-19 and Challenges for Future Disease Response,” *Korea Institute for Health and Social Affairs*, no. 374, pp. 1-8, Mar. 2020.

[6] S. J. Taylor and L. Benjamin, “Forecasting at Scale,” *Facebook*, Menlo Park, vol. 72, no. 1, pp. 37-45, Sep. 2018.

[7] FBProphet [Internet]. Available: https://facebook.github.io/prophet/docs/quick_start.html.

[8] S. W. Kim, “Basic Statistivcs,” Hakjisa, pp. 127, 2016.

[9] API-Reference-scikit-learn [Internet]. Available: https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics.

김재호(Jae-Ho Kim)

수원대학교 컴퓨터학부 학사
수원대학교 컴퓨터학부 석사과정
※관심분야 : 인공지능

김장영(Jang-Young Kim)

연세대학교 컴퓨터과학 공학사
Pennsylvania State Univ. 공학석사
State University of New York 공학박사
University of South Carolina 교수
수원대학교 컴퓨터학부 교수
※관심분야 : Big data, AI, Cloud computing, Networks