

「2022 연구데이터 분석활용 경진대회」

# 코로나19(COVID-19) 대응정책 수립을 위한 업종별 매출 예측 시스템 구축

Team : 이어드림

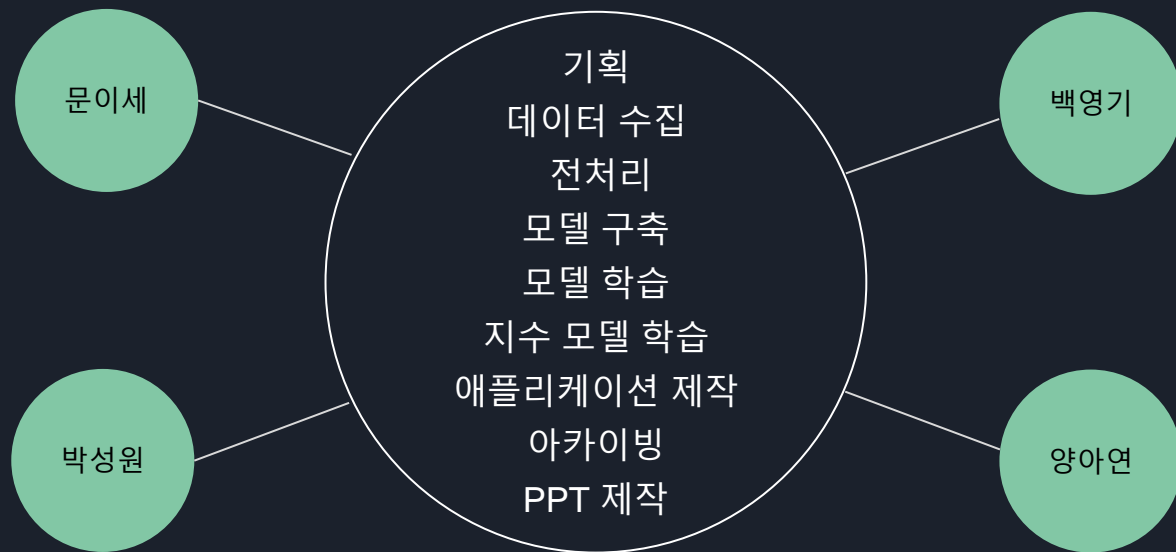


# 목 차

1. Team “YEAR DREAM”
2. 문제 인식
3. 이전 연구
4. 연구 목적
5. 데이터 수집
6. 데이터 전처리
7. 모델링
8. LSTM 모델 학습 결과
9. 위험도 지수 산출
10. 결론
11. DataON 실행 환경 및 결과

# Team “YEAR DREAM”

소속: 중소벤처기업부 “2022년 스타트업 AI기술인력 양성사업” <이어드림 스쿨 2기>



# 단골집이 사라졌어요 !



# 문제 인식

## 신속한 대응 부족

"가게 망한 뒤 지원금 주면 뭐 하나?"...소  
상공인 지원 매뉴얼 마련 시급



정부가 내년 1월 지급을 목표로 하는 '3차 재난지원금'을 놓고 자영업자와 소상공인들의 불만이 높아지고 있다. 사회적 거리두기 강화에 따라 영업에 타격을 입은 건 지난달인데, 정부는 아직까지도 지원 대상과 그 방법 등을 구체화하지 못했기 때문이다.

정부는 지원이 늦어지는 것에 대해 "피해 규모를 정확히 파악해야 제대로 된 선별 지원을 할 수 있다"는 신중론을 펼치고 있다. 하지만 제때 지원이 이뤄지지 않으면 소상공인 등에 게 실질적인 도움이 될 수 없는 만큼, 피해가 발생하면 즉시 지원이 가능한 '지원 매뉴얼'을 지금이라도 마련해야 한다는 지적이 나오고 있다.

2020년 12월

## 적절한 지원기준 부재

YTN

여야, 코로나19 피해 보상 논의...불평등 해소 대책은 온도 차

2021년 01월 15일 11시 54분



[앵커]

코로나19로 피해를 본 업종과 자영업자들의 반발이 커지면서 여야가 일제히 피해 업종들을 만났습니다.

다만, 코로나19 불평등 상황을 해결할 대책을 두고는 서로 다른 방향을 잡았는데요.

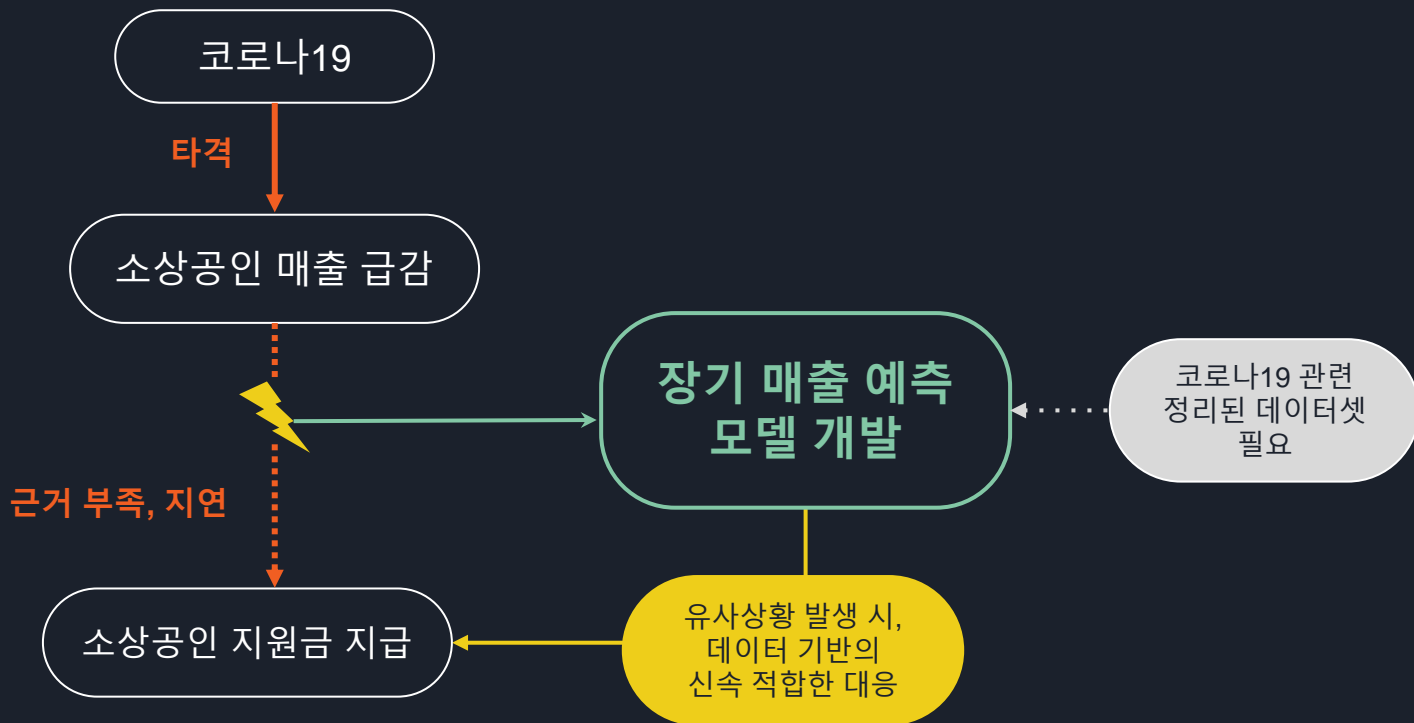
국회 취재기자 연결해 자세한 소식 알아보겠습니다. 송재인 기자!

먼저 더불어민주당 소식부터 알아보죠.

전국의 자영업자들을 만나 코로나19 피해 상황을 점검했죠?

2021년 1월

# 문제 인식



# 이전 연구

## 코로나 관련

- 노윤아, et al. "사회적 변수를 고려한 LSTM 기반 코로나 19 일별 확진자 수 예측 기법." 정보과학회 컴퓨팅의 실제 논문지 28.2 (2022): 116-121.
- 배진수, and 김성범. "머신러닝 모델을 이용한 대한민국 코로나 신규 확진자 예측." 대한 산업공학회지 47.3 (2021): 272-279.

## 매출 관련

- 유현지. "코로나 19 와 서울시 골목상권의 매출액 영향요인에 관한 연구." 한국지역개발학회지 33.3 (2021): 45-75.
- 선충녕, 조민희, and 송사광. "매출 데이터를 이용한 침수 재난 소상공인 운영피해 추정 방법." 한국 정보과학회 학술발표논문집 (2017): 42-44.
- 건물 단위 운영 피해 추정 기술 개발 (Development of Operational Damage Estimate Technology by Building) 한국과학기술정보연구원 연구보고서 (2017)
- 신성호, et al. "딥러닝을 활용한 날씨 빅데이터와 소상공인 매출 분석." (2016).

⇒ 코로나 관련요인으로 소상공인 매출을 예측하려는 시도는 없었음



# 연구 목적

## 1. 서울시 소상공인 업종별 매출 예측

- 업종별 매출 급감 시점 조기 감지 가능

## 2. 매출 급감 예상 업종에 대한 맞춤형 지원 정책 제안에 기여

- 매출 위험도 지수를 통해 업종별 상대적인 매출 변화량 확인
- 적절한 지원 대상 선택 및 객관적 지원 근거 확보





# 데이터 수집

코로나19 확진자 수와 매출에 영향을 줄 것으로 예상되는 요인 탐색

- 집단감염
- 백신 접종률
- 여행경보
- 거리두기 단계
- 인구 이동량
- 확진자 수
- 업종별 신한카드 매출 데이터

- 공휴일/추석연휴
- 요일 정보
- 기상청 날씨 데이터
- 경제통계 100대 지수
- 코로나 신규 확진자 수 증감 데이터 (증감 수, 증감률)

⇒ 총 12개 카테고리에 대한 데이터셋 (요인 323개),  
2018-01-01 ~ 2022-06-30 (1642일) 동안의 데이터

# 데이터 수집 - 활용 데이터 종류



## 질병관리청

- 집단감염
- 거리두기 단계



## 서울특별시 빅데이터 캠퍼스

- 업종별 신한카드 매출 데이터



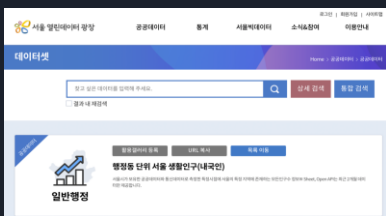
## 공공데이터포털

- 백신 접종률
- 여행경보
- 공휴일/추석연휴
- 확진자 수



## 기상자료개방포털

- 기상청 날씨 정보



## 서울 열린데이터 광장

- 인구가동량



## 한국은행 경제통계시스템

- 경제통계 100대 지수

# 활용 데이터셋 명칭

date	herd_infection	vaccine_1st_rate	vaccine_2nd_rate	vaccine_3rd_rate	vaccine_4th_rate	alarm_lv	continent_eng_nm_Africa	continent_eng_nm_America	continent_eng_nm_Asia
2018-01-01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2018-01-02	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2018-01-03	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2018-01-04	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2018-01-05	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2018-01-06	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2018-01-07	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2018-01-08	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2018-01-09	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2018-01-10	2022-05-29	0.0	87.7	63.8	6.9	0.0	0.0	0.0	0.0
2018-01-11	2022-05-30	0.0	87.8	63.8	7.000000000000000C	0.0	0.0	0.0	0.0
2018-01-12	2022-05-31	0.0	87.8	63.8	7.000000000000000C	0.0	0.0	0.0	0.0
2018-01-13	2022-06-01	0.0	87.8	63.8	7.1	0.0	0.0	0.0	0.0
2018-01-14	2022-06-02	0.0	87.8	63.8	7.1	0.0	0.0	0.0	0.0
2018-01-15	2022-06-03	0.0	87.8	63.9	7.200000000000000C	0.0	0.0	0.0	0.0
2018-01-16	2022-06-04	0.0	87.8	63.9	7.200000000000000C	0.0	0.0	0.0	0.0
2018-01-17	2022-06-05	0.0	87.8	63.9	7.200000000000000C	0.0	0.0	0.0	0.0
2018-01-18	2022-06-06	0.0	87.8	63.9	7.200000000000000C	0.0	0.0	0.0	0.0
2018-01-19	2022-06-07	0.0	87.8	63.9	7.200000000000000C	0.0	0.0	0.0	0.0
2018-01-20	2022-06-08	0.0	87.8	63.9	7.200000000000000C	0.0	0.0	0.0	0.0
2018-01-21	2022-06-09	0.0	87.8	63.9	7.200000000000000C	0.0	0.0	0.0	0.0
2018-01-22	2022-06-10	0.0	87.8	63.9	7.3	0.0	0.0	0.0	0.0
2018-01-23	2022-06-11	0.0	87.8	63.9	7.3	0.0	0.0	0.0	0.0
2018-01-24	2022-06-12	0.0	87.8	63.9	7.3	0.0	0.0	0.0	0.0
2018-01-25	2022-06-13	0.0	87.8	63.9	7.3	0.0	0.0	0.0	0.0
2018-01-26	2022-06-14	0.0	87.8	63.9	7.400000000000000C	50.0	2.0	0.0	1.0
2018-01-27	2022-06-15	0.0	87.8	63.9	7.400000000000000C	0.0	0.0	0.0	0.0
2018-01-28	2022-06-16	0.0	87.8	63.9	7.400000000000000C	0.0	0.0	0.0	0.0
2018-01-29	2022-06-17	0.0	87.8	63.9	7.400000000000000C	0.0	0.0	0.0	0.0
2018-01-30	2022-06-18	0.0	87.8	63.9	7.400000000000000C	0.0	0.0	0.0	0.0
2018-01-31	2022-06-19	0.0	87.8	63.9	7.400000000000000C	0.0	0.0	0.0	0.0
2018-02-01	2022-06-20	0.0	87.8	63.9	7.400000000000000C	0.0	0.0	0.0	0.0
2018-02-02	2022-06-21	0.0	87.8	63.9	7.5	0.0	0.0	0.0	0.0
2018-02-03	2022-06-22	0.0	87.8	63.9	7.5	0.0	0.0	0.0	0.0
	2022-06-23	0.0	87.8	63.9	7.5	0.0	0.0	0.0	1.0
	2022-06-24	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	2022-06-25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	2022-06-26	0.0	87.8	64.0	7.5	0.0	0.0	0.0	0.0
	2022-06-27	0.0	87.8	64.0	7.5	0.0	0.0	0.0	0.0
	2022-06-28	0.0	87.8	64.0	7.6	0.0	0.0	0.0	0.0
	2022-06-29	0.0	87.8	64.0	7.6	75.0	3.0	0.0	0.0
	2022-06-30	0.0	87.8	64.0	7.6	75.0	3.0	4.0	1.0

## [ Feature dataset ]

- 모델 학습을 위해 입력되는 input features 데이터셋
- 2018-01-01 ~ 2022-06-30 (1642일) 동안의 데이터
- 활용 데이터 종류: 외부 데이터 (*이전 슬라이드 (p9) 참고*)
  - ✓ 질병관리청
  - ✓ 공공데이터포털
  - ✓ 서울 열린데이터 광장
  - ✓ 기상자료 개방포털
  - ✓ 한국은행 경제통계시스템
- 외부 데이터 수집하여 하나의 데이터프레임으로 통합

# 활용 데이터셋 명칭

x	Curve1
2018-01-01	807810663.5
2018-01-02	693383655
2018-01-03	1115616175
2018-01-04	1141933209
2018-01-05	1373647263
2018-01-06	1736374519
2018-01-07	1076132407
2018-01-08	751161494.1
2018-01-09	1218021644
2018-01-10	1143644202
2018-01-11	1194563679
2018-01-12	1365626054
2018-01-13	1657414931
2018-01-14	1141925368
2018-01-15	825532664.3
2018-01-16	1228886321
2018-01-17	1245479099
2018-01-18	1254633428
2018-01-19	1387368520
2018-01-20	1716915030
2018-01-21	1088137690
2018-01-22	754012170.2
2018-01-23	1041223605
2018-01-24	1091566594
2018-01-25	1086419041
2018-01-26	1201414293
2018-01-27	1510366520
2018-01-28	998877725.1
2018-01-29	762015102.9
2018-01-30	1120741591
2018-01-31	1184247069

x	Curve1
2018-01-01	108844924
2018-01-02	489227682
2018-01-03	269968412
2018-01-04	2149007777
2018-01-05	220483189
2018-01-06	243123803
2018-01-07	83998167
2018-01-08	288404396
2018-01-09	227181065
2018-01-10	696304716
2018-01-11	439352803
2018-01-12	554711340
2018-01-13	150853135
2018-01-14	121498169
2018-01-15	402441418
2018-01-16	282512465
2018-01-17	562093510
2018-01-18	1137207614
2018-01-19	232185556
2018-01-20	190250796
2018-01-21	111415170
2018-01-22	280818827
2018-01-23	244754198
2018-01-24	773928234
2018-01-25	408284174
2018-01-26	354608644
2018-01-27	201133783
2018-01-28	109718490
2018-01-29	684629775
2018-01-30	372209279
2018-01-31	764928151

x	Curve1
2018-01-01	11498272.1
2018-01-02	327280628.2
2018-01-03	314121401.7
2018-01-04	214713815.6
2018-01-05	198338185.1
2018-01-06	302422594.5
2018-01-07	16180178.9
2018-01-08	267044743.7
2018-01-09	205936777.9
2018-01-10	244235467.4
2018-01-11	187805721.2
2018-01-12	178155671.4
2018-01-13	281367334.8
2018-01-14	5354290.7
2018-01-15	285751106.1
2018-01-16	193063901.1
2018-01-17	265865926.5
2018-01-18	193061946.1
2018-01-19	200371024.4
2018-01-20	295980371.4
2018-01-21	13242025.8
2018-01-22	255041993.1
2018-01-23	191887317.8
2018-01-24	215570630.4
2018-01-25	171124657.8
2018-01-26	180480724.9
2018-01-27	282522600.6
2018-01-28	6802143
2018-01-29	257666614.8
2018-01-30	185739742.5
2018-01-31	308830175.5

## [ Target dataset ]

- 모델 예측에 사용되는 target 데이터셋
- 2018-01-01 ~ 2022-06-30 (1642일) 동안의 각 업종의  
일별 카드매출 데이터
- 활용 데이터 종류: 외부 데이터 (*이전 슬라이드 (p9) 참고*)
  - ✓ 서울특별시 빅데이터 캠퍼스

.....

# 데이터 전처리

데이터명	전처리 전	전처리 후
거리두기 단계	노래방 22시 이후 영업제한 (24시간 중 8시간 제한)	수치화 $8H / 24H = 0.333$
백신 접종률	누적 접종률 50%	$50 / 100 = 0.5$
여행경보 단계	총 4단계 중 2단계	$2 / 4 = 0.5$
여행 제한 국가	텍스트 데이터	one-hot encoding
날씨 데이터, 인구이동량 등..	원본 데이터	scikit-learn을 이용한 minmax 스케일링
코로나 신규 확진자 수 증감 데이터	누적 확진자 수	전일 대비 증감 수 계산, 전일 대비 증감률(%) 계산



# 모델링

- 사용한 Model 종류들
  - Machine Learning Model
    - Random Forest
    - XGBoost
    - Auto-regressive Integrated Moving Average (ARIMA)
    - Prophet
  - Deep Learning Model
    - Long Short-Term Memory (LSTM)
- 사용한 평가지표 : MSE



# 모델링 - 선정

- 모델별 '노래방 업종' MSE 비교

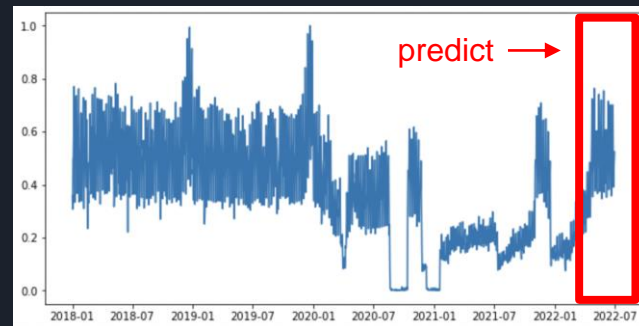
Machine Learning Model				Deep Learning Model
Random Forest	XGBoost	ARIMA	Prophet	<b>LSTM</b>
0.0258	0.0131	0.0570	0.08173	0.0020

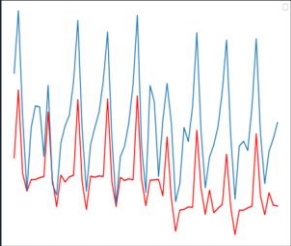
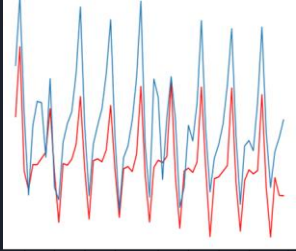
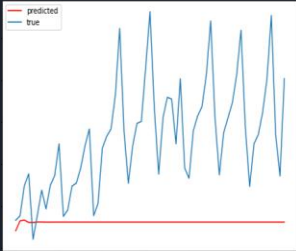
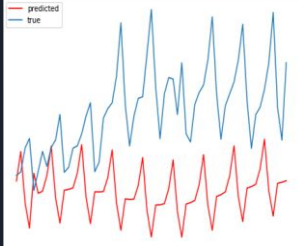
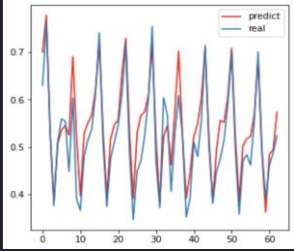


# 모델링 - 선정

- 모델별 '노래방 업종' 예측 그래프 비교 (일별)

predict / real



Machine Learning Model				Deep Learning Model
Random Forest	XGBoost	ARIMA	Prophet	LSTM
				

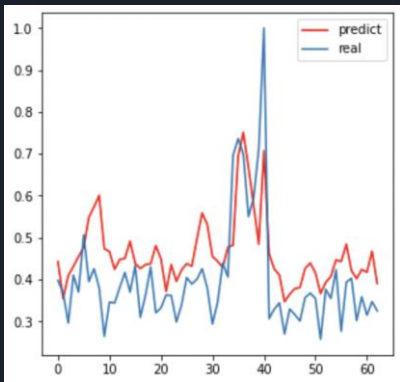
# LSTM 모델 학습 결과

다른 업종별 모델 학습, 예측

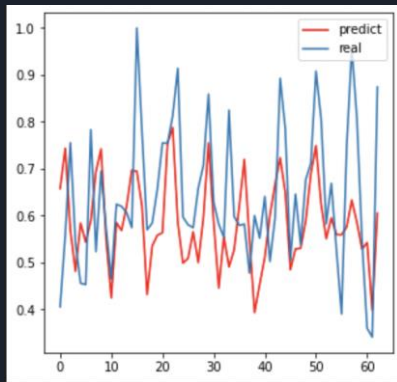
- 생활잡화(SB015), 실외골프/스키(SB026), 세탁소(SB043), 가구(SB057)

predict / real

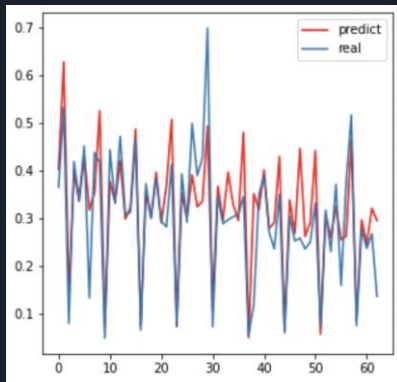
생활잡화(SB015)



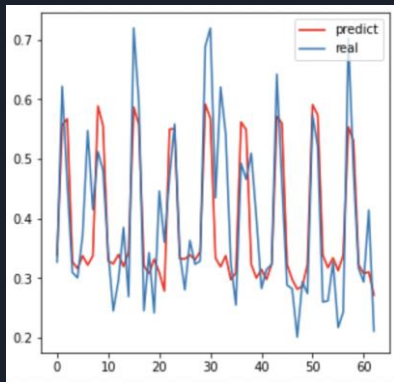
실외골프/스키(SB026)



세탁소(SB043)



가구(SB057)

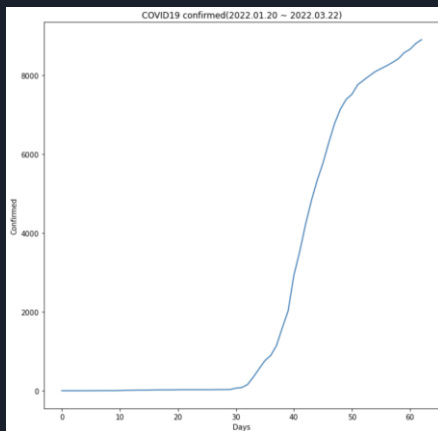


# 위험도 지수 산출

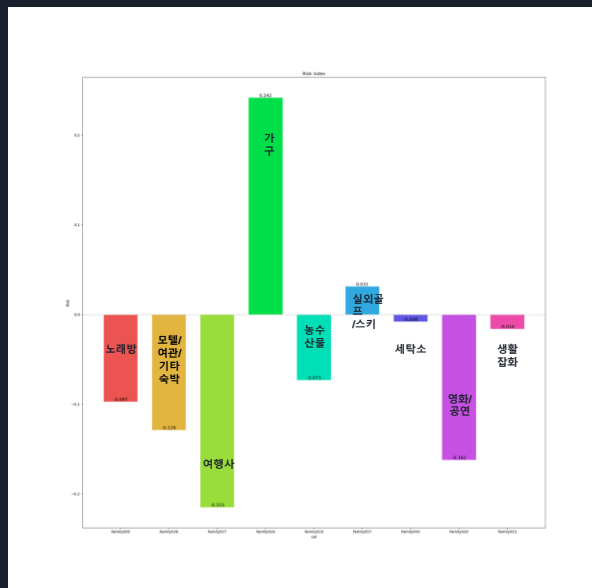
2020년 1주차~9주차 코로나19 대유행 기간 - 업종별 매출 위험도 지수

- 예측 대상 기간동안의 업종별 매출 추이를 '위험도 지수(risk index)'로 표현
- 위험도 지수 산출 방법  
(R1 : 4~6주차 매출액/1~3주차 매출액, R2 : 7~9주차 매출액/4~6주차 매출액)

$$RiskIndex = \sqrt{(R_1) * (R_2)} - 1$$



동 기간동안 확진자 수





# 결론

## 차별성

- 코로나 팬데믹 상황에서 복합적인 요인을 고려하여 직접 매출을 예측한 모델
- 데이터셋 구축
  - 규제 정책과 같이 수치화 되지 않은 비정형 데이터를 수치화
  - 산재되어 있던 데이터를 하나로 취합

## 우수성

- 매출 예측 결과를 바탕으로 위험도 지수 만들어 객관적인 지표로 사용할 수 있게 함

## 활용성 - 연구 확장 가능성

- 전국 단위 예측 모델로 확장 가능
  - DataOn 부산 ‘소상공인 매출 집계 데이터’ 활용 등
- 예측 목표 기간을 늘려 장기적 전망 확인

# DataON 실행 환경

## 애플리케이션 (1) predict\_sales\_with\_pandemic

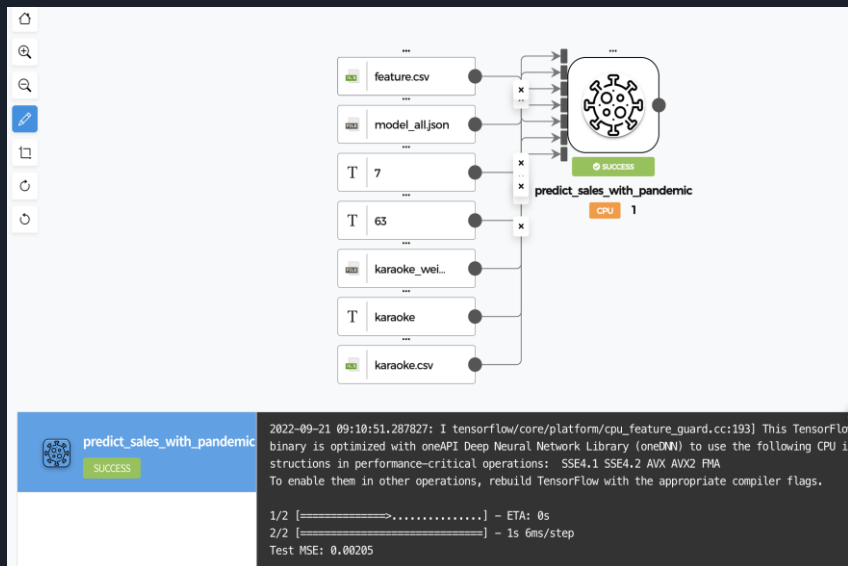
기능: 학습된 모델을 사용하여 매출 예측

입력 데이터: 7개

- model\_path : best model structure json file
- target\_name : 예측하려는 업종명
- target\_path : target dataset
- weight\_path : best model의 weight
- feature\_path : 모델 학습시 사용되는 feature dataset
- predict\_days : 예측하려는 일 수 (ex. 63)
- combined\_days : 예측값을 합산하여 출력할 일 수(ex. 7)

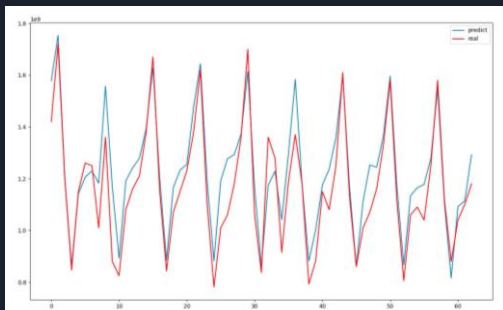
출력 데이터: 3개

- 주별 예측값 : (2) visualize\_sales\_risk\_index에 사용
- 주별 예측 그래프
- 일자별 예측 그래프



# DataON 실행 결과

## 애플리케이션 (1) predict\_sales\_with\_pandemic



resultkaraoke\_predict\_by\_weeks.csv

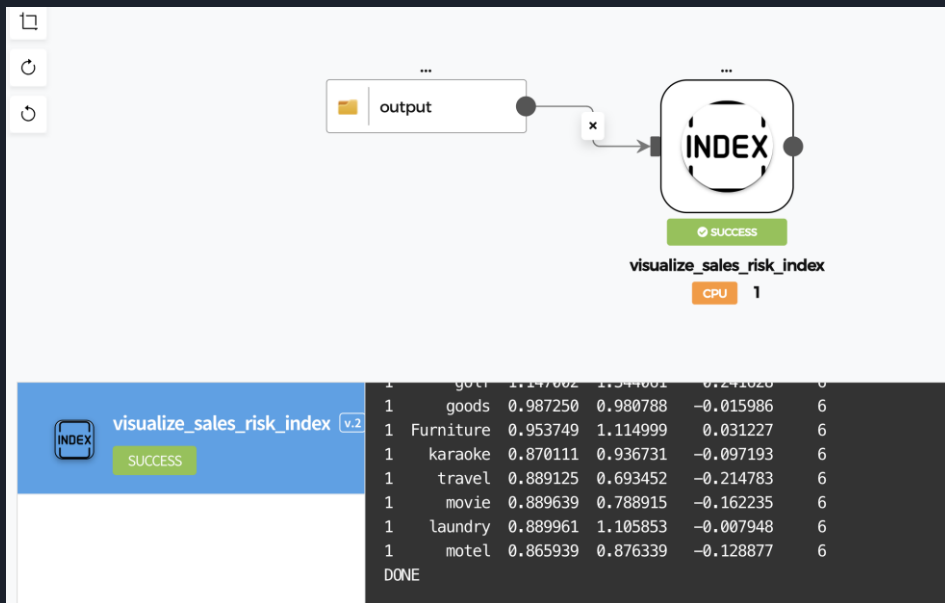
	A	0	Family
1	0	8967239000.0	karaoke
2	1	8512291300.0	karaoke
3	2	8759930000.0	karaoke
4	3	8953184000.0	karaoke
5	4	8456766000.0	karaoke
6	5	8355613000.0	karaoke
7	6	8594948000.0	karaoke
8	7	8471506000.0	karaoke
9	8	8243752400.0	karaoke
10			

✓ 선정된 모델(**LSTM**)로 매출액 예측 결과,  
좌측 상단 그래프와 같이 실제값을 잘 예측하는  
것을 보여줌

✓ **MSE = 0.002** 높은 예측 성능을 보여줌

# DataON 실행 환경

## 애플리케이션 (2) visualize\_sales\_risk\_index



기능: 예측된 매출액을 사용하여 위험도 지수 생성

입력 데이터: 1개

- inpfiler1 : 예측된 매출액 .csv 파일

출력 데이터: 3개

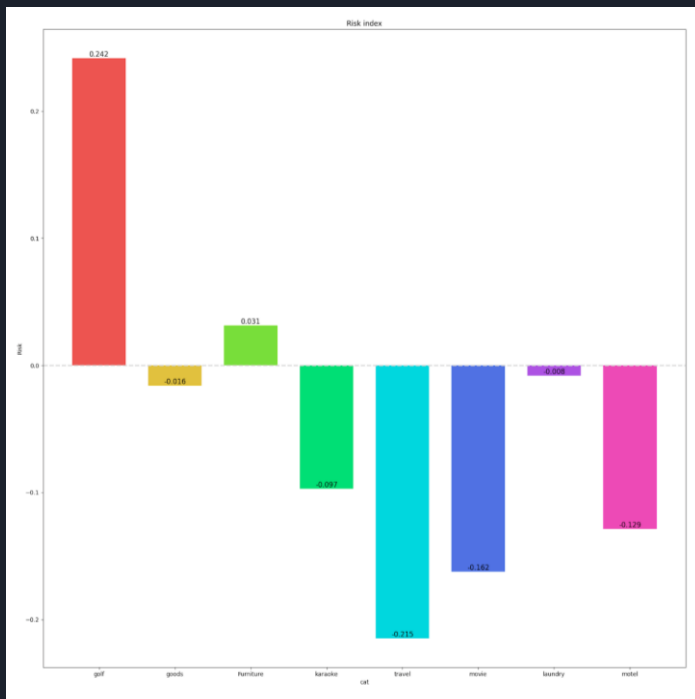
- 위험도 지수 그래프

- 예측 그래프(단위:1주)

- 예측 그래프(단위:3주)

# DataON 실행 결과

애플리케이션 (2) visualize\_sales\_risk\_index



- ✓ 업종별 위험도 지수를 이용해  
앞으로의 매출 전망 예측 가능
- ✓ 지원정책 수립 시 가장 큰 비율로 감소한  
업종을 우선적으로 지원할 수 있도록  
참고 자료로 활용 가능





**감사합니다!**



# Q & A