문화 · 관광 업종의 특징에 따른 코로나-19 영향 분석

DATA060
이상언

INDEX

1) 배경분석

업종별로 타격 차이 有

"코로나19, 대다수 업종에 큰타격…디스플레이,제약, 통신,온라인 유통은긍정적"[출처] Copyrights 디지털세정신문

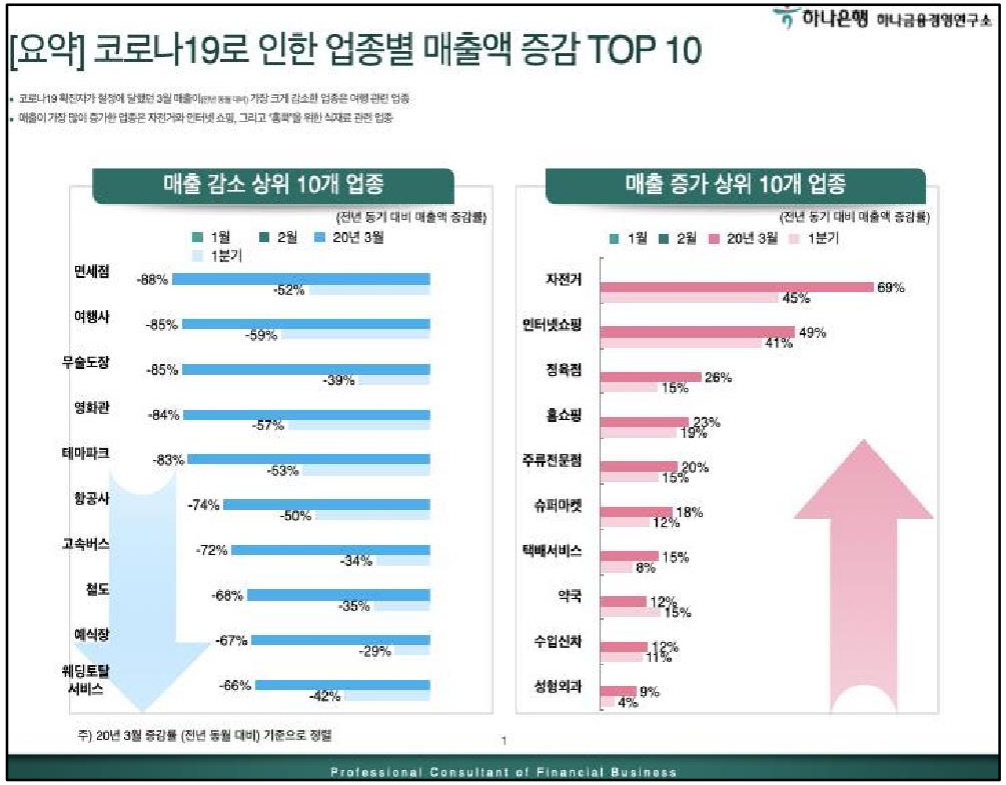
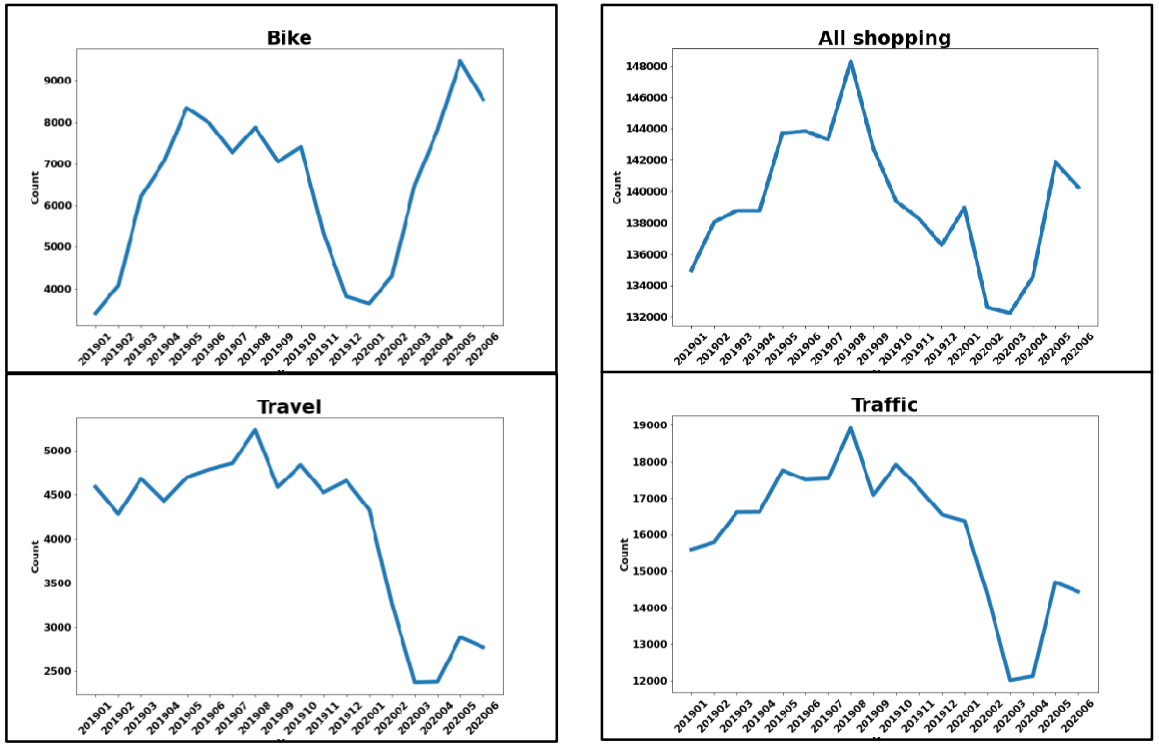


Fig. 코로나19로 인한 업종별 매출액 증감 TOP 10

[출처] 한국세정신문 하나금융경영연구소, 코로나19 소비 변화 보고서

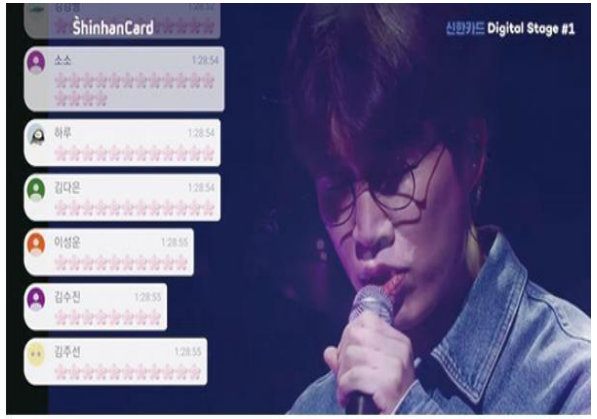
신한카드데이터를 사용하여 만든 업종별 매출건수 차트
(긍정적- 자전거, 종합 쇼핑/ 부정적- 여행사, 교통)



1) 배경분석

② 문화· 관광 산업의 성공적인 언택트화 전략

"실황보다 더뜨거운 '방구석 랜선 콘서트'…댓글로 '떼창' 부르고 '야광봉' 응원까지"



△지난 4월 26일 신한카드 유튜브 공식 계정을 통해 실시간 방송된 밴드 '소란'의 콘서트 장면. 팬들은 채팅창을 볕 모양 이모티콘으로 도배하며 '야광봉 응원'에 나섰다. <신한카드 제공>

신한카드는 지난 4월부터 신한카드 디지털스테이지라는 이름으로 랜선 콘서트를 이어가고 있다. 신한카드 유튜브 공식계정을 통해 공연을 라이브로 송출하는 방식이다. 4월 26일 밴드 소란의 무대를 시작으로 5월 8일 차이콥스키 협주곡 콘서트를 선보였다. 이후에도 국악, 클래식, 대중음악을 아우르는 다양한 장르의 공연을 계속 이어갈 계획이다.

반응은 뜨겁다. 소란의 랜선 콘서트는 동시접속자 1300명이 약 2시간 동안 꾸준히 유지됐다. 공연 중 전체 접속자 수는 7000명이 넘었고 댓글은 7만 5000개에 달한 것으로 집계됐다.

서울관광재단, 코로나19 발생 이후 여가·관광 트렌드 변화 분석

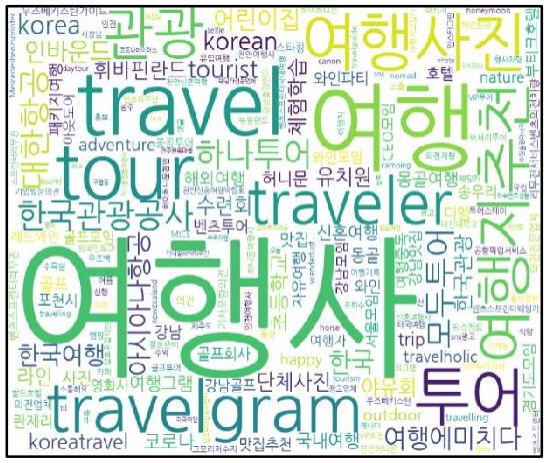
▲코로나19로 인한 집콕현상 장기화로 증가한 '홈' 연관어

[출처]매일경제 티티엘뉴스

2) SNS 워드클라우드

인스타그램의 해시태그를 수집하여 워드클라우드로 시각화

코로나 이전(2017년 1월~ 2020년 1월)



코로나 이후(2020년 2월 이후)

2) SNS 워드클라우드

코로나 이후 증가 키워드
코로나 이후 감소 키워드

SNS 워드클라우드 : 영화

오프라인 관람(CGV, 메가박스, 롯데시네마) 감소 및 온라인 관람(넷플릭스) 증가

코로나 이전

코로나 이후

2) SNS 워드클라우드

코로나 이후 증가 키워드

코로나 이후 감소 키워드

SNS 워드클라우드 : 여행사

해외여행(해외여행, 유럽여행) 감소 및 국내여행(한국여행, 국내여행, 한국관광) 증가

코로나 이전

코로나 이후

2) SNS 워드클라우드

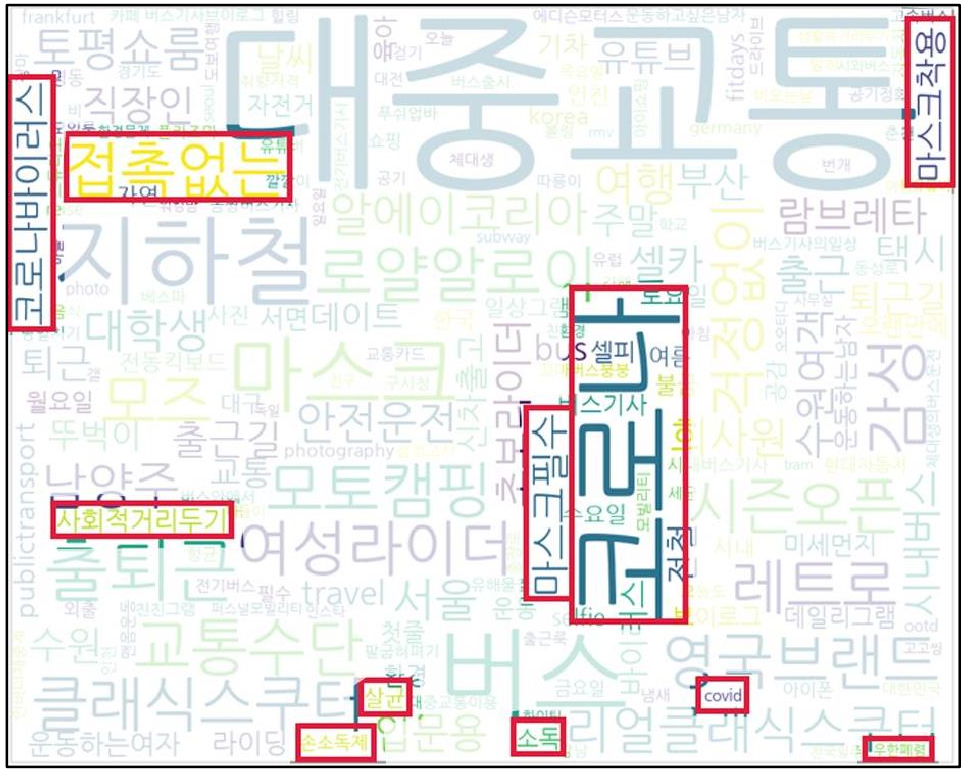
코로나 이후 증가 키워드

코로나 이후 감소 키워드

SNS 워드클라우드: 대중교통

코로나 관련 키워드 등장(코로나, 코로나 바이러스, 마스크, 접촉없는, 소독 등)

코로나 이전



코로나 이후

2) SNS 워드클라우드

SNS 워드 클라우드의 결과를 종합한 결과

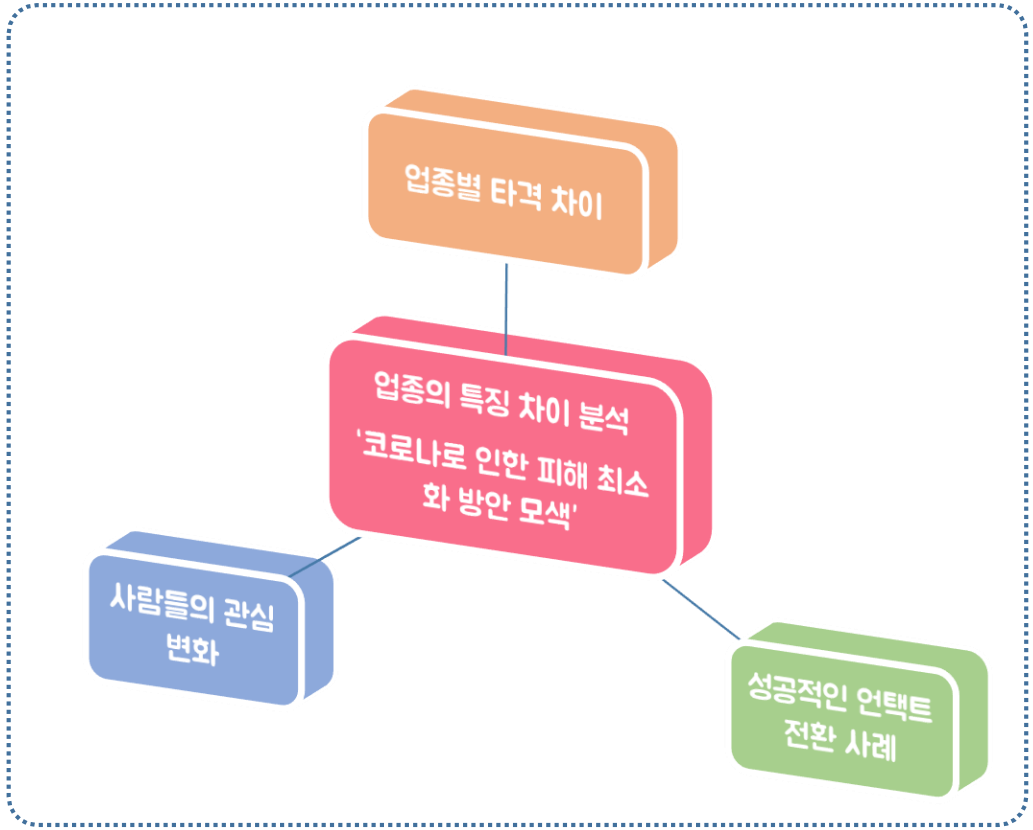
영화
오프라인 관람 (cgv,메가박스,롯데시네마) → 온라인 관람 (넷플릭스)

여행사
해외여행 (해외여행, 유럽여행) → 국내여행 (한국여행, 국내여행, 한국관광)

대중교통
위·생방·역비접촉 (마스크, 접촉없는, 살균, 소독, 사회적 거리두기 등)

대중들의 코로나에 의한 인식· 관심 변화

3) 종합 주제 선정



1) 전체적인 프로세스 흐름도

신한카드 데이터
기간 : 2017년 1월~ 2021년 6월
개수 : 약 4200만 개
출처 : 신한카드 데이터 내 국인 지출 데이터만 사용

설문조사 데이터
기간 : 2021년 2월
대상 : 성인남녀 100명
출처 : 직접작성한 설문 업종별 특징에 대한 설문

SNS 데이터
기간 : 2017년 1월~ 2021년 6월
출처 : 인스타그램 웹 크롤링 인스타그램의 키워드 별 게시글 수

확진자 데이터
기간 : 2020년 1월~ 2021년 6월
출처 : 질병관리본부
월, 지역별 코로나19 확진자 데이터

데이터 전처리

- 신한카드 데이터는 모든 데이터 전처리에 포함
- 악기연주, 스포츠경기관람, 스키 제외
- 광역시, 경기, 서울, 제주도 지역만 사용

업종별 특징

데이터분석에 효과적인 형태로 수치화

사용건수 기울기 변화율

코로나19 심각도 별 사용건수 증감율

월별 확진자수, 사용건수

모델을 통한 검증

분산분석
코로나19 피해/회복 정도 분류그룹별로 차이가 있는 특징 변수 추출

랜덤포레스트 (+ SVM)
코로나19 피해/회복의 중요특징 변수 확인

회귀분석
코로나19 심각도 가중치 반영 특징변수의 유의성 검증
특징변수의 회귀계수확인

LSTM (시계열분석)
코로나19 확진자 증감에 따른 업종별 피해/회복 추세 예측

2) 사용 데이터 소개

설문 데이터

변수명	변수 설명
q01	실내/야외
q02	같이하는 사람 수
q03	의생활
q04	식생활
q05	주생활
q06	혼자가능 여부
q07	같이가능 여부
q08	시설의 접근성
q09	대면/비대면
q10	활동에 걸리는 시간
q11	진입장벽
q12	주기성
q13	준비하는 정도
q14	인구밀집도
q15	활동성

Fig. 설문조사

- q08: 접근성이 높은 정도를 0~1로 표현
- q11: 숙련도가 필요한 정도를 0~1로 표현
- q12: 한 주간 활동 횟수
- q13: 활동하기 위해 준비가 필요한 정도를 0~1로 표현
- q14: 활동하는 동안 인구 밀집도를 0~1로 표현
- q15: 신체를 활발히 움직이는 정도를 0~1로 표현

특징 데이터

변수명	변수 설명
q16	가격대
q17	문화
q18	여행
q19	스포츠
q20	취미오락
q21	이동거리
q22	관심도
q23	20대
q24	30대
q25	40대
q26	50대
q27	60대
q28	남녀비율
q29	시간대 07시~19시
q30	시간대 19시~02시
q31	시간대 02시~07시

Fig. 신한카드, SNS 데이터

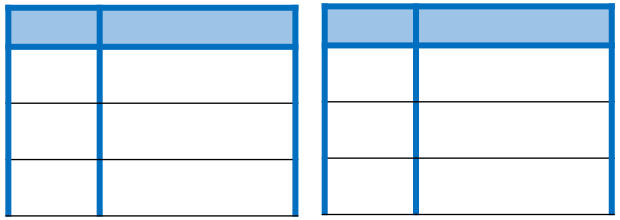
- q16: (전체 사용액수)/(전체 사용건수)
- q17~q20: 해당 업종의 대분류(신한카드 데이터 기준)
- q21: 거주지와 카드 이용 가맹점 주소 간의 거리
- q22: SNS 게시글 개수
- q23~q27: (해당 나이의 사용건수)/(전체 사용건수)
- q28: (남성의 사용건수)/(전체 사용 건수)
- q29~q31: (해당 시간대 별 사용건수)/(전체 사용 건수)

3) 데이터 전처리

사용건수 데이터

LSTM에 사용

신한카드 데이터 | 확진자 데이터



join

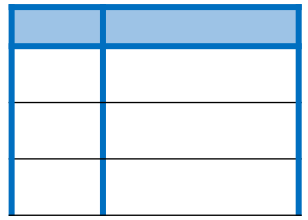
변수명	변수 설명
gb2	가맹정 업종 소분류
ta_ym	이용년월
cnt	이용건수의 합
cvd	전국확진자 수

사용건수증감률, 월별 코로나 심각도 데이터

회귀분석에 사용

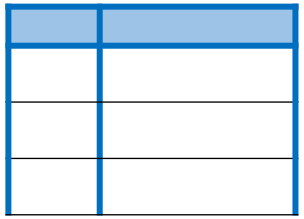
변수명	변수 설명
q01	실내/ 야외
q02	같이하는 사람 수
q03	의생활
…	…
q30	시간대 19시~02시
q31	시간대 02시~07시
varcnt	사용건수의 증감률
severity	코로나 심각도

function

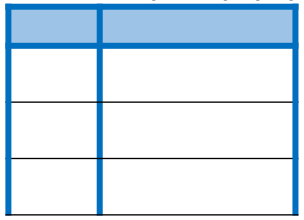


join

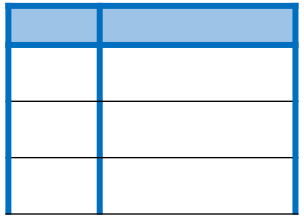
신한카드 데이터



업종별 특징 데이터



확진자 데이터



- varcnt : (21년의 해당월 사용건수)/ (17,18,19년의 해당월 사용건수)
- severity : (월지역확진수)/ (전국확진자수)+(월전국확진자수)/ (최고확진자수)
- severity * (월지역사용건수)/ (월전국사용건수) 만큼 반영가중치 조정

3) 데이터 전처리

사용건수기울기 변화율 데이터

변수명	변수 설명
v2	지역
q01	실내/ 야외
q02	같이하는 사람 수
q03	의생활
…	…
q30	시간대 19시~02시
q31	시간대 02시~07시
slop13	1~3월 사용건수의 기울기 변화율
slop46	4~6월 사용건수의 기울기 변화율

Group by

join

신한카드 데이터

업종별 특징 데이터

- slop## :(17,18,19년의 해당기간 사용건수기울기 평균)- (20년의 해당기 간 사용건수 기울기)
- 업종들의 사용건수 그래프를 보았을 때 대체적으로 1~3월이 감소, 4~6월이 증가
- -> 1~3월을 피해시기, 4~6월을 회복시기로 간주하여 1~3월과 4~6월 기간을 분리

"분류분석(랜덤포레스트,SVM)과 분산 분석을 위해 A, B, C로 labeling

- 1~3월 labeling 기준
  기울기 변화 -56% 이하 : C 기울기 변화 -56%~+-36% : B 기울기 변화 -36% 이상 : A
- C : 75, B:75, A : 74
- 라벨 분류가 1/3씩 되도록 기준을 설정 (4~6월 labeling도 같은 방식 적용)

Example)

년	2017		2018		2019		2020	
월	1	3	1	3	1	3	1	3
사용건수	1000	800	1200	1200	1500	1350	1500	1200
기울기	-20%		0%		-10%		-20%	
평균	Mean = -10%							

1) 분산분석

업종별 특징 데이터 스키마

변수명	변수설명
q01	실내/ 야외
q02	같이하는 사람 수
q03	의생활
q04	식생활
q05	주생활
q08	시설의 접근성
q09	대면/ 비대면
q10	활동에 걸리는 시간

변수명	변수설명
q11	진입장벽
q12	주기성
q13	준비하는 정도
q14	인구 밀집도
q15	활동성
q16	가격대
q17	문화
q18	여행

변수명	변수설명
q19	스포츠
q20	취미오락
q21	이동거리
q22	관심도
q23	20대 비율
q24	30대 비율
q25	40대 비율
q26	50대 비율

변수명	변수설명
q27	60대 비율
q28	남녀 비율
q29	시간대07시~19시
q30	시간대19시~02시
q31	시간대02시~07시

(1) 1~3월 코로나19 영향에 의한 사용건수 피해의 정도로 나눈 기준(A, B, C)

문제 제기

각각의 변수(q01~q31)가 1~3월 피해의 정도로 나눈 기준(A, B, C)에 대해 차이가 있는가?

(1) 실내/야외(q01)에 대한 분석

정규성 분석

n≥ 30이지만 같은 특성이 9개씩 25개 존재하므로 정규성 분석 실시

-> Shapiro-wilk test로 정규성 분석

정규성 검정

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	label_c	통계량	자유도	유의확률	통계량	자유도	유의확률
q01	A	.178	75	.000	.812	75	.000
	B	.283	74	.000	.779	74	.000
	C	.272	75	.000	.788	75	.000

p-값 < 0.05 -> 귀무가설 기각 -> 정규성을 만족하지 않음.

위와 같은 방법으로 각 변수 q02~q31에 대한 정규성 검정 시행-> 모든 p-값이 <0.05

-> 변수 모두 Kruskal-wallis test 실시

1) 분산분석

업종별 특징 데이터 스키마

변수명	변수 설명
q01	실내/ 야외
q02	같이하는 사람 수
q03	의생활
q04	식생활
q05	주생활
q08	시설의 접근성
q09	대면/ 비대면
q10	활동에 걸리는 시간

변수명	변수 설명
q11	진입장벽
q12	주기성
q13	준비하는 정도
q14	인구 밀집도
q15	활동성
q16	가격대
q17	문화
q18	여행

변수명	변수 설명
q19	스포츠
q20	취미오락
q21	이동거리
q22	관심도
q23	20대 비율
q24	30대 비율
q25	40대 비율
q26	50대 비율

변수명	변수 설명
q27	60대 비율
q28	남녀 비율
q29	시간대07시~19시
q30	시간대19시~02시
q31	시간대02시~07시

(1) 1~3월 코로나19 영향에 의한 사용건수 피해의 정도로 나눈 기준(A, B, C)

분산 분석

-가설

H_0 : $\mu_1 = \mu_2 = \mu_3$ (A, B,C그룹에 대한 업종의 실내/야외 사용 비율의 평균은 같다.)

H_1 : $\mu_1 \neq \mu_2 \neq \mu_3$ (A, B,C그룹에 대한 업종의 실내/야외 사용 비율의 평균에 차이가 존재한다.)

가설 검정 요약

	귀무가설	검정	Sig.	의사결정
1	q01의 분포는 label_c의 범주에서 동일합니다.	독립표본 Kruskal-Wallis 검정	.729	귀무가설을 유지합니다.

p-값 > 0.05이므로 귀무가설 채택

-> A, B,C그룹에 대한 업종의 시설의수의 평균에 차이가 존재하므로 시설의수에 대한 척도는 유의하지 않다고 할 수 있다.

위와 같은 방법으로 각 변수q02~q31에 대한 유의성 검정 시행

H_0 기각: q02, q04, q12, q14, q18, q20, q21, q22, q27, q28, q29, q30, q31

A, B,C그룹에 대한 변수(같이하는 사람 수, 주생활 여부, 주기성, 인구밀집도, 여행 여부, 취미오락 여부, 이동거리, 관심도, 60대 비율, 시간대07~19시, 시간대19~02시, 시간대02~07시)의 평균은 다르다고 할 수 있다!

1) 분산분석

업종별 특징 데이터 스키마

변수명	변수 설명
q01	실내/ 야외
q02	같이하는 사람 수
q03	의생활
q04	식생활
q05	주생활
q08	시설의 접근성
q09	대면/ 비대면
q10	활동에 걸리는 시간

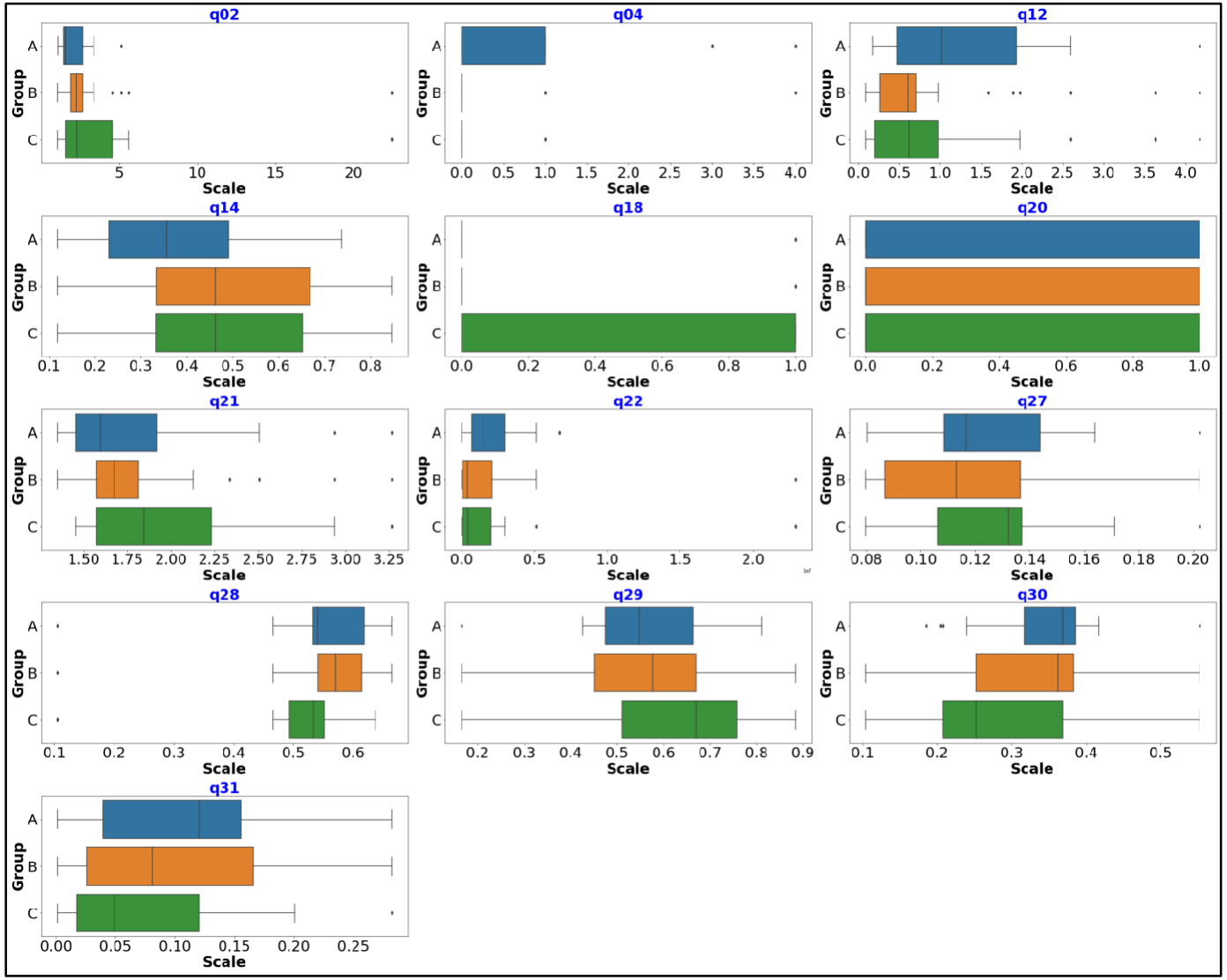
변수명	변수 설명
q11	진입장벽
q12	주기성
q13	준비하는 정도
q14	인구 밀집도
q15	활동성
q16	가격대
q17	문화
q18	여행

변수명	변수 설명
q19	스포츠
q20	취미오락
q21	이동거리
q22	관심도
q23	20대 비율
q24	30대 비율
q25	40대 비율
q26	50대 비율

변수명	변수 설명
q27	60대 비율
q28	남녀 비율
q29	시간대 07시~19시
q30	시간대 19시~02시
q31	시간대 02시~07시

피해 Boxplot

Boxplot을 통해 피해에 대한 각 변수의 차이를 알아봄



1) 분산분석

업종별 특징 데이터 스키마

변수명	변수 설명
q01	실내/ 야외
q02	같이하는 사람 수
q03	의생활
q04	식생활
q05	주생활
q08	시설의 접근성
q09	대면/ 비대면
q10	활동에 걸리는 시간

변수명	변수 설명
q11	진입장벽
q12	주기성
q13	준비하는 정도
q14	인구 밀집도
q15	활동성
q16	가격대
q17	문화
q18	여행

변수명	변수 설명
q19	스포츠
q20	취미오락
q21	이동거리
q22	관심도
q23	20대 비율
q24	30대 비율
q25	40대 비율
q26	50대 비율

변수명	변수 설명
q27	60대 비율
q28	남녀 비율
q29	시간대07시~19시
q30	시간대19시~02시
q31	시간대02시~07시

(2) 4~6 월 코로나19 영향에 의한 사용건수 회복의 정도로 나눈 기준(A, B, C)

문제 제기

각각의 변수(q01~q31)가 4~6월 회복의 정도로 나눈 기준(A, B, C)에 대해 차이가 있는가?

(1) 실내/야외(q01)에 대한 분석

정규성 분석

n≥ 30이지만 같은 특성이 9개씩 25개 존재하므로 정규성분석 실시
-> Shapiro-wilk test로 정규성분석

정규성 검정

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	label_c	통계량	자유도	유의확률	통계량	자유도	유의확률
q01	A	.303	74	.000	.788	74	.000
	B	.211	76	.000	.783	76	.000
	C	.187	74	.000	.828	74	.000

p-값 < 0.05 -> 귀무가설 기각 -> 정규성을 만족하지 않음.
위와 같은 방법으로 각 변수q02~q31에 대한 정규성검정 시행-> 모든 p-값이 <0.05

-> 변수 모두 Kruskal-wallis test 실시

1) 분산분석

업종별 특징 데이터 스키마

변수명	변수 설명
q01	실내/ 야외
q02	같이하는 사람 수
q03	의생활
q04	식생활
q05	주생활
q08	시설의 접근성
q09	대면/ 비대면
q10	활동에 걸리는 시간

변수명	변수 설명
q11	진입장벽
q12	주기성
q13	준비하는 정도
q14	인구 밀집도
q15	활동성
q16	가격대
q17	문화
q18	여행

변수명	변수 설명
q19	스포츠
q20	취미오락
q21	이동거리
q22	관심도
q23	20대 비율
q24	30대 비율
q25	40대 비율
q26	50대 비율

변수명	변수 설명
q27	60대 비율
q28	남녀 비율
q29	시간대07시~19시
q30	시간대19시~02시
q31	시간대02시~07시

(2) 4~6 월 코로나19 영향에 의한 사용건수 회복의 정도로 나눈 기준(A, B, C)

분산 분석

-가설

H_0 : $\mu_1 = \mu_2 = \mu_3$ (A, B,C그룹에 대한 업종의 실내/야외 사용 비율의 평균은 같다.)

H_1: $\mu_1 \neq \mu_2 \neq \mu_3$ (A, B,C그룹에 대한 업종의 실내/야외 사용 비율의 평균에 차이가 존재한다.

가설 검정 요약

	귀무가설	검정	Sig.	의사결정
1	q01의 분포는 label_c의 범주에서 동일합니다.	독립표본 Kruskal-Wallis 검정	.239	귀무가설을 유지합니다.

p-값 > 0.05이므로 귀무가설 기각

-> A, B,C그룹에 대한 업종의 시설의수의 평균에 차이가 존재하므로 시설의수에 대한 척도는 유의하지 않다고 할 수 있다..

위와 같은 방법으로 각 변수q02~q31에 대한 유의성검정 시행

H_0 기각: q14, q15, q18, q20, q21, q24, q28

A, B,C그룹에 대한 변수(인구밀집도, 활동성, 여행여부, 취미오락 여부, 이동거리 여부, 30대 비율, 남녀비율) 의 평균은 다르다고 할 수 있다!

1) 분산분석

업종별 특징 데이터 스키마

변수명	변수 설명
q01	실내/ 야외
q02	같이하는 사람 수
q03	의생활
q04	식생활
q05	주생활
q08	시설의 접근성
q09	대면/ 비대면
q10	활동에 걸리는 시간

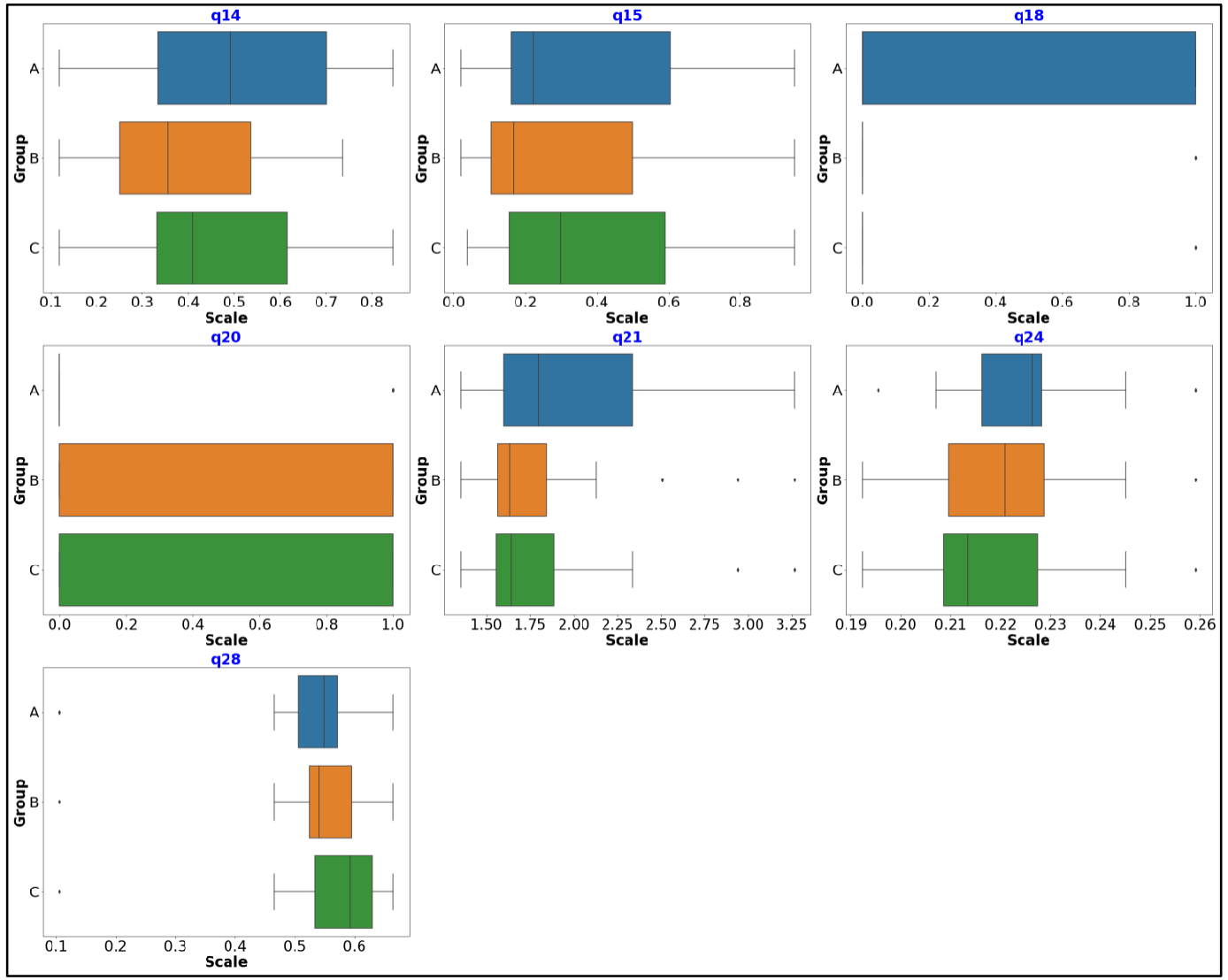
변수명	변수 설명
q11	진입장벽
q12	주기성
q13	준비하는 정도
q14	인구 밀집도
q15	활동성
q16	가격대
q17	문화
q18	여행

변수명	변수 설명
q19	스포츠
q20	취미오락
q21	이동거리
q22	관심도
q23	20대 비율
q24	30대 비율
q25	40대 비율
q26	50대 비율

변수명	변수 설명
q27	60대 비율
q28	남녀 비율
q29	시간대07시~19시
q30	시간대19시~02시
q31	시간대02시~07시

회복 Boxplot

Boxplot을 통해 회복에 대한 각 변수의 차이를 알아봄



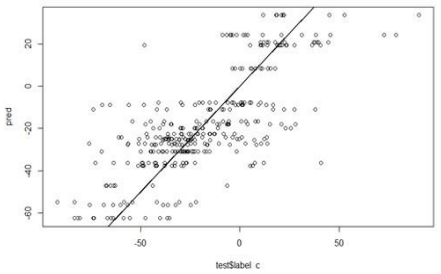
2) 회귀

변수별 회귀 계수 및 p-value

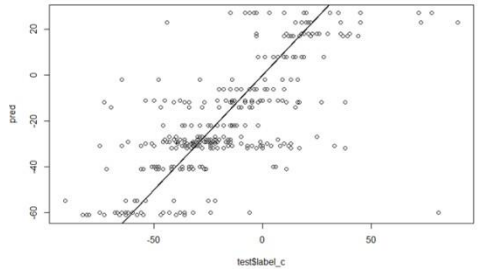
변수	회귀계수	p-value
실내/야외	-0.122106	0.169162
같이하는사람 수	-0.240930	5.93e-09 ***
의	0.372195	2e-16 ***
식	-0.220221	8.67e-08 ***
주	-0.207459	0.002597 **
시설의 수	0.609119	2e-16 ***
대면/비대면	0.308263	7.93e-11 ***
활동시간	0.866895	2.16e-09 ***
진입장벽	0.332988	2e-16 ***
주기성	-0.302052	1.66e-14 ***
준비하는정도	-0.016853	0.855963 ***
인구밀집도	-0.364498	2e-16 ***
활동성	0.581656	2e-16 ***
가격대	-0.488624	2.85e-05 ***
문화	0.186900	1.77e-08 ***
여행	0.394493	0.000104 ***
스포츠	-0.143706	4.53e-06 ***
이동거리	-1.076550	2.24e-10 ***
관심도	-0.669892	2e-16 ***
20대	0.004019	0.969578
30대	-0.474641	4.20e-11 ***
40대	0.136165	0.385867
50대	-0.646518	0.021947 *
60대	0.770648	3.95e-05 ***

- 사용 건수 증감률, 월별 코로나 확진자 수 데이터 사용
- 코로나 심각도에 따라 반영 가중치 조정
- 꽤 많은 변수의 p-value가 상당히 낮은 값을 보임을 통해, 변수들의 설명력을 확인.
  (회귀 계수의 절대값이 큰 변수들은 색으로 표시)
- R^2 = 0.6172로 유의미한 회귀모델임을 알 수 있음

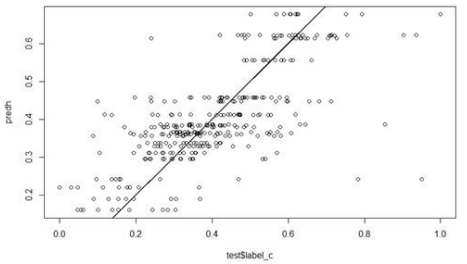
모델 검증



다중회귀 R^2 = 0.6172



SVM R^2 = 0.6181



Neural Network R^2 = 0.6226

SVM, 신경망 모델으로도 분석을 해본 결과 R^2이 다중회귀모델과 비슷함 -> 다중회귀모델의 신뢰성 검증

2) 회귀 분석

데이터 스키마 정의

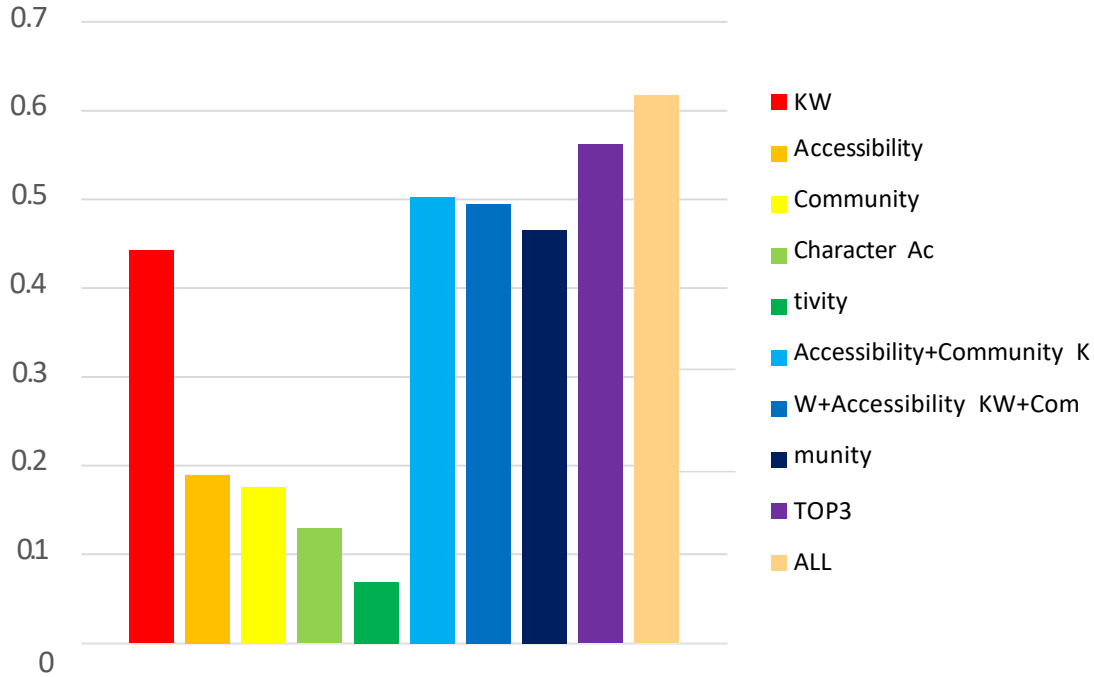
변수명	변수 설명
q01	실내/ 야외
q02	같이하는 사람 수
q03	의생활
q04	식생활
q05	주생활
q08	시설의 접근성
q09	대면/ 비대면
q10	활동에 걸리는 시간

변수명	변수 설명
q11	진입장벽
q12	주기성
q13	준비하는 정도
q14	인구 밀집도
q15	활동성
q16	가격대
q17	문화
q18	여행

변수명	변수 설명
q19	스포츠
q20	취미오락
q21	이동거리
q22	관심도
q23	20대 비율
q24	30대 비율
q25	40대 비율
q26	50대 비율

변수명	변수 설명
q27	60대 비율
q28	남녀 비율
q29	시간대07시~19시
q30	시간대19시~02시
q31	시간대02시~07시

변수 그룹별 R^2 막대 그래프

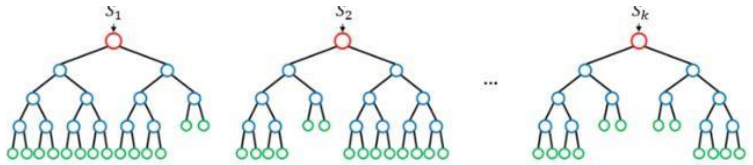


- KW: 분산분석(Kruskal-Wallis)에서 유의미할 가능성이 높은 변수 집합
- Accessibility: 접근성 변수 집합
- Community: 타인 접촉 관련 변수 집합
- Character: 의/식/주 및 대분류 변수 집합
- Activity: 활동에 직접 연관된 변수 집합

- KW > Accessibility > Community > Character > Activity
- Accessibility+Community > KW+Accessibility > KW+Community
- KW+Acessibility+Community(TOP3)는 ALL과 R^2이 비슷함을 확인

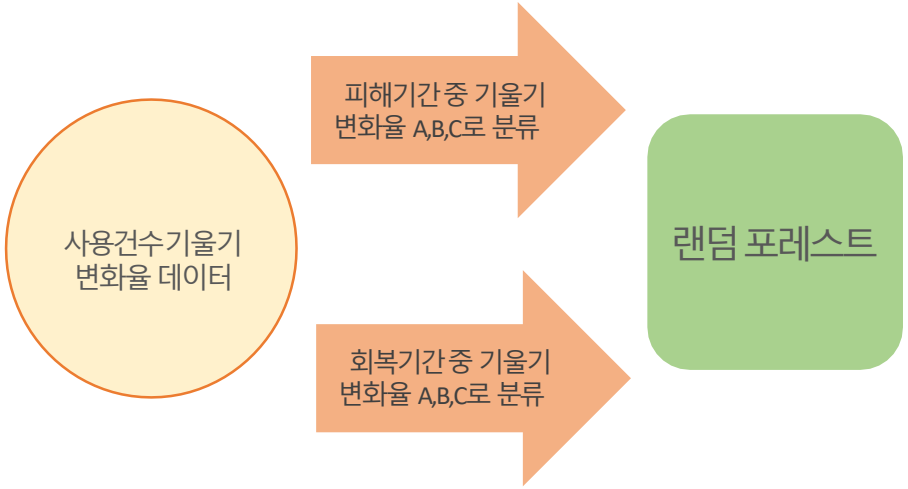
3) 랜덤 포레스트

랜덤 포레스트란?



- 다수의 의사결정트리의 앙상블을 통해 얻은 예측결과

랜덤 포레스트 사용 목적

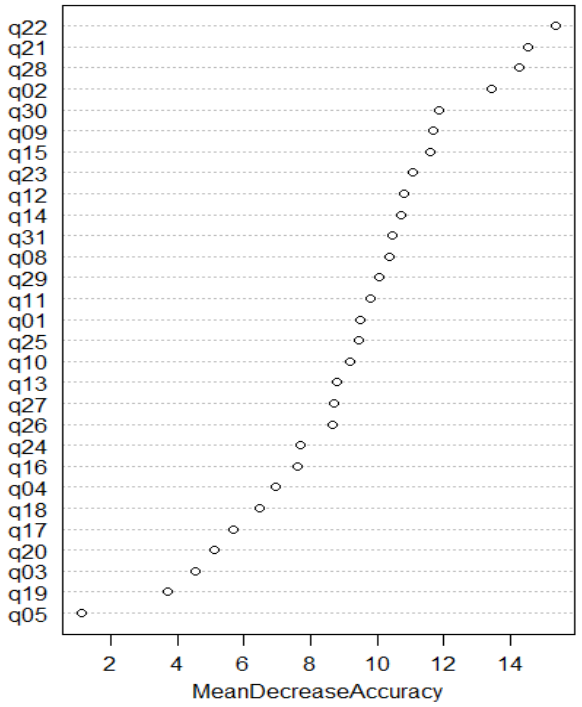


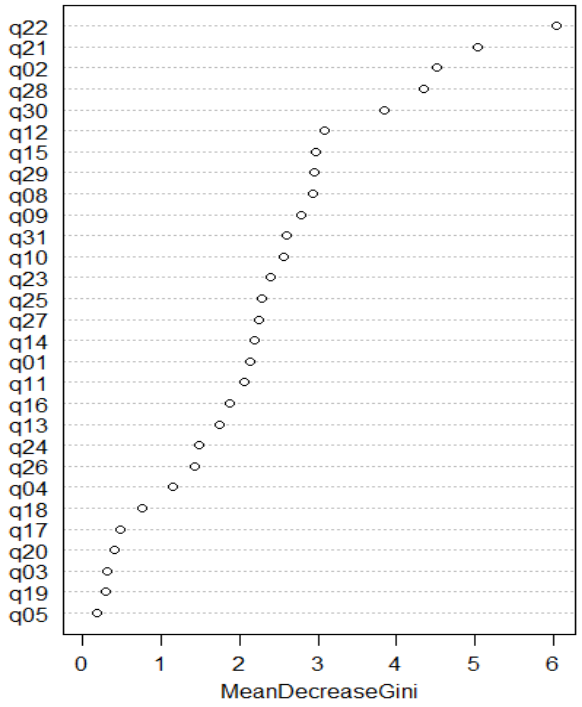
- 랜덤 포레스트의 정확도에 영향력이 큰 변수 확인

결과 해석방법

ntree = 500

mtry = $floor(\sqrt{(\text{설명변수의 수})})$





- 관련 변수 값을 다른 값으로 대체하였을 때 정확도가 감소하는 정도(Mean Decrease in Accuracy, MDA)
- 관련 변수 값을 두개로 나누었을 때 순수도(purity) 값이 감소하는 정도(GINI) (Mean Decrease Gini, MDG)

3) 랜덤 포레스트

업종별 특징 데이터 스키마

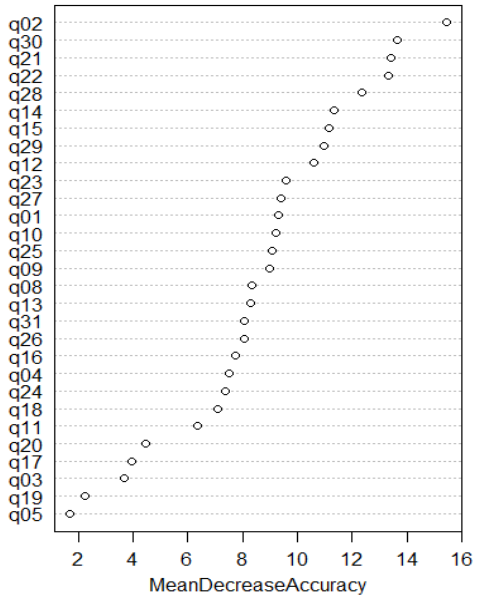
변수명	변수 설명
q01	실내/ 야외
q02	같이하는 사람 수
q03	의생활
Q04	식생활
Q05	주생활
Q08	시설의 접근성
Q09	대면/ 비대면
Q10	활동에 걸리는 시간

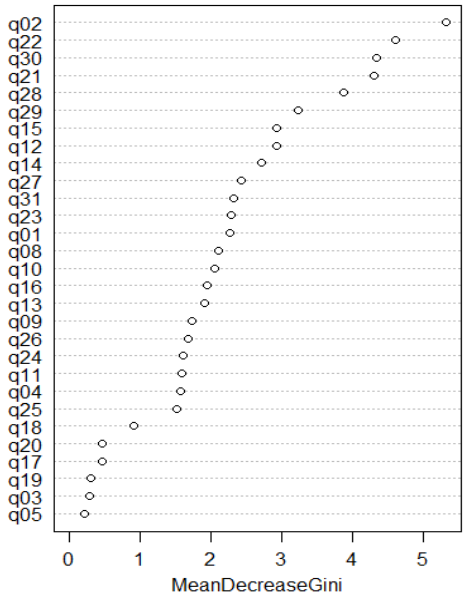
변수명	변수 설명
q11	진입장벽
q12	주기성
q13	준비하는 정도
q14	인구 밀집도
q15	활동성
q16	가격대
q17	문화
q18	여행

변수명	변수 설명
q19	스포츠
q20	취미오락
q21	이동거리
q22	관심도
q23	20대 비율
q24	30대 비율
q25	40대 비율
q26	50대 비율

변수명	변수 설명
q27	60대 비율
q28	남녀 비율
q29	시간대07시~19시
q30	시간대19시~02시
q31	시간대02시~07시

랜덤 포레스트 결과 - 피해 기간(1~3월)





OOBestimate of error rate: 37.5%

Confusion matrix:

	A	B	C	Recall
A	56	12	7	0.746
B	17	35	22	0.473
C	3	23	49	0.653
Prec.	0.736	0.500	0.628	

Macro F1-Score :0.622

- 예측 성공률이 62.5% 정도로 나온 것을 확인
- A를 C로, C를 A로 예측한 경우는 거의 없음
- q02, q30, q21, q22, q28, q14, q15 중요하게 작용
- 회복 기간의 랜덤 포레스트 분석보다 에러율 낮음

3) 랜덤 포레스트

업종별 특징 데이터 스키마

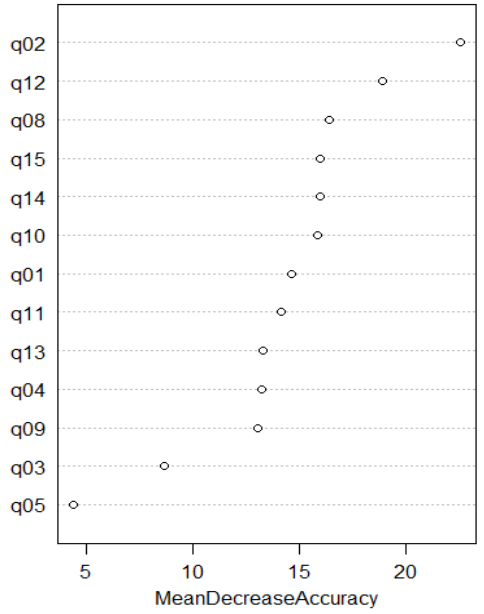
변수명	변수 설명
q01	실내/ 야외
q02	같이하는 사람 수
q03	의생활
q04	식생활
q05	주생활
q08	시설의 접근성
q09	대면/ 비대면
q10	활동에 걸리는 시간

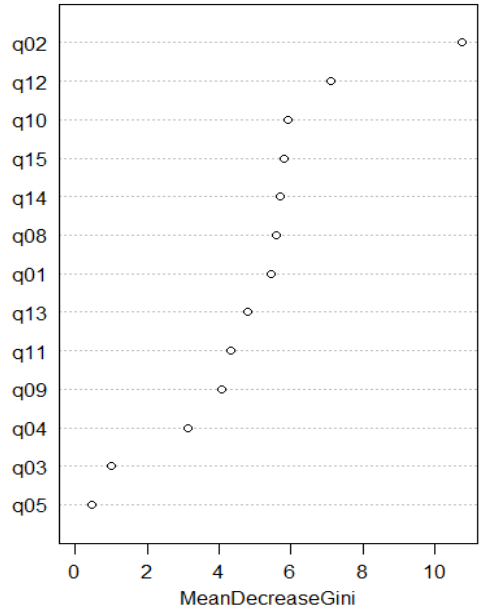
변수명	변수 설명
q11	진입장벽
q12	주기성
q13	준비하는 정도
q14	인구 밀집도
q15	활동성
q16	가격대
q17	문화
q18	여행

변수명	변수 설명
q19	스포츠
q20	취미오락
q21	이동거리
q22	관심도
q23	20대 비율
q24	30대 비율
q25	40대 비율
q26	50대 비율

변수명	변수 설명
q27	60대 비율
q28	남녀 비율
q29	시간대07시~19시
q30	시간대19시~02시
q31	시간대02시~07시

랜덤 포레스트 결과(q01~q15) - 피해 기간(1~3월)





OOBestimate of error rate: 38.39%

Confusion matrix:

	A	B	C	Recall
A	56	13	6	0.746
B	17	36	21	0.473
C	4	25	46	0.653
Prec.	0.727	0.493	0.630	

Macro F1-Score : 0.615

- 설문특징데이터(q01~q15) 만 가지고 랜덤포레스트 분석을한 경우에도 F-1 Score가 크게차이 나지 않음.
- 설문 결과가 유의미함을 알 수 있음.

3) 랜덤 포레스트

업종별 특징 데이터 스키마

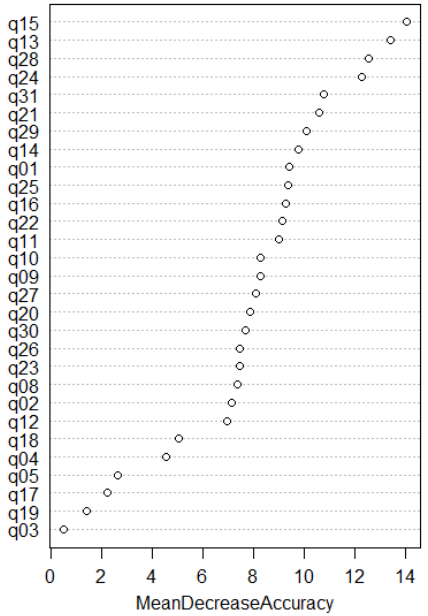
변수명	변수 설명
q01	실내/ 야외
q02	같이하는 사람 수
q03	의생활
q04	식생활
q05	주생활
q08	시설의 접근성
q09	대면/ 비대면
q10	활동에 걸리는 시간

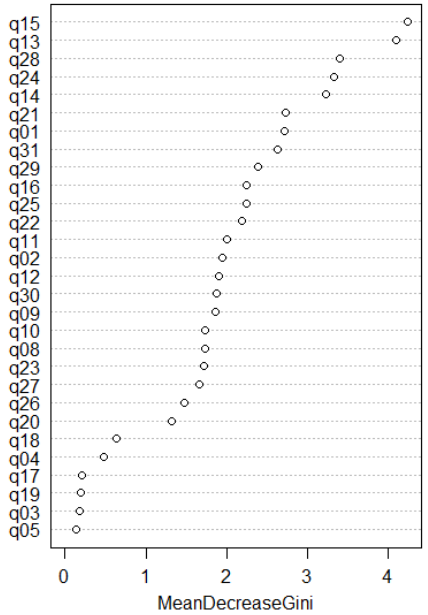
변수명	변수 설명
q11	진입장벽
q12	주기성
q13	준비하는 정도
q14	인구 밀집도
q15	활동성
q16	가격대
q17	문화
q18	여행

변수명	변수 설명
q19	스포츠
q20	취미오락
q21	이동거리
q22	관심도
q23	20대 비율
q24	30대 비율
q25	40대 비율
q26	50대 비율

변수명	변수 설명
q27	60대 비율
q28	남녀 비율
q29	시간대 07시~19시
q30	시간대 19시~02시
q31	시간대 02시~07시

랜덤 포레스트 결과 - 회복 기간(4~6월)





OOBestimate of error rate: 39.73%

Confusion matrix:

	A	B	C	Recall
A	49	18	7	0.662
B	15	45	16	0.592
C	11	22	41	0.554
Prec.	0.653	0.529	0.640	

- 예측 성공률 60.27%
- q15, q13, q28, q24, q31, q21, q29, q14가 중요하게 작용 (피해 기간의 중요 변수들과는 약간 차이가 존재)
- 피해 기간의 랜덤 포레스트 분석보다는 높은 에러율

Macro F1-Score :0.622

3) 랜덤 포레스트

업종별 특징 데이터 스키마

변수명	변수 설명
q01	실내/ 야외
q02	같이하는 사람 수
q03	의생활
q04	식생활
q05	주생활
q08	시설의 접근성
q09	대면/ 비대면
q10	활동에 걸리는 시간

변수명	변수 설명
q11	진입장벽
q12	주기성
q13	준비하는 정도
q14	인구 밀집도
q15	활동성
q16	가격대
q17	문화
q18	여행

변수명	변수 설명
q19	스포츠
q20	취미오락
q21	이동거리
q22	관심도
q23	20대 비율
q24	30대 비율
q25	40대 비율
q26	50대 비율

변수명	변수 설명
q27	60대 비율
q28	남녀 비율
q29	시간대07시~19시
q30	시간대19시~02시
q31	시간대02시~07시

SVM 결과를 통한 랜덤 포레스트 모델 검증

피해기간(1~3월)

OOBestimate of error rate: 32.20%

Confusion matrix:

	A	B	C	Recall
A	56	13	6	0.746
B	13	46	15	0.621
C	3	22	50	0.666
Prec.	0.777	0.567	0.704	

Macro F1-Score : 0.679

회복기간(4~6월)

OOBestimate of error rate: 35.77%

Confusion matrix:

	A	B	C	Recall
A	49	18	7	0.662
B	11	52	13	0.684
C	11	20	43	0.581
Prec.	0.690	0.577	0.682	

Macro F1-Score : 0.643

- SVM 모델으로 기울기 변화율 데이터 분류 결과
- 랜덤 포레스트와 비슷한 양상 -> 랜덤 포레스트 모델의 신뢰성 확인

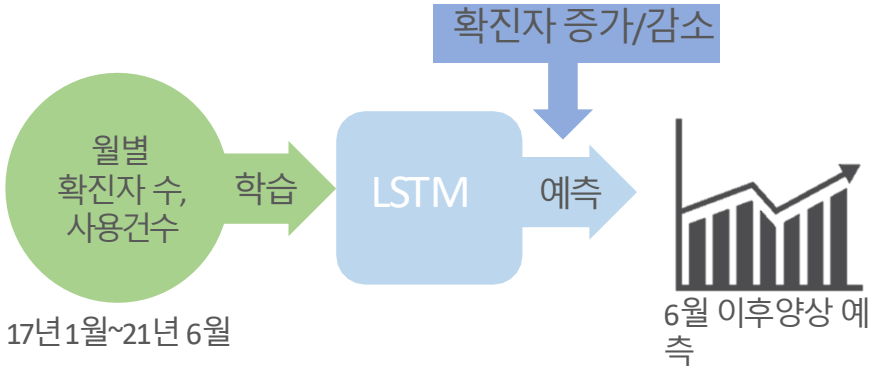
4) LSTM (시계열 분석)

LSTM (시계열 분석)이란?

- Long Short-Term Memory
- 시계열 데이터 학습에 최적화 된 딥러닝 모델

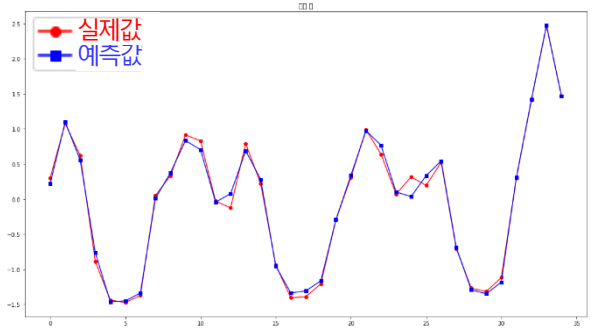
LSTM (시계열 분석) 사용 목적

- 2017년 1월~ 2020년 6월의 월별 확진자 수와 사용건수 를 학습하여 그 이후의 확진자 수에 대한 사용건수를 예측
- 2020년 7월 이후 코로나19 확진자 수가 증가할 때 업종의 피해정도를 예상
- 2020년 7월 이후 코로나19 확진자 수가 감소할 때 업종의 회복정도를 예상

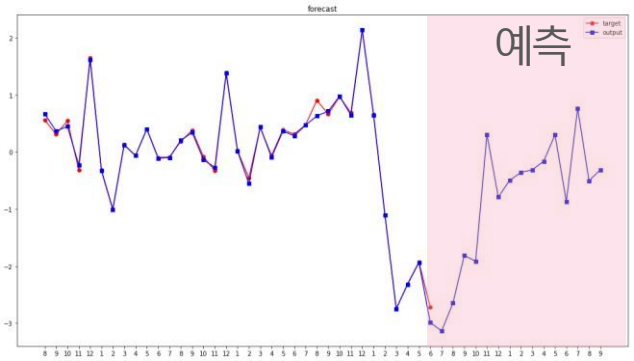


LSTM (시계열 분석) 모델 구축 및 활용

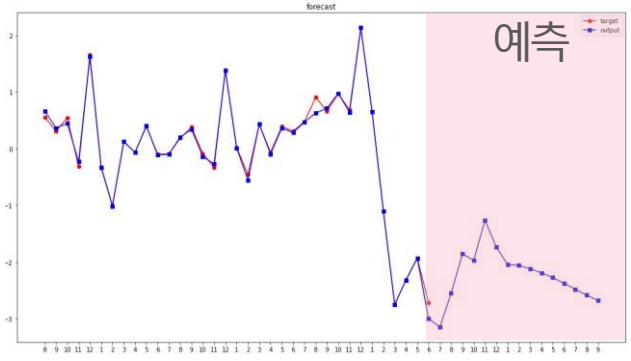
- 전국 확진자 데이터만 사용



- 2020년 6월까지의 시계열 예측 값과 실제 값 비교
- 신뢰할 수 있는 모델 성능 확인
- 2020년 7월 이후 확진자 수를 변화 시켜가며 예측결과 확인



<확진자 수가 감소 하는 경우 예측결과>

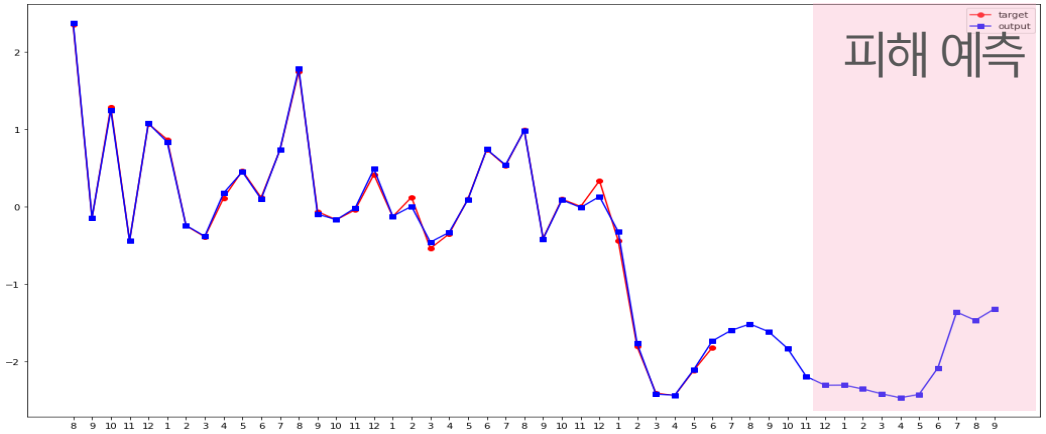


<확진자 수가 증가 하는 경우 예측결과>

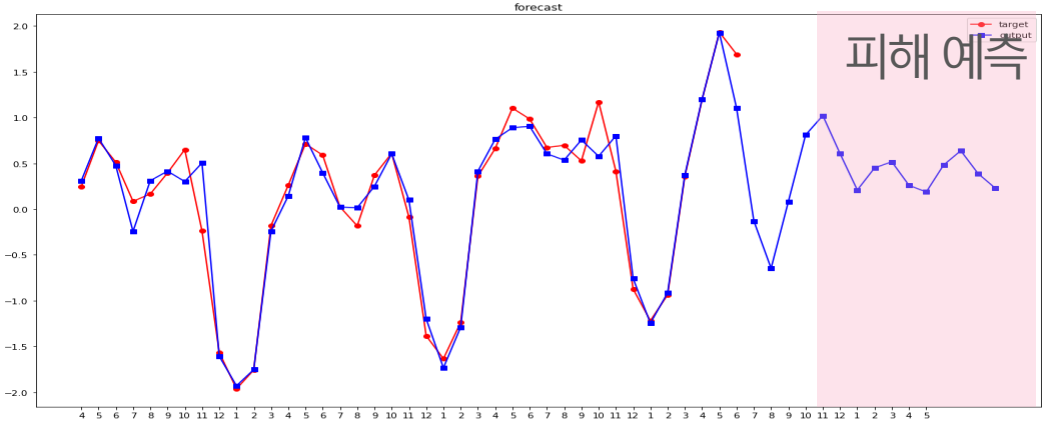
4) LSTM (시계열 분석)

피해 예측(확진자 수 증가)

- 피해가 클것으로 예측 되는 업종 예 (공연관람)

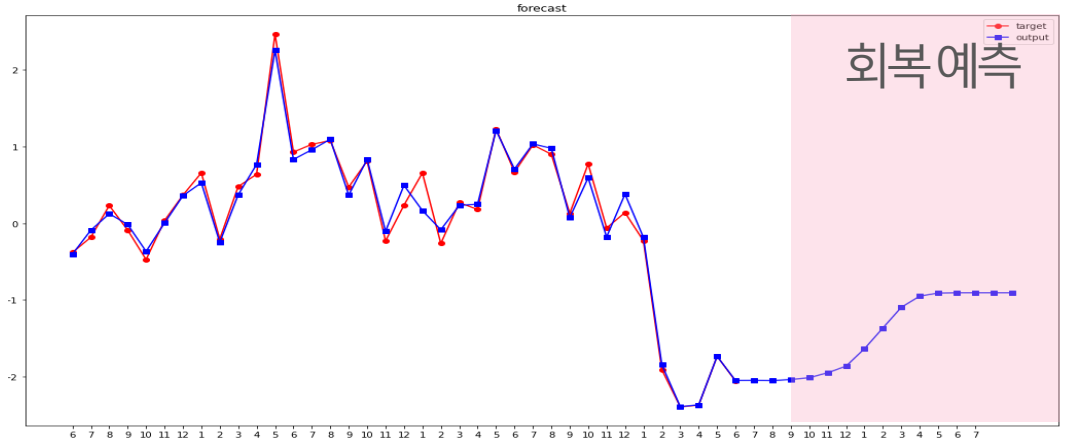


- 피해가 작을 것으로 예측 되는 업종 예 (골프)

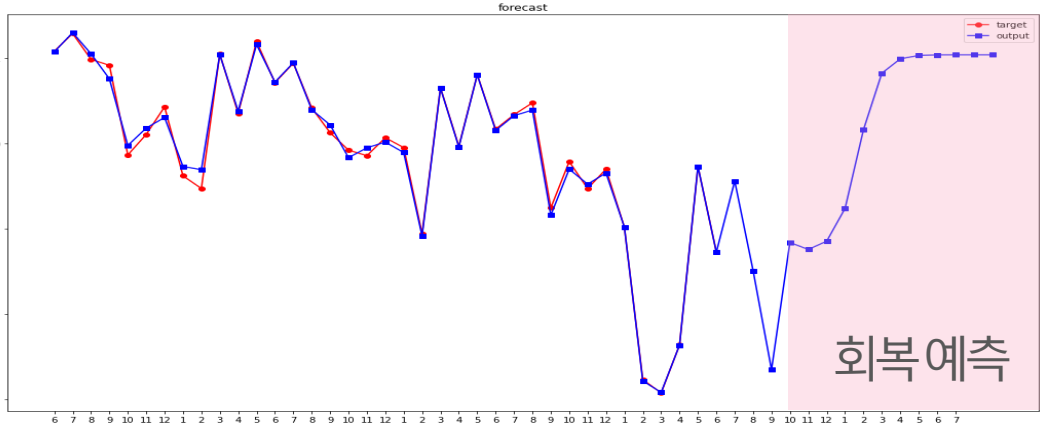


회복 예측(확진자 수 감소)

- 회복이 더딜것으로 예측 되는 업종 예 (여행사)



- 회복이 빠를 것으로 예측 되는 업종 예 (미용)



피해가 큰 업종의 예

공연관람

피해 감소에 유의미한 특징 변수

- 이동거리 (회귀계수-)
- 인구밀집도 (회귀계수-)
- 같이하는 사람 수 (회귀계수-)

소규모 지역 공연

회복이 더딘 업종의 예

여행사

회복 증대에 유의미한 특징 변수

- 활동성 (회귀계수+)
- 인구밀집도 (회귀계수-)
- 이동거리 (회귀계수-)
- 준비정도 (회귀계수-)

자전거여행패키지