데이터 통합 개인 2차

주문 수 분석 모델을 통한 치킨집 운영 방안 기준 제시

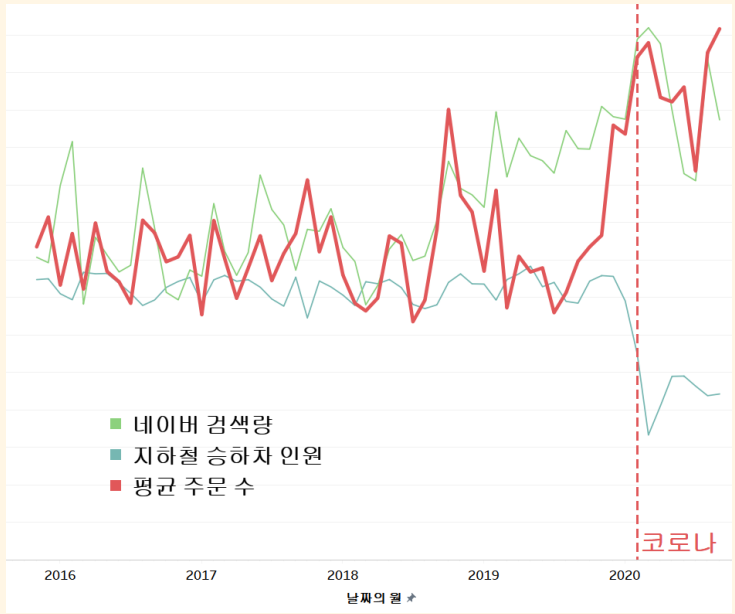
2014131026 수학과 한인욱

Contents

1 Research Question

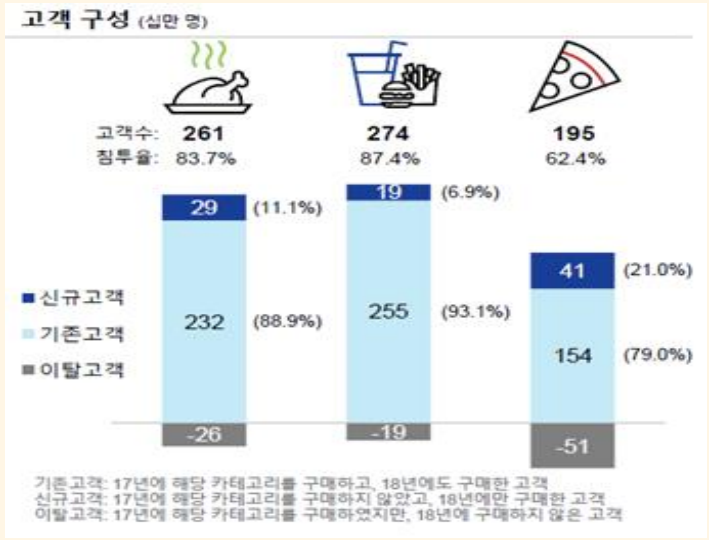
수집 데이터

- 날짜
- 강수량mm
- 강우여부
- 공휴일
- 구글검색량
- 네이버검색량
- 도마1동인구
- 도마2동인구
- 방학여부
- 버스승하차인원
- 법동인구
- 복날
- 사망자
- 영업가게
- 완치자
- 요일
- 유튜브조회수
- 전국사망자
- 전국완치자
- 전국확진자
- 주말여부
- 주문수
- 지하철승하차인원
- 최고기온
- 최저기온
- 평균기온
- 확진자



(단위: %)

		자담 치킨	ㄱ 치킨	ㄴ 치킨	ㄷ 치킨	ㄹ 치킨	ㅁ 치킨	ㅂ 치킨	ㅅ 치킨	ㅇ 치킨	ㅈ 치킨
응답자 전체	선호도	51.1	68.0	57.8	61.3	59.4	48.6	39.4	47.3	45.6	36.5
해당 브랜드 경험자	재구입 의향	88.7	85.8	77.4	82.5	81.5	70.9	81.4	77.5	80.4	72.8
	추천 의향	91.9	80.5	75.5	78.5	73.5	67.8	71.7	72.7	73.6	63.2

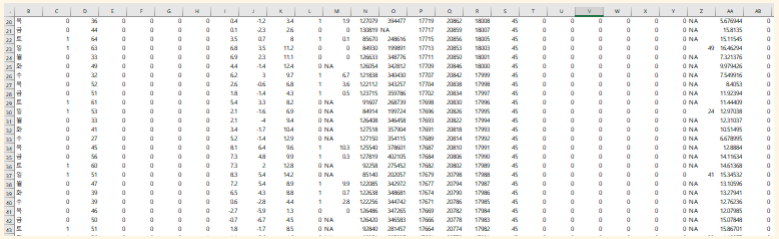


Research Question : 주문수 분석 모델을 통한 치킨집 운영 방안 기준 제시

- 같은 조원 중에서 부모님이 치킨집을 운영하고 있는 조원이 있었다. 그래서 프로젝트가 가게에 도움이 되고자 조원 모두 Research Question을 치킨집과 관련되게 잡으려 하였고, 그에 따라 병합하게 된 데이터들도 치킨집의 위치와 관련된 대전 중심으로 수집하게 되었다.
- 모으고 난 데이터에는 당연하게도 치킨집 매출을 예측 할 만한 변수가 많았다. 특히나 가운데 그림을 보면 알 수 있듯이, 네이버 검색량과, 지하철 승하차 인원이 매출과 매우 관련 있어 보였다. 즉 이 데이터를 이용해, 매출을 예측할 수 있는 모델을 만들고, 그 모델을 통해서 가게의 전반적인 운영에 도움을 주는 것이 가능해 보였다.
- 게다가 요즘 시기는 치킨집에 대해 매우 중요한 시기이다. 왜냐하면 맨 오른쪽 그림을 보면 알 수 있듯이, 치킨은 기존 고객의 비율과, 충성도가 큰 상품이다. 이는 단골 확보가 매우 중요하다는 의미이다. 그러므로 코로나 시대로 늘어난 신규 고객을 단골로 만들고, 장기적으로 매출을 늘릴 제일 중요한 시기인 것이다.
- 그러므로 이 중요한 시기에 언제 쉴지, 어떤 전략을 써야 할지 등은 매우 중요하다. 하지만 이런 판단에 대해 근거가 필요할 것이다. 이러한 근거가 될 수 있는것이 바로 통계 모델이다. (자세한 방법은 맨 뒤 10page에 수록)
- 그러므로 Research Question 을 위와 같이 잡았다. 수집한 데이터로도 충분히 예측이 가능해 보였고, 지금 시기에 제일 필요하기 때문이다.

2 데이터 선택 및 설명

조별 데이터(1698*27)



변수선택 기준

1. **지속적인 Update 가 가능한 변수**
2. **의미가 겹치지 않는 변수**
3. **코로나 경향을 올바르게 파악 가능한 변수**
4. **너무 Sparse 하지 않은 변수**

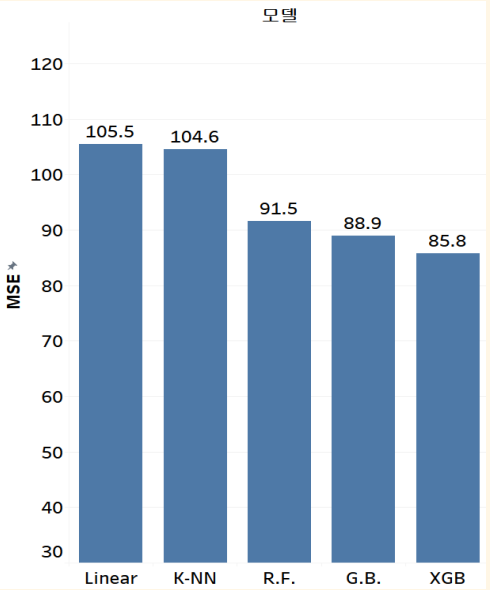
변수선택

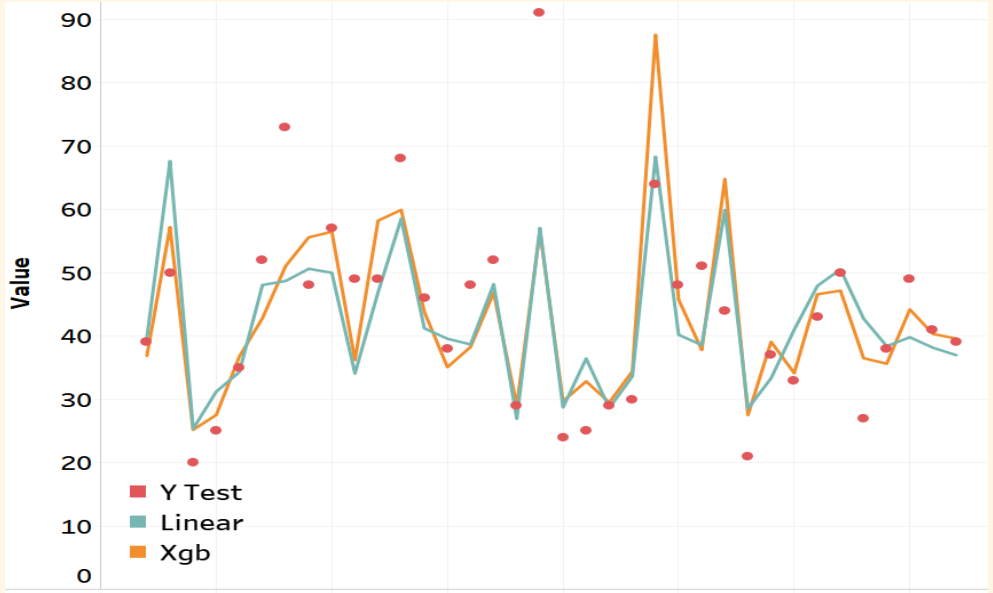
: 범주형 변수 : 수치형 변수 : NA 를 채운 데이터

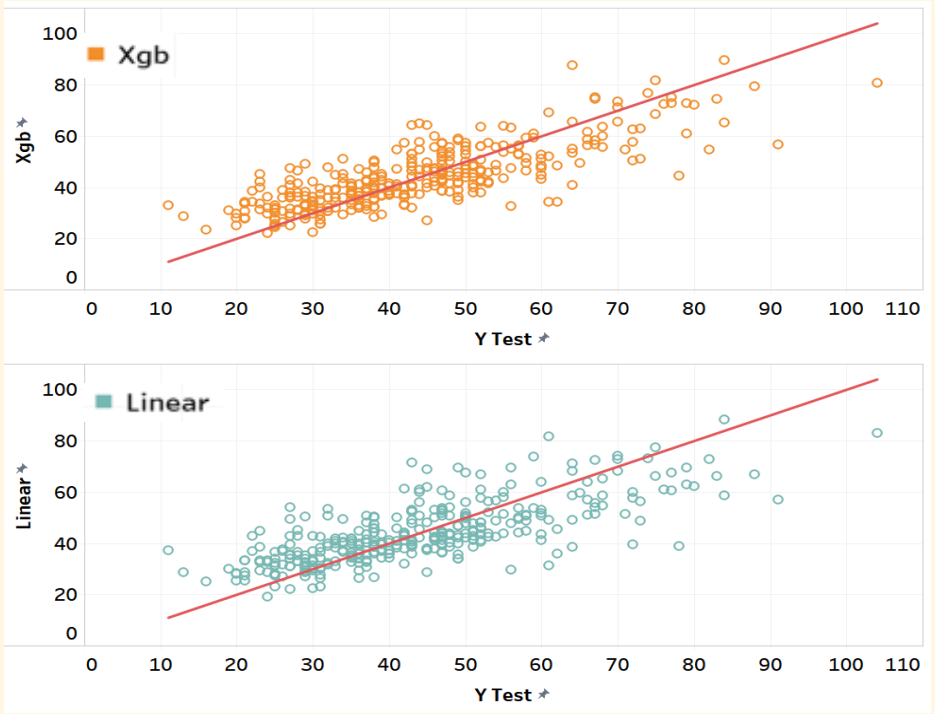
변수	요일	월	공휴일	방학여부	평균 기온	강수량	지하철 승하차 인원	영업 가게 수	네이버 검색량	유튜브 조회수	주문 수
모집단	시간 데이터				대전				전국		Y 변수
출처	날짜 에서 추출	날짜에서 추출	휴일 정보 데이터	배재대 학사 일정	기상청	기상청	CIA 가정 대전 도시 철도공사	인허가 데이터	모델 생성 네이버 트랜드	모델 생성 유튜브조회수 수집 사이트	Bhc 치킨집 포스기

- 시간적 특성을 반영하기 위해 많은 시간 데이터를 넣었다.
- 정보가 비슷한 경우는 삭제하였다.(평균,최고,최저기온)
- 선형적으로 연관이 컸던 변수는 삭제하였다. (인구, 영업가게 수의 상관계수는 0.95)
- NA 는 CIA가정으로 채우거나, 모델을 만들어서 채웠다.(조별 3차 보고서 참고)
- 코로나와 관련된 변수가 없는데, 그 이유는 옆 기사를 보면 알 수 있듯이 요즘에는 확진자가 증가함에도 사람들이 안전 불감증에 빠지면서 코로나에 대한 행동과 인식이 바뀌고 있다. 그러므로 코로나 확진자 정보는 사람들의 행동을 올바르게 측정하지 못한다고 판단해 삭제하였다.
- 코로나 정보 대신에 지하철 승하차 인원, 네이버 검색량이 사람들의 코로나에 대한 행동과 인식에 대한 정보를 가지고 있다고 생각하였다. 사람들이 외출을 많이 하면 지하철 승하차 인원이 늘 것이고, 집에 많이 있으면 인터넷 쓰는 시간이 늘어나 네이버 검색량이 늘 것이라고 생각하였다.
- 최종적으로 11개 변수를 선택하였다.

3 모델링 방법 탐색







모델의 비교

- 시계열 데이터이지만, 시계열 모델을 쓰기에는 Domain 지식이 부족한점, 모델이 불안정한 점이 있어서 회귀 모델을 쓰기로 하였다.
- 임시로 데이터에 정규화를 모두 진행하고 Linear regression, K-NN regression, Random forest, Gradient Boosting, XGB 를 사용해 보았다.
- Test MSE가 별로 차이가 나지 않음을 볼 수 있다. 예측이 어떤 차이가 있는지 보기 위해 실제 Y 값과 그 예측을 비교해 보았다(가운데 그림). 그림을 보게 되면 값들의 추세는 어느정도 예측하고 있으나, 매우 크거나 작은 값은 제대로 예측하지 못하고 있다.
- 전체적인 예측의 추세를 보기 위해 맨 오른쪽 그림같이 실제 Y 값과, 예측의 그래프를 그려보았다. 빨간색 라인이 실제 값이므로, 빨간색 라인에 가깝게 예측하는 것이 정확도가 높은 것이다. 두 모델 모두 대부분 실제값이 작을때는 크게 예측하고, 실제값이 클 때에는 작게 예측하고 있다.

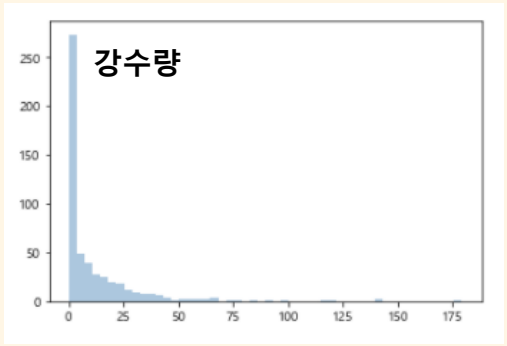
오차의 해석

- 그 이유는 단체행사, 불규칙한 단축영업, 스포츠 경기, 등 의 경우로 추측된다.
- MT, 기숙사 행사 등의 행사가 있거나, 스포츠경기(한일전, 야구, 등)의 경우 데이터에 정보가 없어 모델이 예측할 수 없으므로 실제보다 낮게 예측하게 된다.
- 개인적인 사정으로 단축 영업한 경우도 있다고 한다. 이 경우도 데이터에 정보가 없어 매출을 실제보다 높게 예측하게 된다.

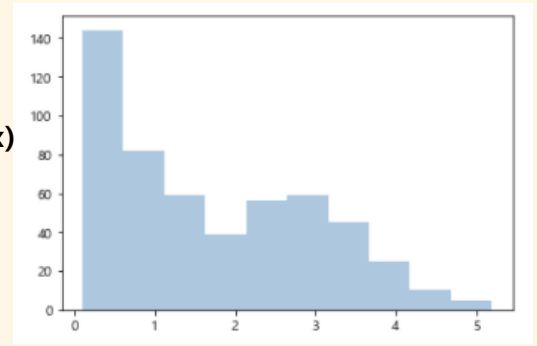
모델의 결정

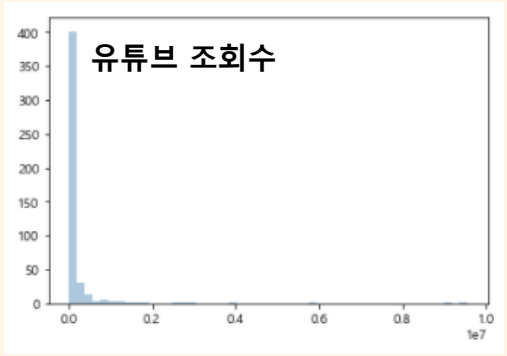
- 여러 모델 모두 비슷한 추이의 데이터를 예측하는 것으로 봐서, 현재 모인 데이터 상에서는 regression 을 이용 하였을 때 이 정도가 한계점이라고 생각했다.
- 즉 현재 데이터 상에서 예측력이 다른 모델과 크게 차이가 나지 않고, 해석력이 매우 좋으며, 변수의 유의성도 검정할 수 있고, 학부 수준에서는 다른 모델보다 이해도가 높았던 선형 회귀를 이용하기로 하였다.
- 선형회귀를 이용해 각 계수 및 유의성을 참고하면 변수등이 매출에 어떻게 영향을 끼치는지 해석할 수 있을 것이다.

4 데이터 변환

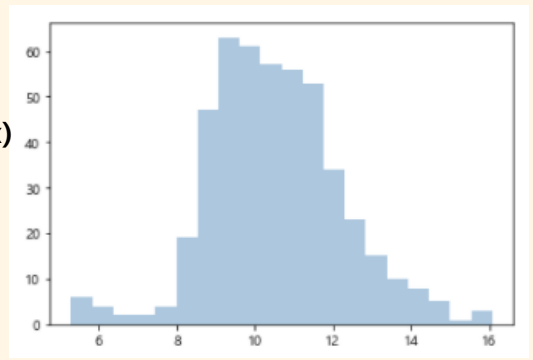


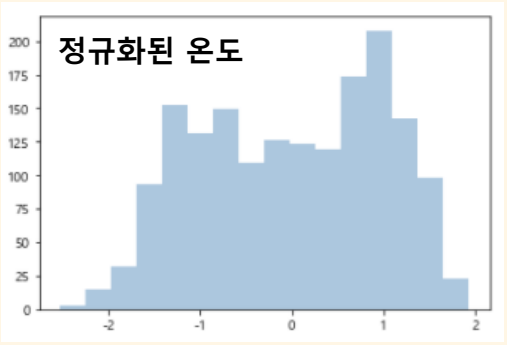
Log(1+x)



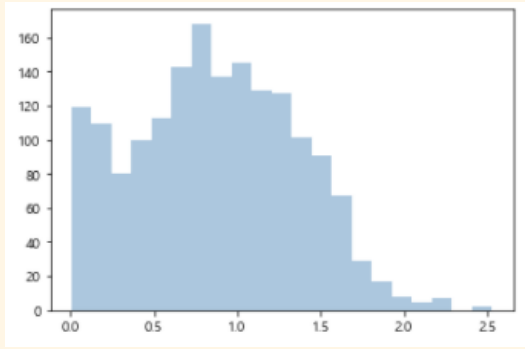


Log(1+x)





절댓값



- 모델을 선형 회귀로 정하였으므로, 그에 알맞게 변수를 변환해서 사용해야 했다.
- 모델이 어떤 변환을 하였을 때 Test mse 가 잘 나오는지, R-square 값이 커지는지, 계수가 유의해 지는지 등을 검사하면서 좋은 변환을 찾았다.
- 그 결과 옆 그림과 같았다. 강수량의 경우 사람이 느끼는 체감은 강수량의 절대치와는 다르기 때문에 log 변환을 해 주었다.
- 조회수 역시, 신제품별로 조회수가 극명하게 갈리는데에 비해, 체감 인기는 이에 비례하지 않기 때문에 log 변환을 해 주었다.
- 온도의 경우 '극단적인 온도(춥거나, 덥거나)' 에서 치킨 주문수가 감소하였다. 즉 이를 반영하기 위해선 온도가 평균보다 매우 높거나 낮을 때 큰 값을 가지게 하는 게 좋을 것이다. 그러므로 정규화 이후, 절댓값을 취하였다.
- 나머지 변수에 대해서는 정규화를 진행하였다.
- 위의 결과를 반영하여 선형 회귀를 실행한 결과 test MSE 가 97까지 낮아졌고, 각 계수들은 이전보다 훨씬 유의성이 좋아졌다.

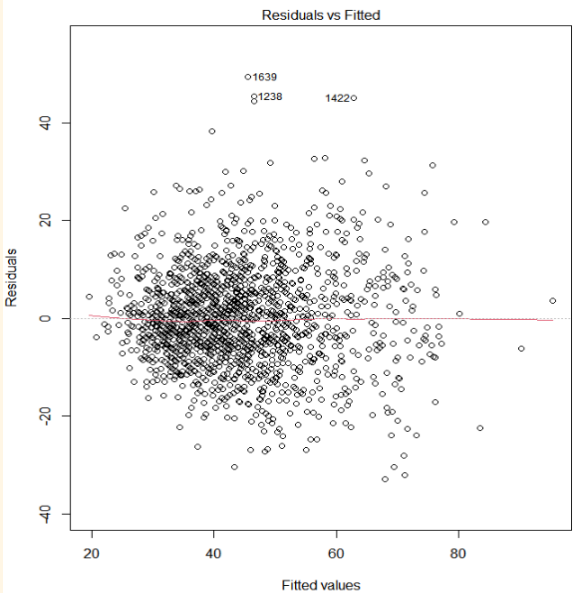
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	48.2270	1.8658	25.849	< 2e-16	***
holyday	-7.1903	1.3384	-5.372	8.88e-08	***
vacation	-2.4302	1.0389	-2.339	0.019439	*
temp	-0.4425	0.7708	-0.574	0.566024	
rain	0.2703	0.2576	1.049	0.294137	
subway	-7.8390	0.4420	-17.736	< 2e-16	***
store	0.6627	0.2999	2.210	0.027273	*
naver	7.6385	0.3823	19.980	< 2e-16	***
youtube	0.2128	0.2516	0.846	0.397872	
thu	-3.8478	0.9650	-3.987	6.97e-05	***
wed	-5.1716	0.9961	-5.192	2.34e-07	***
mon	-2.8963	0.9923	-2.919	0.003563	**
sun	-15.1885	1.2449	-12.201	< 2e-16	***
sat	-7.9571	1.0140	-7.847	7.53e-15	***
tue	3.3730	0.9592	3.516	0.000449	***

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	52.71542	1.80672	29.177	< 2e-16	***
holyday	-7.59412	1.33901	-5.671	1.67e-08	***
vacation	-1.91265	1.04746	-1.826	0.068029	.
temp	-2.87777	0.84914	-3.389	0.000718	***
rain	0.09013	0.26466	0.341	0.733494	
subway	-8.06924	0.45109	-17.888	< 2e-16	***
store	0.70916	0.29948	2.368	0.017998	*
naver	7.40854	0.39183	18.908	< 2e-16	***
youtube	0.75039	0.26800	2.800	0.005170	**
thu	-4.04591	0.96381	-4.198	2.84e-05	***
wed	-5.42749	0.99211	-5.471	5.17e-08	***
mon	-3.19961	0.99092	-3.229	0.001267	**
sun	-15.69419	1.25259	-12.529	< 2e-16	***
sat	-8.16394	1.01284	-8.060	1.43e-15	***
tue	3.13165	0.95696	3.272	0.001088	**

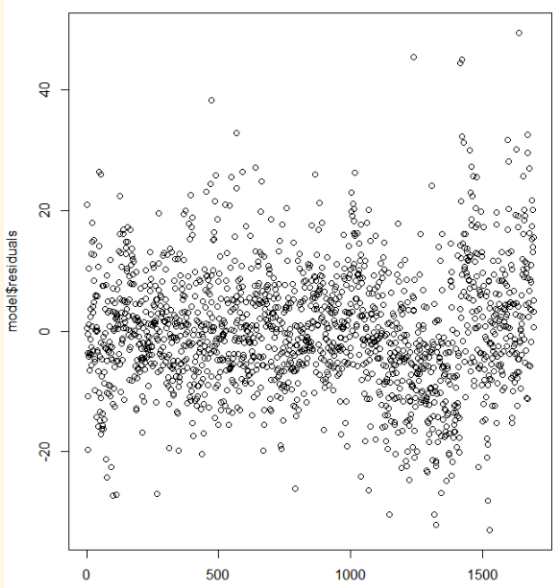
** 나머지 계수들은 지면상 생략하였다. Rain(강수)의 유의성이 더 나빠진 듯 하지만, 이는 Subway가 강수량의 설명력을 가져갔기 때문이다. Subway를 빼고 돌리게 되면, 변환 후의 Rain이 훨씬 유의하게 나온다.

5 선형회귀 가정사항 체크

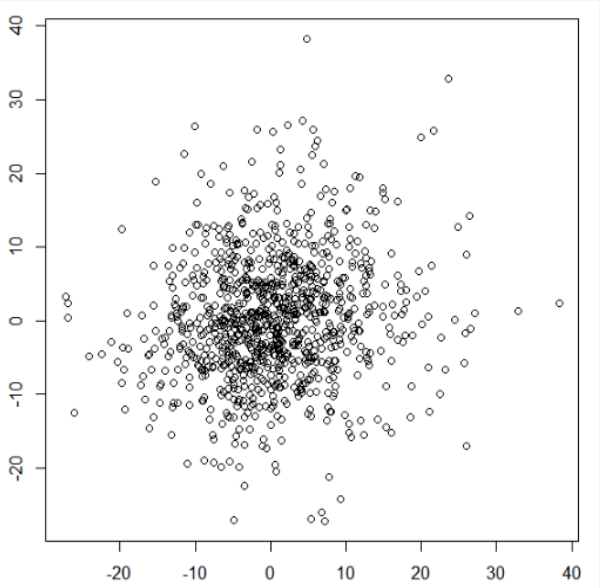
Fitted VS Residual



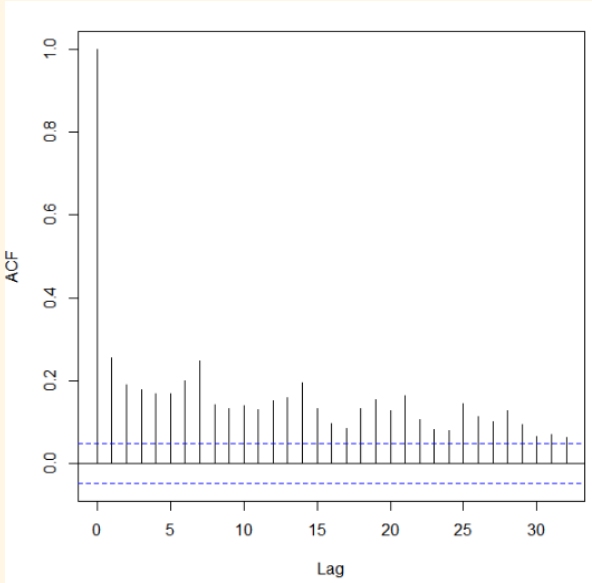
Time VS Residual



e_{i-1} VS e_i



ACF of Residual



1. 선형성

Fitted 값과 residual 의 추이를 비교해 보았다. Residual 값의 의 추이는 빨간색 선으로, 0 에 근접하며 수평에 가까운 선형으로 나타났다. 이는 잔차가 퍼져있긴 하지만 Hyperplane 위에서 양과 음 값으로 고르게 나타난다는 것이다. 이는 y 이 X값들에 대해 약하게 나마 선형성을 가지고있는듯 하다.

2. 오차의 등분산성

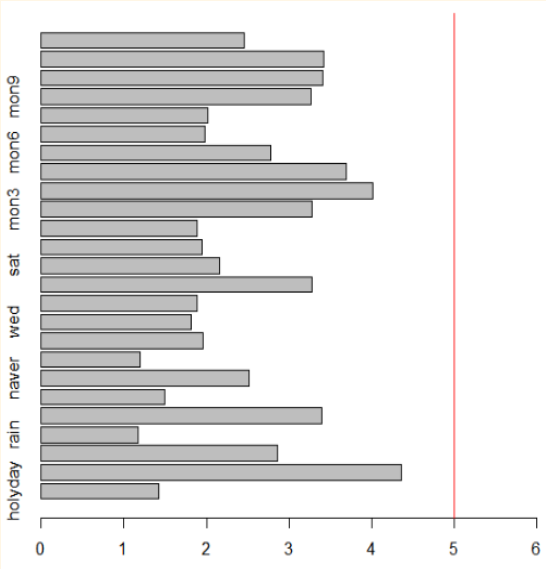
Fitted 값과 residual 의 추이를 보면 fitted value 가 20~40 까지는 깔때기 모양을 보이는 듯 하다가 그 이후에는 약간 안정되는듯 하다. (위로 동떨어진 세 개 데이터는 이상치라고 판단하였다.) 그러므로 등분산성을 약간 위배하고 있다.

3. 오차의 독립성

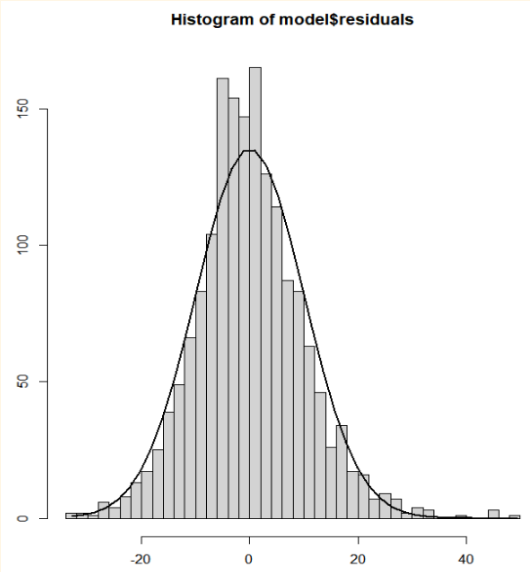
시계열 데이터 이므로 시간에 따른 잔차 변화를 중심적으로 보았다. 시간별로 그려보니 1000~1400번째에서 잔차가 떨어지는 추세를 관찰 가능하였다. 그 뒤에 1 차이나는(하루 차이) 잔차의 scatter 분포를 그려보니 큰 패턴이 파악되지는 않았고, ACF 값도 최대값이 0.2 부근으로, 큰 상관은 없다고 판단하였다. 즉 1000~1400번째 잔차에 대해 추이가 나타남에 따라 독립성을 약하게 위반하고 있다.

6 선형회귀 가정사항 체크

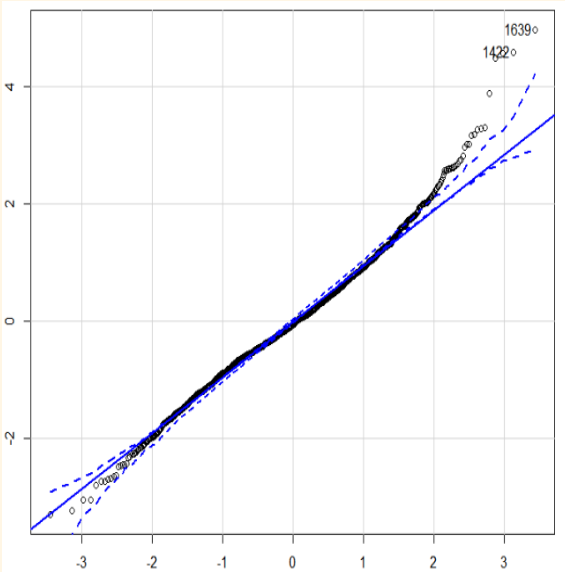
변수들의 VIF값



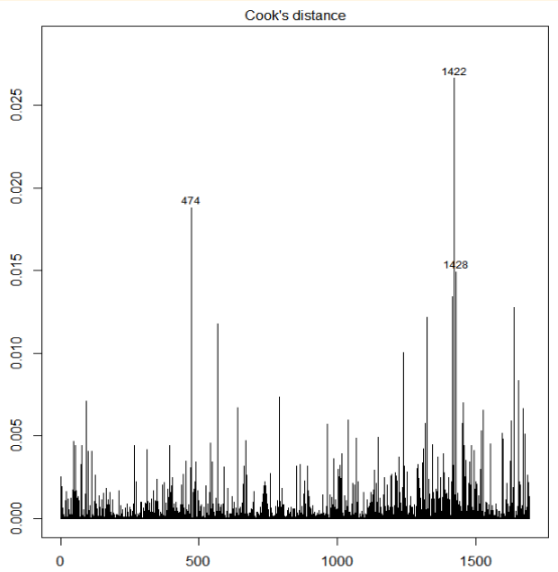
잔차 히스토그램과 Normal



QQ Plot



Cooks Distance



4. 다중 공선성의 여부

VIF 의 값을 수치형 변수에 대해 계산해 본 결과 vif 값이 5 이상의 다중공선성이 보이는 변수는 없었다.

5. 오차의 정규성

잔차 히스토그램과, 그와 비슷한 Normal 분포를 위에 그려보았다. 잔차가 Normal 보다 약간 더 중앙이 뾰족해 보이는듯 하다. 그리고 QQ plot 에서도 직선의 모양이라기 보다는, 양 끝이 들려 있다. 그에 따라 95% CI 의 Interval 에서 벗어난 점을 볼 수 있다. 즉 정규성을 약간 위배하고 있다.

6. 이상치의 여부

Cooks distance 그래프에 따라 값이 매우 큰 몇 개의 영향치들을 관찰할 수 있었다. 이 영향치들을 조사한 결과, 100~90건의 기록적인 매출을 기록한 날 이였다. 내가 모든 데이터 상에서는 이런 매출 증가에 대한 설명을 할 수 없었다. 나중에 다른 변수가 있는 데이터를 붙이게 되면 이런 이상치들을 설명할 수도 있으므로, 이상치를 제거하지 않고 유지하기로 하였다.

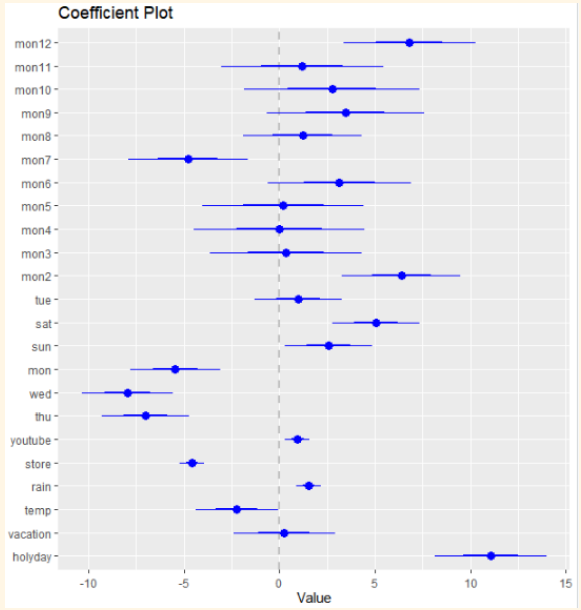
7 선형회귀의 변수해석

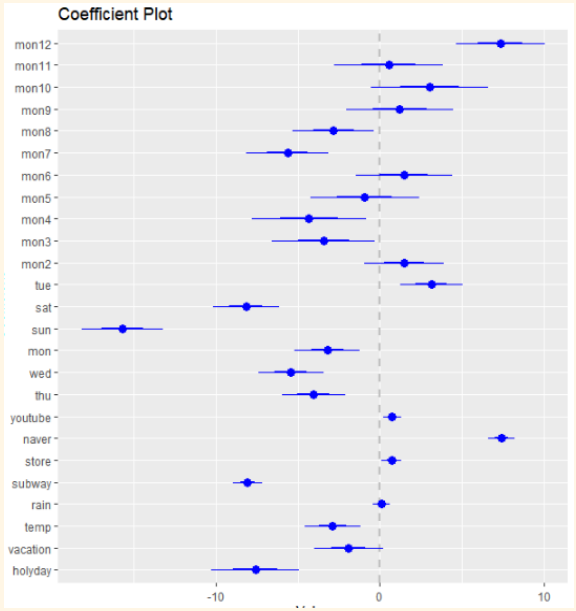
Subway, naver 변수 제외

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 45.17969    2.24806  20.097  < 2e-16 ***
holyday     11.06307    1.45682   7.594 5.13e-14 ***
vacation     0.24123    1.33201   0.181  0.85631
temp        -2.23613    1.08253  -2.066  0.03901 *
rain         1.61759    0.33173   4.876 1.18e-06 ***
store       -4.58442    0.31596 -14.510  < 2e-16 ***
youtube      0.93422    0.32627   2.863  0.00424 **
thu         -7.01956    1.14380  -6.137 1.05e-09 ***
wed         -7.98027    1.19421  -6.682 3.19e-11 ***
mon         -5.45771    1.17883  -4.630 3.94e-06 ***
sun          2.57212    1.14876   2.239  0.02528 *
sat          5.04592    1.14601   4.403 1.14e-05 ***
tue          0.99543    1.14221   0.871  0.38361
mon2         6.37575    1.54435   4.128 3.83e-05 ***
mon3         0.35191    1.98613   0.177  0.85939
mon4        -0.02002    2.22996  -0.009  0.99284
mon5         0.20689    2.11081   0.098  0.92193
mon6         3.13321    1.86887   1.677  0.09382 .
mon7        -4.80475    1.56546  -3.069  0.00218 **
mon8         1.20781    1.55685   0.776  0.43798
mon9         3.44838    2.06457   1.670  0.09505 .
mon10        2.74404    2.28916   1.199  0.23081
mon11        1.18762    2.12129   0.560  0.57565
mon12        6.80093    1.72682   3.938 8.54e-05 ***
```

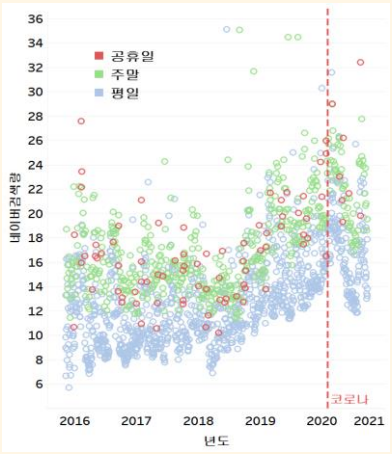
Subway, naver 변수 포함

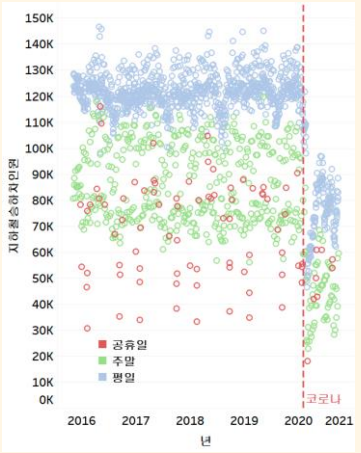
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.67053    1.82461  28.867  < 2e-16 ***
holyday     -7.59412    1.33901  -5.671 1.67e-08 ***
vacation    -1.91265    1.04746  -1.826 0.068029 .
temp        -2.87777    0.84914  -3.389 0.000718 ***
rain         0.08509    0.24986   0.341 0.733494
subway      -8.06924    0.45109 -17.888  < 2e-16 ***
store        0.70916    0.29948   2.368 0.017998 *
naver        7.40854    0.39183  18.908  < 2e-16 ***
youtube      0.75039    0.26800   2.800 0.005170 **
thu         -4.04591    0.96381  -4.198 2.84e-05 ***
wed         -5.42749    0.99211  -5.471 5.17e-08 ***
mon         -3.19961    0.99092  -3.229 0.001267 **
sun        -15.69419    1.25259 -12.529  < 2e-16 ***
sat         -8.16394    1.01284  -8.060 1.43e-15 ***
tue          3.13165    0.95696   3.272 0.001088 **
mon2         1.47595    1.22231   1.208 0.227405
mon3        -3.44157    1.56330  -2.201 0.027838 *
mon4        -4.31891    1.75350  -2.463 0.013877 *
mon5        -0.94178    1.65596  -0.569 0.569622
mon6         1.45866    1.46699   0.994 0.320209
mon7        -5.63487    1.24176  -4.538 6.09e-06 ***
mon8        -2.83126    1.23407  -2.294 0.021900 *
mon9         1.20820    1.62319   0.744 0.456777
mon10        3.02400    1.79565   1.684 0.092355 .
mon11        0.52752    1.66594   0.317 0.751548
mon12        7.33251    1.35527   5.410 7.20e-08 ***
```

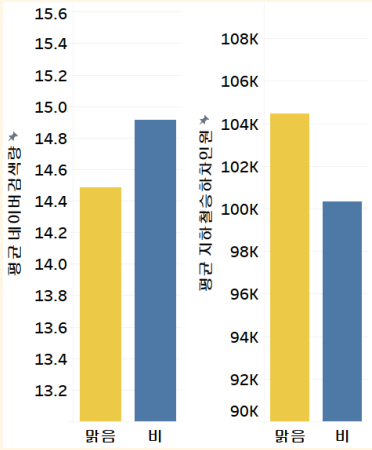




시간과 강수 영향을 받는 네이버 검색량과 승하차 인원







- Naver, subway 는 시간과 강수의 영향을 많이 받는 변수이다. 그러므로 이를 빼고 돌리면 각 요일과 강수에 대한 효과를 볼 수 있다. 그리고 이 계수는 우리의 관찰 결과와 같다. (주말,휴일이나, 비가 오면 더 잘팔림)
- Subway와 naver변수를 넣으면 시간과 강수 관련된 계수가 크게 변화한다. 이는 Subway 와 naver 변수가 시간과 강수의 설명력을 많이 가져간것 으로 해석된다.
- 즉 최종 모델에서 시간과 강수에 관련된 변수의 계수는 독립적으로 해석하면 위험하다. (유동성 감소 및 검색량 증가의 현상도 같이 고려해야하기 때문)
- 그리고 가정의 위배상황, 변수의 오차(전국 데이터의 사용, NA 를 채운 데이터 사용)으로 인한 에러의 증가로, 모델의 계수 유의성이 더 안 좋아질 것이고, 95% CI 의 길이도 더 길어질 것이다. 그리고 계수의 값도 약간 변할 수 있다.
- 그에 따라 유의성이 애매한 값들 (3월,4월,영업가게 수 등) 은 사실 유의하지 않을 수 있다.

8 모델링 Summary

NA를 채운 변수 사용
- 네이버 트랜드, 지하철 승하차 인원은 일부 NA를 채운 데이터이다.

유튜브 조회수
- 일별 조회수를 모델을 만들어 근사해 추정

오차의 정규성 위배
- 잔차가 정규성을 약간 위배한다.

측정에서의 오차

모델 가정의 오차

전국 변수 사용
- 네이버 검색량, 유튜브 조회수는 전국을 기준으로 측정

오차의 등분산성 위배
- 잔차가 약간의 깔때기 모양을 보인다.

오차의 독립성 위배
- 코로나 직전에 잔차의 추세가 보임

갑작스런 휴업
- 개인적인 사정으로 일부만 장사하는 경우 주문수가 매우 적었다.

단체 주문
- 행사, 스포츠 경기 등으로 인해 주문이 많아진 경우 모델이 예측할 수 없었다.

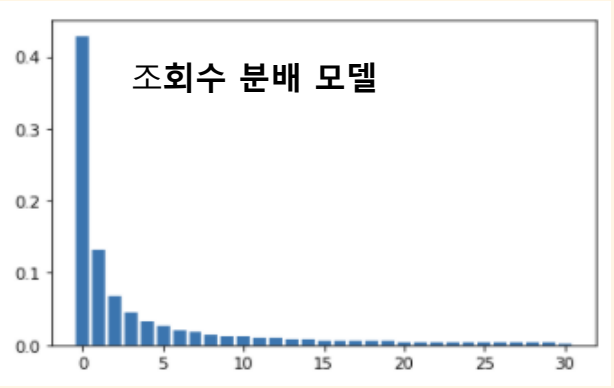
이상치의 오차

선형 모델

- 오차들이 위와 같이 쌓였기 때문에 실제 선형 회귀의 변수들의 유의성은 모델의 결과보다 더 안 좋을 것이고, 실제 계수의 값도 조금 달라질 수 있다.
- 다만, 네이버 검색량, 승하차, 요일, 휴일, 7월과 12월의 경우는 p값이 매우 낮았다. 이 변수들은 오차를 감안 하더라도 유의하다고 생각할 수 있다.
- 네이버와 승하차 변수가 시간관련 변수와 강수량 변수의 설명력을 많이 잡아먹었다. 그에 따라서 변수 해석시 계수를 독립적으로 해석하면 위험하다. 예시로 비의 경우 유의하지 않아 보이나, 지하철 승하차를 감소시키는 효과가 있음을 고려하면 유의한 변수일 것이다.
- 네이버 검색량(인터넷 트랜드)와 승하차 인원(유동성), 그리고 시기(휴일 및 요일)가 매출에 큰 영향을 끼치고 있다. 그리고 기후와 유튜브 조회수(신제품 트랜드)도 매출에 작은 영향을 끼치고 있다.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	52.67053	1.82461	28.867	< 2e-16	***
holyday	-7.59412	1.33901	-5.671	1.67e-08	***
vacation	-1.91265	1.04746	-1.826	0.068029	.
temp	-2.87777	0.84914	-3.389	0.000718	***
rain	0.08509	0.24986	0.341	0.733494	
subway	-8.06924	0.45109	-17.888	< 2e-16	***
store	0.70916	0.29948	2.368	0.017998	*
naver	7.40854	0.39183	18.908	< 2e-16	***
youtube	0.75039	0.26800	2.800	0.005170	**
thu	-4.04591	0.96381	-4.198	2.84e-05	***
wed	-5.42749	0.99211	-5.471	5.17e-08	***
mon	-3.19961	0.99092	-3.229	0.001267	**
sun	-15.69419	1.25259	-12.529	< 2e-16	***
sat	-8.16394	1.01284	-8.060	1.43e-15	***
tue	3.13165	0.95696	3.272	0.001088	**
mon2	1.47595	1.22231	1.208	0.227405	
mon3	-3.44157	1.56330	-2.201	0.027838	*
mon4	-4.31891	1.75350	-2.463	0.013877	*
mon5	-0.94178	1.65596	-0.569	0.569622	
mon6	1.45866	1.46699	0.994	0.320209	
mon7	-5.63487	1.24176	-4.538	6.09e-06	***
mon8	-2.83126	1.23407	-2.294	0.021900	*
mon9	1.20820	1.62319	0.744	0.456777	
mon10	3.02400	1.79565	1.684	0.092355	.
mon11	0.52752	1.66594	0.317	0.751548	
mon12	7.33251	1.35527	5.410	7.20e-08	***

9 모델의 이용

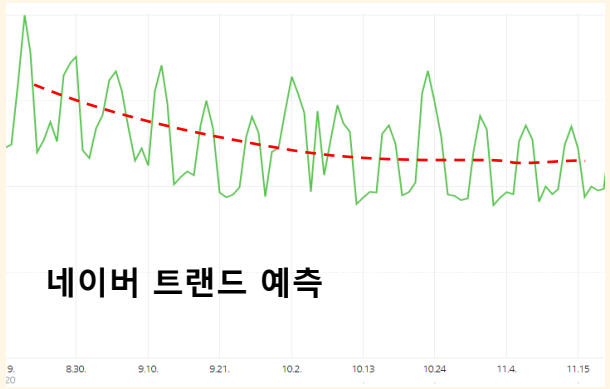


예측

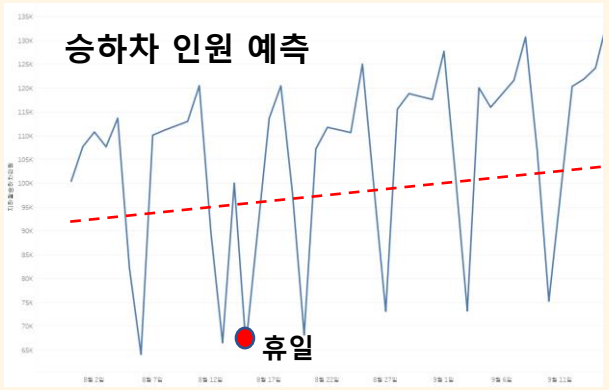
변수	휴일,월 등	평균 기온	강수량	지하철 승하차 인원	영업 가게 수	네이버 검색량	유튜브 조회수
예측	쉽게 채울수 있음	기상청 예보	기상청 예보	시계열 데이터에서 예측	저번달 인허가 데이터	시계열 데이터에서 예측	유튜브 조회수 모델에서 계산

예측

주문 수
BHC 배재대점 치킨 주문수



공휴일,강수 보정



단기간 예측 가능

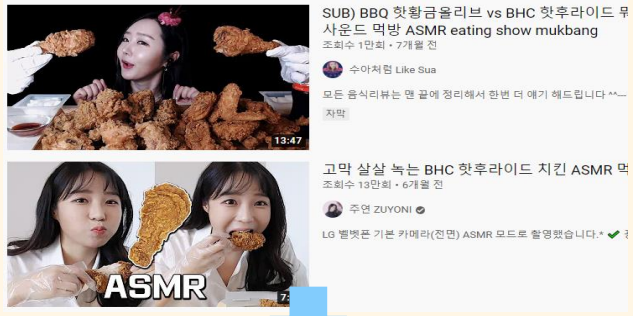
- 네이버 트랜드, 승하차 인원의 경우 그 추이가 시간순으로 보면 단순한 추이(빨간색) 와 계절성(일주일 간격)이 보인다. 이전 두 달 정도의 데이터를 이용하면, 시간에 대한 추이를 단기적으로는 어느정도 예측 가능할 것이다.
- 그리고 위 예측 추세에서 공휴일과 강수의 보정이 필요하다. 평균적으로 비가 오면 승하차 인원은 4000명이 감소하고, 네이버 검색량은 0.43이 증가했고, 공휴일인 경우 승하차 인원은 39000명 감소하고 네이버 검색량은 2.742 증가했다. 이를 보정해주면 될 것이다.
- 인허가 데이터의 업데이트 주기는 한달이다. 그리고 인허가가 난 이후 어느정도 기간이 지나야 실제 영업이 가능해 질 것이다. 즉 영업 가게 수 데이터는 저번 달 인허가 데이터를 사용하여도, 예측치로서 충분할 것이다.
- 유튜브 조회수의 경우, 이미 우리가 누적 조회수 모델을 조별 보고서에서 Least square method 로 만들었기 때문에, 그 모델을 이용하여 채워 넣을 수 있다.
- 그러므로, 10일 정도의 단기 매출은 어느정도 예측이 가능할 것이다.

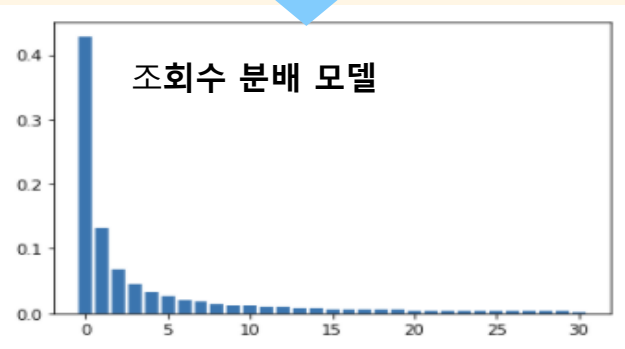
정상적인 매출이라는 기준점이 될 수 있음

- 이 모델을 사용한다면, 이 모델의 예측치를 '정상적' 으로 간주할 수 있다. 만약 감소폭이나 증가폭이 예측치보다 매우 달라진다면 어떤 원인 때문인지 조사해 보고 그에 따라 조치를 취할 수 있을 것이다.
- 예시로 코로나가 점점 호전된다면, 재택근무가 줄면서 주문수의 감소가 있을 것이다. 그 감소폭에 대해 모델의 예측치와 비교함 으로서 감소폭이 정상적인 감소인지 감시가 가능하다.

10 모델의 실제 적용 예시

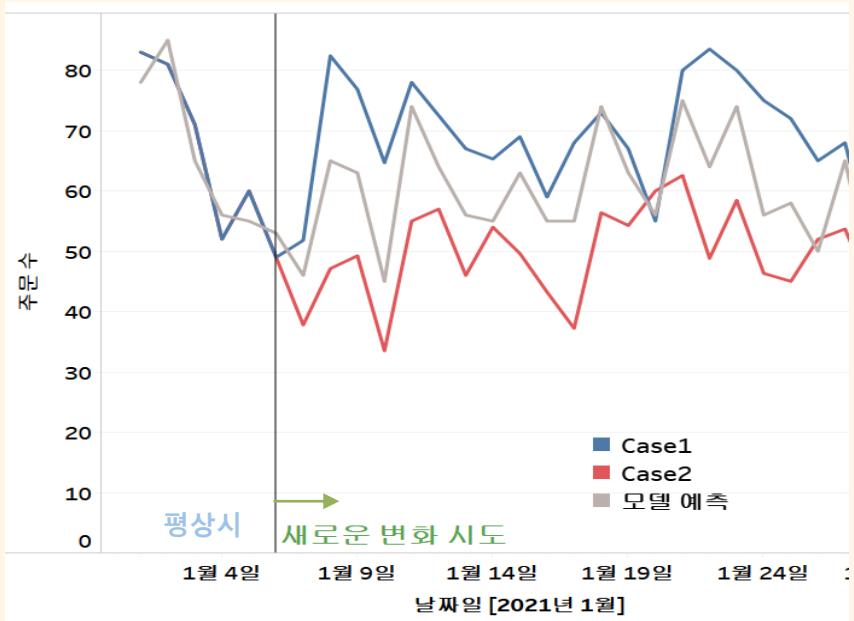
신제품의 효과가 궁금해





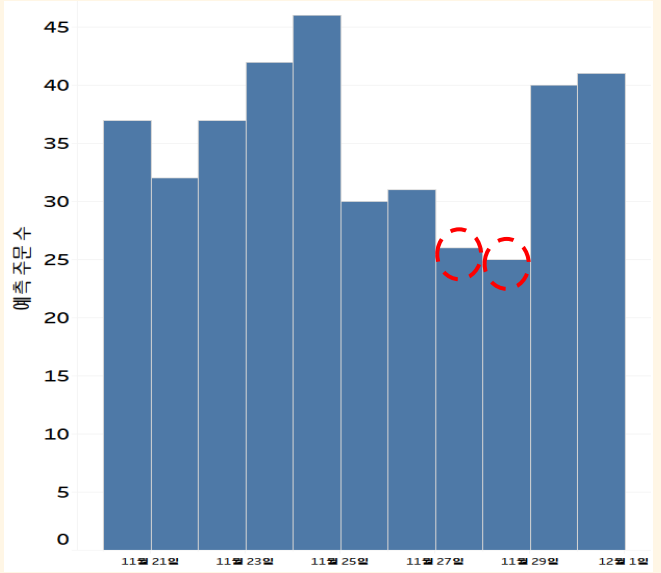
- 신제품이 출시되면, 그 메뉴를 리뷰하는 유튜브 영상이 나올 것이다.
- 조회수 상위 5개 동영상의 첫날의 조회수를 더하고, 모델의 비율에 따라 뒤 30일까지 예측한다.
- 알맞은 변환 이후 우리의 모델에서의 계수와 곱해 어느 정도의 주문 수 증가가 있을 것이라 예측 가능하다.
- 그에 따라 재고 관리 등이 가능할 것이다.

새로운 방법의 효과가 궁금해



- 다른 양념을 쓰거나, 배달 어플 플랫폼을 바꾸는 등의 변화를 해보고 그 효과가 궁금할 수 있다.
- 우리 모델은 새로운 변화 전에 대한 fitting 이기 때문에, 이 모델의 예측과 비교했을 때에 위에서 시도한 방법이 유효한지 예측 값과 비교해보면 알 수 있을 것이다.
- Case1 : 모델의 예측보다 대체로 크다. 그렇다면 새로운 고객이 늘었다는 것이고, 변화된 판매 전략이 유효했음을 의미한다.
- Case2 : 주문수가 약간 증가하는 추세긴 하지만 모델의 예측보다 대체로 작다. 즉 이는 새로운 변화가 부정적 이였고, 다시 원래의 방법으로 돌아가야 함을 의미한다.

장사를 언제 쉬어야 할까?



- 장사를 하다 보면, 병원 예약, 가족여행 등의 쉬는 날이 꼭 필요하다.
- 하지만 쉬는 날에는 그 날의 고객들은 다른 브랜드의 치킨을 먹게 되고, 재구매율이 높은 치킨 특성상 기존 고객에서 이탈될 수 있다.
- 모델을 이용하여 뒤 10일중에서 제일 매출이 적은 날을 특정지을 수 있고, 그 날 장사를 쉰다면, 기존 고객 이탈을 최소화 시킬 수 있을 것이다.

Reference

1page 치킨 충성도 기사 https://www.newswire.co.kr/newsRead.php?no=906558

1page 치킨 재구매 https://www.bigdata-finance.kr/dataset/datasetView.do?datastId=SET1600006

2page 대전 안전불감증 기사 http://www.chungnamilbo.co.kr/news/articleView.html?idxno=574791