데이터 통합
개인 1차 보고서

2014131026 한인욱

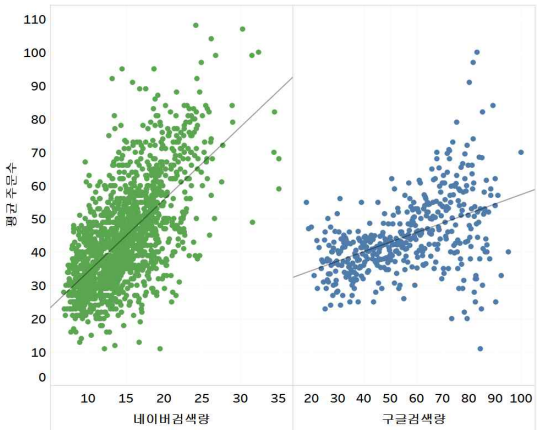
<목차>

1. Research Question

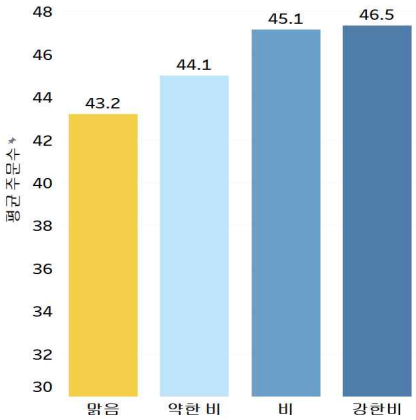
우리 조는 처음부터 Research Question을 치킨집 매출에 관한 분석으로 정하고 프로젝트를 진행하였다. 즉 그에 따라 Research Question은 '치킨집 매출 분석 및 최대화 방안'으로 정하였다.

이를 위해서는 단순히 어려운 모델을 이용해 prediction의 정확도를 높이는게 중요한게 아니라, 모델의 해석이 얼마나 가능한지, 모델이 이후에도 계속 데이터가 update 되면서 발전 가능한지, 이 모델을 가지고 어떤 일을 할 수 있는지 등을 염두해 가면서 연구를 진행해야 할 것이다.

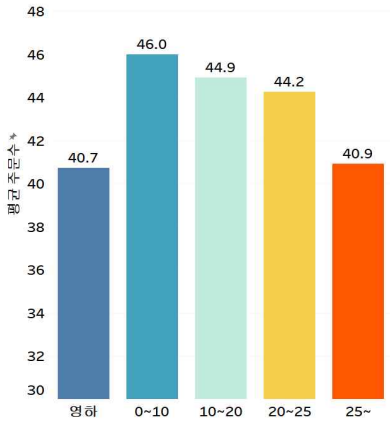
2. 데이터 살펴보기



[그림 1] 검색량과 주문수

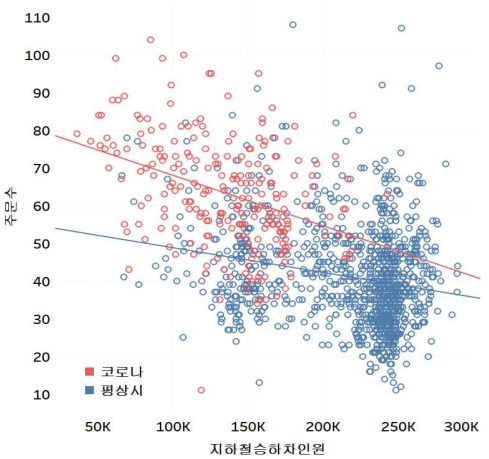


[그림 2] 강수량과 주문수

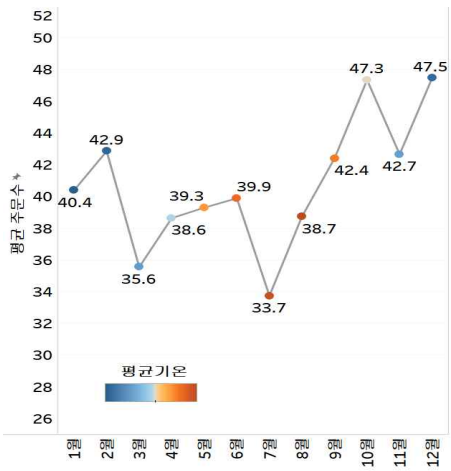


[그림 3] 온도와 주문수

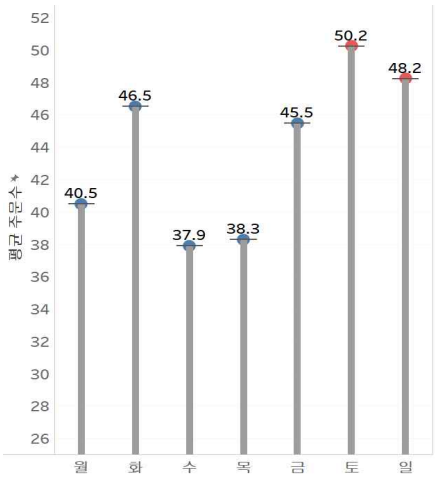
- 주문수와 검색량이 유의한 관계가 있어 보이고, 네이버가 더 뚜렷한 관계를 보이고 있다.
- 강수량이 유의해 보인다. 비가 오면 밖에 나가기 싫어져서 배달을 시키는 것으로 예상된다.
- 영하의 추운 날 이거나, 25도 이상의 더운 날은, 치킨 판매량이 저조함을 볼 수 있다.



[그림 4] 승하차 인원과 주문수



[그림 5] 월별 주문수 (색 : 기온)



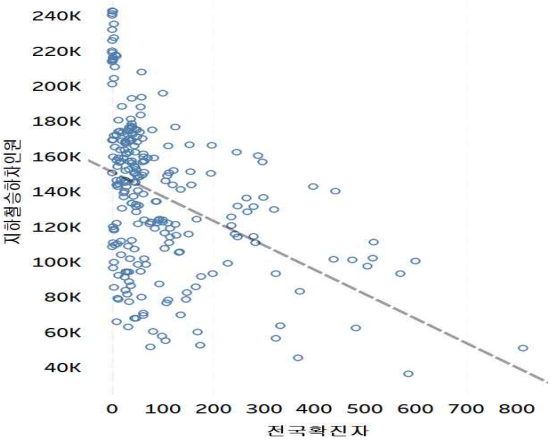
[그림 6] 요일과 주문수

- 지하철 승하차 인원은 일별 주문수와 반비례 관계가 있어 보인다.
- 월별 판매량 차이가 확연하다. 겨울철이 다른 달에 비해 더 많이 팔리는 듯하다.
- 요일 판매량은 주말이 높다. 그리고 화요일이 금요일보다 판매량이 높은 것은 주목할 특징이다.

데이터를 살펴보았을 때, 많은 변수가 판매량과 얽혀있음을 알 수 있다. 이런 변수들을 모두 고려하는 게 중요할 것이다. 특히 그중에서 시간 변수들(월, 요일)과의 관계가 뚜렷한 것을 볼 수 있는데, 이런 시간적인 변수들을 최대한 반영하는 모델을 짜야 할 것이다.

3. 데이터 변수선택

- **선택기준 1** : 미래에도 관측이 되는 변수여야 한다.
변수가 의미가 있으려면 변수가 이후에도 계속 예측이 되고 예측치 형성에 도움이 되어야 한다. 그러나 코로나와 관련된 변수는 중요하긴 하지만, 종식된 이후에는 관찰되지 않을 것이다. 내가 만든 모델이 이 시점 이후에 데이터를 업데이트하며 계속 개선이 가능하려면, 미래에 관찰되지 않는 변수는 제외해야 할 것이다. 그러므로 코로나 변수를 삭제하였다.
대신에 옆 그림과 같이 '지하철 승하차 인원' 이 어느 정도 코로나 확진자 정보를 가지고 있으므로, 이를 사용하기로 하였다.

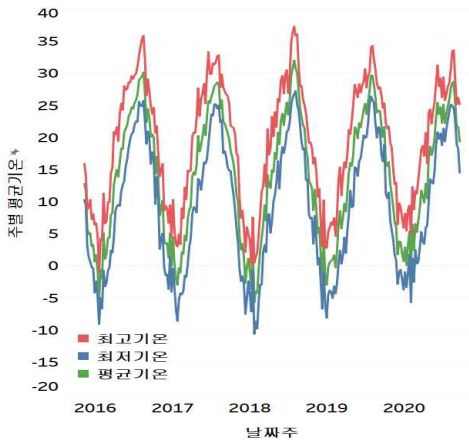


[그림 7] 전국확진자와 승하차 인원 그래프

- **선택기준 2** : 시간적 요소를 최대한으로 고려 가능해야 한다..
데이터가 시계열 데이터이지만, 시계열 모델은 그 모델이 매우 복잡하고 불안정하여 내 수준에서는 다루기 어렵다고 판단했기 때문에 회귀모델을 사용하기로 하였다. 그러므로 회귀모델로 최대한의 시간적 요소를 고려할 수 있도록 기존의 날짜 변수에서 월, 요일의 변수를 추출해 변수를 늘렸다.

- **선택기준 3** : 겹치는 의미의 변수는 제거한다.
최저, 최고, 평균기온은 모두 기온을 측정하는 변수이고 [그림8]을 보면 그 추이가 비슷하다. 이 셋을 통합하여 평균기온만 사용하기로 하였다.

- **선택기준 4** : 정보가 너무 적은 변수는 제거한다.
복날의 경우 1년에 겨우 1번 돌아오는 변수라서, 너무 정보가 적다고 판단하여 제외하였다.



[그림 8] 주별 평균기온 그래프

이에 따라 1차적으로 사용할 변수는 월(categorical), 요일(categorical), 평균기온, 강수량, 구글검색량, 네이버검색량, 지하철 승하차인원, 공휴일(categorical), 방학여부(categorical), 인구, 영업가게를 사용하기로 하였다.

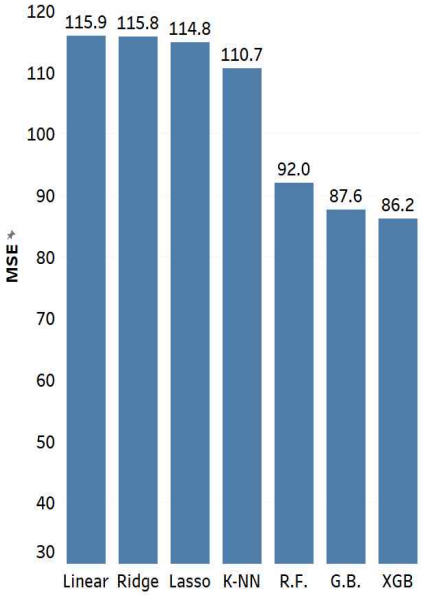
4. 방법론(모델) 탐색

시계열 분석은 매우 어렵고, 불안정하여 회귀분석을 사용하기로 하였다. 모델 후보는 multilinear / lasso / ridge / k-nn regression / xgboost / random forest / gradient boosting 이다.
NA imputation는 임시로 '인구'와 '구글 검색량'은 결측치 앞,뒤 값의 등차수열(선형으로)로 채워넣는 방법을 썼고[그림 9], 나머지 변수에 대해서는 Nearest Neiborhood imputation(N=5)을 사용하였다.

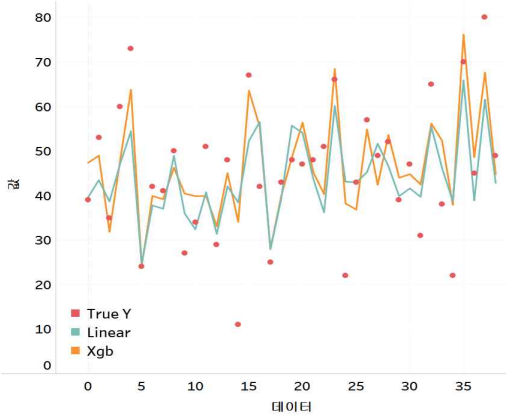
24	24
NA	27
NA	30
NA →	33
NA	36
NA	39
NA	42
45	45

[그림 9] 예시

강수량의 경우사람이 느끼기에 50mm가 10mm의 5배라고 느끼지 않을 것이다. 그러므로 로그 변환($\log(1+x)$) 변환을 통해서 이러한 부분을 보정하려 하였다. 그 이후 나머지 수치형 변수들에 대해서는 standard scaling을 하였다. 나머지 세팅은 [그림 10] 의 설명과 같다.



[그림 10] 적합된 모델의 MSE
※Train,Test set은 은 8:2 로 분리
※월, 일 데이터는 one-hot vector 변환
※hyperparameter는 5-cross validation 에서 MSE가 제일 낮게 나오는 값 선택
※machine learning 모델은 default설정과 비슷하게 설정
※Ridge : λ = 59.636
※Lasso : λ = 0.052
※K-NN regression : Neighborhood = 9
※Random forest = default 설정
※Gradient boost = default 설정
※XGB : n_eastimator = 150, learningrate = 0.08 , maxdepth =3, 나머지는 default 설정
※MSE 는 Test set에 대한 MSE



[그림 11] linear예측, xgb예측, True값 그림

그 결과 MSE 는 [그림 10]와 같이 나온다. multi Linear, Ridge, Lasso 같은 전통적 Linear 모델의 MSE 와, 머신러닝 방법의 MSE 가 크게 차이가 나지 않음을 볼 수 있다. [그림 11]은 True 데이터(빨간점) 40개에 대한 xgb모델과(주황색)과 LINEAR 모델(초록색)의 예측치를 그린 그래프이다. xgb모델이 조금 더 좋은 예측을 하고 있지만 linear 모델도 어느정도 비슷하게 경향을 따라가고 있음을 알 수 있다.

Linear model이 해석이 쉬운 것은 물론, 예측도 어느정도 경향을 올바르게 예측하고 있으므로 나눈 Linear model을 사용하기로 하였다.

그리고 모델의 해석과 이용의 편이성을 위해 변수들을 4개 지표로 압축하여, 4개의 지표를 이용한 linear 모델로 사용하기로 하였다.

5. 지표 탐색

여기서 지표는 '치킨집 주문수'라는 변수를 예측하는데에 도움이 되어야 할 것이다. 그러한 변수는 크게 날씨와, 주변상권, 유동성, 인터넷 유행 4가지 유형으로 나누어 볼 수 있을 것이다. 이 4가지 유형의 지표값이 높을수록 치킨 주문량이 높아지게 지표를 만들기로 하였다.

1. 날씨 지표

- 강수량, 기온으로 구성된 지표. 비가 오거나[그림2], 온도가 적당한 날[그림3]에 치킨이 많이 팔렸으므로, 그런 날에 높은 값을 가져야 할 것이다.

2. 주변상권 지수

- 인구수, 주변 가게수로 구성된 지표. 인구수가 높고, 주변 가게수가 적을수록 장사가 잘 될 것이므로, 큰 값을 가져야 할 것이다.

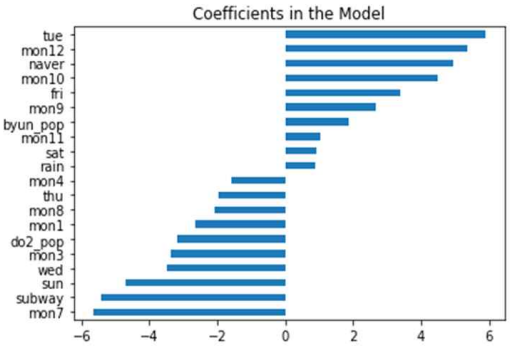
3. 유동성 지수

- 지하철 유동인구로 구성된 지표. 유동인구가 많을수록[그림 4] 치킨배달이 적으므로, 유동인구가 많으면 작은 값을 가져야한다.

4. 유행 지수

- 네이버, 구글검색량. 유튜브 조회수가 클수록[그림1] 큰 값을 가지게 할 것이다.

각 지수들의 구체적인 식은, 적합시킨 모델의 계수와, 그래프, correlaltion 등을 활용하여 결정할 계획이다.

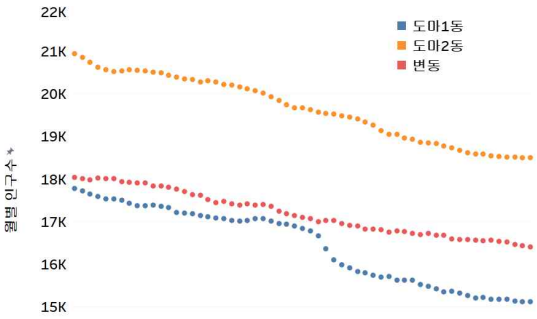


[그림 12] Multi Linear 모델의 계수

6. Imputation 방식 탐색

1. 인구 Imputation

인구수데이터는 월별 데이터이다. 인구수는 갑자기 튀어오르지 않고 그 추세가 명확한 시계열 데이터이다 [그림 13]. 즉 앞 뒤 데이터를 이용해 등차수열으로 imputation 하면 될 것으로 예상된다 [그림 14]



[그림 13] 월별 인구수 추이

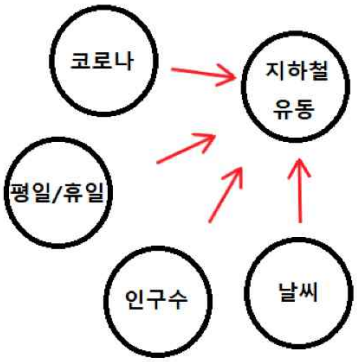
날짜	인구수
6월 1일	10000
6월 2일	NA
...	...
...	...
6월 29일	NA
6월 30일	NA
7월 1일	10030

인구수
10000
10001
...
...
10028
10029
10030

[그림14] 인구수 imputation

2. 지하철 유동 Imputation

지하철 유동에 영향을 끼치는 변수는 인구, 날씨, 주말(토일)여부, 코로나 여부가 있다고 가정하고, 이 네 개가 주어지만 나머지에 대해서 독립이라고 가정하자. (CIA) 그러면 인구, 날씨지수, 주말 여부(one hot vector), 코로나여부(one hot vector) 로 Micro approach를 이용, linear regression based imputation으로 지하철 유동 변수를 imputation 가능할 것이다.



3. 구글 Trend Imputation

구글 trend는 '전국'의 주별 데이터이므로, 이를 imputation 하려면 기존의 '대전' 데이터를 활용하면 안 될 것이다. imputation 하기 위해 population 이 일치하는 '전국' 데이터를 더 찾아보고 이를 이용해 imputation 할 계획이다.

결측치가 모두 채워진 complete 데이터가 일부분 존재하고 있으므로, imputation 방법론들이 True 값을 잘 예측하는지를 Test 할 수 있다. 즉 많은 방법론이 True 값에 얼마나 근접하게 imputation 하는지 시험해 보고, 그에 따라서 제일 좋은 성능을 보이는 방법론을 선택할 계획이다.

7. 결론 및 계획

임시로 적합 시킨 모델들의 MSE 가 생각보다 높게 나왔는데 이는

1. 임시로 실행했던 NA imputation이 적절치 못하였다.
2. scaling 방식 및 변수통합 방식이 적절치 못하였다.
3. 아직 데이터가, y를 제대로 예측할 만큼 충분히 모이지 못했다.

의 가능성이 있을 것이다. 이를 해결하기 위해 다양한 imputation 방법들을 사용해보고, 가장 좋은 방법을 이용하여 정확성을 높이려고 한다. 그리고 추가로 더 많은 데이터를 탐색해보고, 유의한 변수일 경우 추가하여 그 정확도를 높일 것이다.

2차 개인 과제에는 사용할 변수와 scaling 방법, Imputation 방법을 최종적으로 확정 짓고, 큰 상관관계를 보이거나 비슷한 변수들도 통합할 것이다. 그리고 그 변수들을 이용해, 4가지의 지표를 만들고 이를 이용해 linear model을 최종적으로 확정할 계획이다. 치킨집 사장님은 이 4가지 지표를 모니터링, 예측 하면서 치킨집 매출을 예측할 수 있고, 그에 따라 언제 쉬어야 할지, 재고를 어느정도 갖춰야 할지 등에 관한 결정에 도움을 줄 수 있을 것이다.