[ 인공지능 개발자 과정 - 1차 분석]

**가전제품 판매량 예측**

- 날씨와 SNS 언급량 데이터를 중심으로

[5조]

김민서

김민지

남현철

최은수

한찬규

**목 차**

# 서 론…………………………………………………………………………….…………. 3

* 1. 프로젝트 개요 ……………………………………………………………………….…….. 3
  2. 인력 구성…………………………………………………………………………...……3
  3. 프로젝트 일정…………………………………………………………………...………4
  4. 적용기술​…………………………………………………………………………….……4

# 본 론……………………...………………………………………………………….…….….5

* 1. 데이터 수집……………………...………………………………………………..….…..5
  2. 데이터 확인 및 전처리 ……………..……………………………………………...……6
  3. 최종 변수 선택 및 분석 모형 설계……………………………………………….…….10
  4. 예측 모델 성능 비교 …………………………………………………………………....11

# 결 론……………………...……………………………………………………….………….13

* 1. 프로젝트 의의…………………………………………..…………………………...…...13
  2. 모델 보완점………………………………………………………..…………………......13
  3. 진행 소감…………………………………………………………….……………...........14

별 첨) 발표 자료 ………………………………………………………………

# 서 론

## 프로젝트 개요

### 프로젝트 배경

최근 국내 기후환경은 역대급 무더위, 이른 장마, 3한 4미 등 용어가 등장할 정도로 급변하면서 관련 기사와 각종 문제들이 떠오르고 있습니다. 기상 환경들은 삶에 직접적인 영향을 미치는 만큼 소비자들의 의사결정에도 영향을 줄 것입니다.

정보통신정책연구원에 따르면, 2011년 16.8% 였던 SNS 이용률이 2018년 48.2%로 증가하는 등 SNS 이용률이 급증하고 있습니다. 동 연구원 "SNS의 이용과 개인의 사회관계 변화 분석" 보고서에 따르면 SNS를 이용한 대화에서 이슈에 대한 문제해결 도출 경험은 전 연령대에서 50% 이상입니다. 모바일 기기와 SNS의 확산은 우리 사회 변화에 영향을 미치는 주요 변수로 떠오르고 있습니다.

기상과 SNS의 각종 지표를 사용해 판매량을 사전에 예측해 다방면으로 활용할 수 있는 인사이트를 도출하고자 합니다.

### 프로젝트 목표

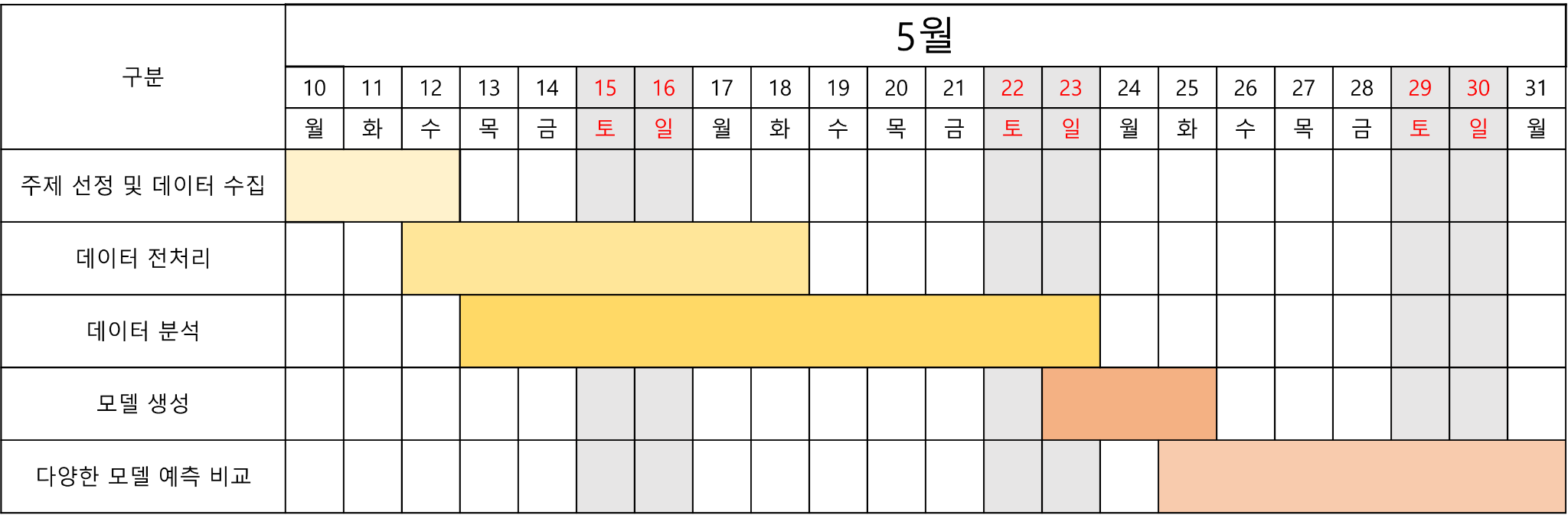
최근 3년간 폭염과 미세먼지 등은 다양한 사회문제를 만들어 낼 만큼 큰 이슈였습니다. 날씨와 비교적 연관성이 클 것이라고 생각되는 에어컨, 가습기, 제습기, 공기청정기를 선택했습니다. 강수량 등 기상 정보와 SNS의 언급 횟수가 소비자의 실제 수요에 어떤 영향을 얼마나 주는 지 확인하고 지도학습 기법을 사용해 판매량 예측 모델을 만들어 새로운 데이터가 주어졌을 때 실 판매량을 예측하는 것을 목표로 설정했습니다.

## 인력 구성



## 

## 프로젝트 일정



## 적용기술

| **구분** | **소프트웨어** | | **버전** |
| --- | --- | --- | --- |
| **Language** | **R** | | **4.0.5** |
| **IDE** | **R studio** | | **1.4.1106** |
| **라이브러리** | **시각화** | **ggplot2** | **3.3.3** |
| **corrplot** | **0.88** |
| **gridExtra** | **2.3** |
| **분석** | **randomForest** | **4.6-14** |
| **glmnet** | **4.1-1** |
| **caret** | **6.0-88** |
| **neuralnet** | **1.44.2** |
| **xgboost** | **1.4.1.1** |

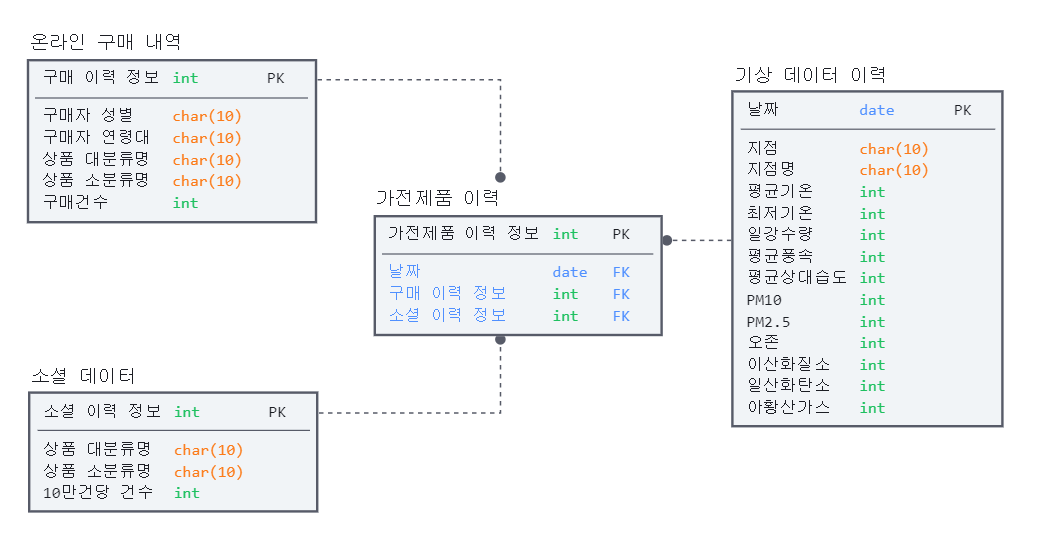
# 본 론

## 데이터 수집

소셜 데이터, 온라인 구매이력 데이터, 날씨 데이터를 날씨마루, 공모전과 기상청을 통해 수집했습니다. 모든 데이터는 2018.01.01 ~ 2019.12.31의 2년간 수집된 자료입니다.

먼저 소셜 데이터는 ㈜바이브컴퍼니에서 제공한 자료로 블로그, 커뮤니티, 인스타그램을 분석하여 각 품목별 10만건당 언급 건수를 변수로 포함하고 있으며 여기서 10만건당 언급건수는 서로 다른 유저 수를 보유한 채널별 문서수의 격차를 보정하기 위해 상대적 문서수로 나타낸 값입니다. 각 채널별의 버즈량을 10만건 대비한 문서수로 살펴봄으로써 채널간의 문서수 차이나 채널안에서 발생하는 전체 문서수의 변동성에 비해 안정적으로 해당 키워드에 해당하는 문서 수를 확인할 수 있습니다. 온라인 구매이력 데이터 구매변수 (구매 날짜, 구매자 성별, 구매자 연령대, 품목별 구매 건수), 날씨 데이터는 일별 날씨변수(평균기온, 최저기온, 최고기온, 일강수량, 평균풍속, 평균상대습도, PM10, PM2.5, 오존, 이산화질소, 일산화탄소, 아황산가스)들을 포함하고 있습니다.

주어진 데이터를 정규화한 ERD는 다음과 같습니다.



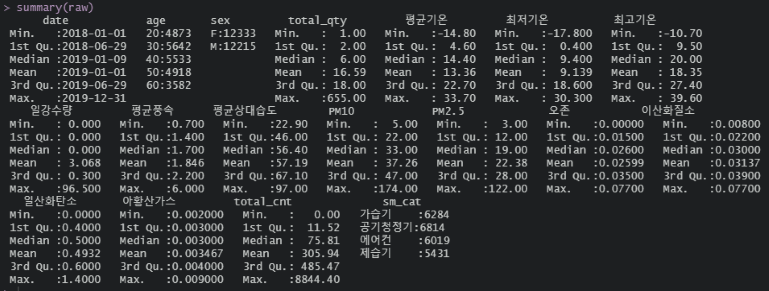
↓

**구매건수 (판매량) 예측**

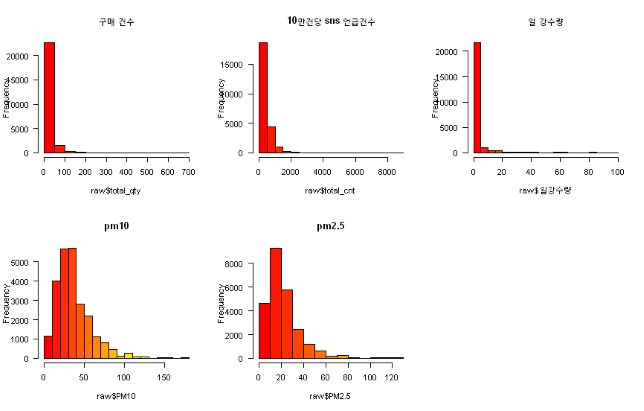
## 데이터 확인 및 전처리

#### 데이터 전처리

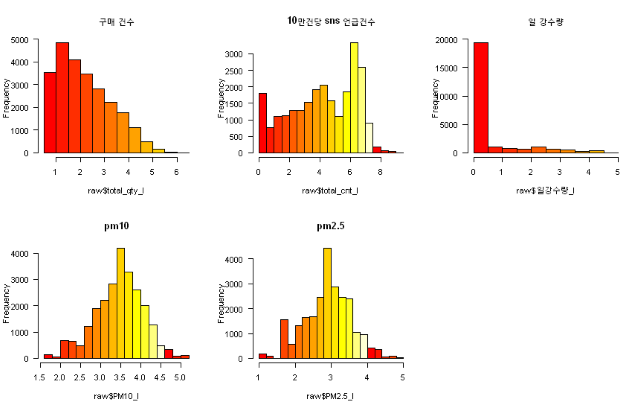
먼저 나눠진 3가지 데이터를 통합하였습니다. 데이터를 합치는 과정에서 발생한 결측치는 제거하였으며 date변수는 날짜형으로, 구매 성별, 구매 나이는 범주형 변수로 변경 후 summary를 통해 각 변수들의 특성을 확인했습니다.



위 결과를 보면 범주형 변수에서의 심한 데이터 불균형은 발생하지 않은 것으로 보이며, 판매 건수(total\_qty), 일강수량, 10만건당 언급건수(total\_cnt), 미세먼지(PM10, PM2.5)에서 치우침 및 이상치가 발생한 것으로 보입니다. 해당 데이터를 시각화해 확인하겠습니다.

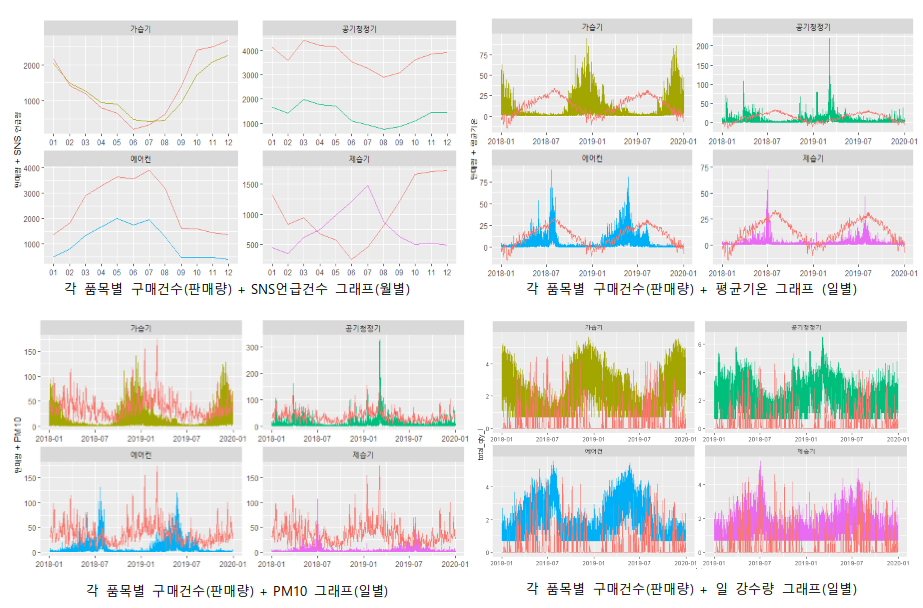


확인 결과 구매건수, SNS언급건수, 일 강수량에서는 심각한 치우침이, 미세먼지에서는 치우침이 발견되었습니다. 머신러닝 알고리즘은 정규분포나 가우시안 분포에서 더욱 쉽게 패턴을 찾을 수 있으므로 해당 변수들에 log를 취하여 분포 모양을 정규 분포에 가깝게 변경했습니다.

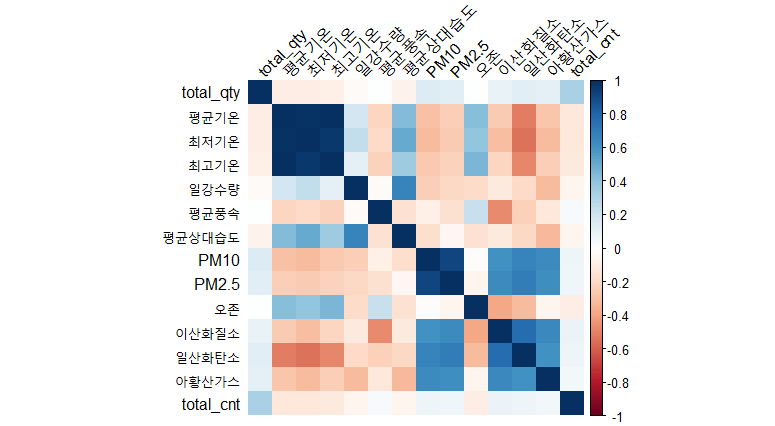
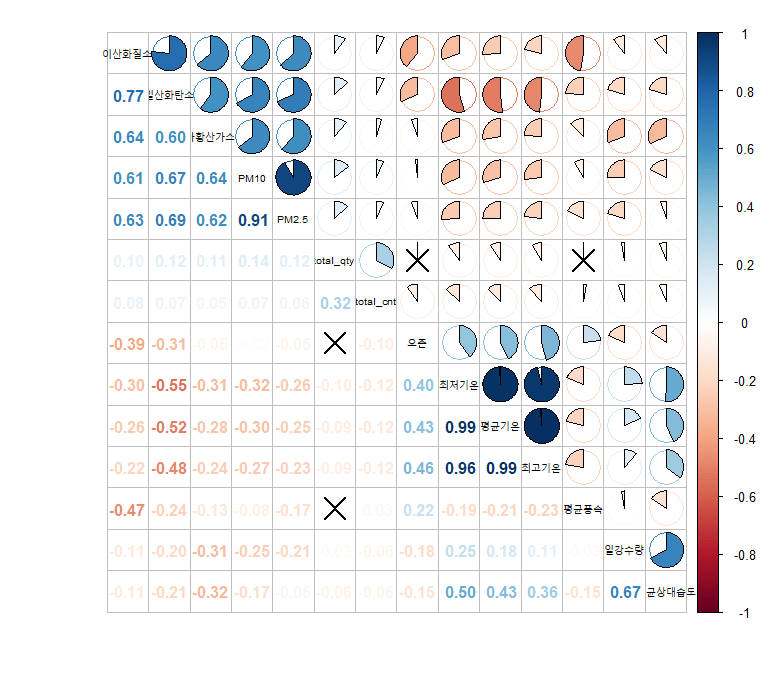


#### 탐색적 분석 (EDA)

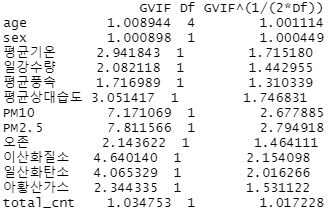
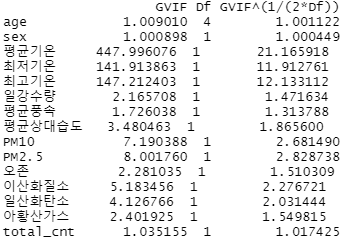
먼저 품목별 데이터의 대략적인 모양을 보고 상관관계 파악 및 필요한 파생변수를 추가하도록 하겠습니다. 전체적인 구매건수와 다른 변수들의 모양은 아래와 같습니다.



데이터들을 살펴본 결과 SNS언급량과 구매건수 사이에 밀접한 연관이 있어 보입니다. 또, 계절의 영향이 있다고 판단 season (봄: 3,4,5월, 여름: 6,7,8월, 가을: 9,10,11월, 겨울:12,1,2월)변수를 추가했습니다. 일별로 예측하기는 무리가 있다고 판단 월로 묶은 month변수를 추가하고 혹시 요일의 영향이 있는지를 보기 위해 weekday변수를 추가 했습니다.

그 다음 연속형 변수간 상관관계를 파악하기 위해 상관관계 그래프를 그려보았습니다. 

확인 결과 미세먼지와, 기온들 사이에서 상관관계가 0.8이상(왼쪽 그림의 하단부 수치가 스피어만 상관계수)으로 나오므로 다중공선성이 의심됩니다. 다중공선성 발생 여부를 수치적으로 확인해 보겠습니다.



모델을 생성하여 VIF를 확인해보니 기온과 미세먼지 관련 변수에서 10이 넘는 숫자가 나옵니다. 평균기온, 최저기온, 최고기온 중 평균을 선택하고 나머지 두 변수를 삭제합니다. 미세먼지의 경우 VIF는 10 이하 지만 5이상이며 스피어만 상관계수가 0.8이상이므로 PM2.5를 삭제하도록 하겠습니다.

또한 데이터에는 다양한 범주형 변수들이 있으므로 카이제곱 검정으로 각 범주형 변수 집단들의 독립성 검정을 실시해 범주형 변수들간의 상관관계를 파악합니다. 귀무가설은 “두 집단은 서로 독립이다.” 이고 대립가설은 “독립이 아니다.”입니다. 아래는 카이제곱 검정을 진행한 결과 나온 p-value입니다.

|  | **Age(나이)** | **Sex(성별)** | **Weekday(요일)** | **Month(구매월)** | **Season(계절)** |
| --- | --- | --- | --- | --- | --- |
| **Age(나이)** | - | **0.000431** | 0.9125 | 0.1752 | **0.002303** |
| **Sex(성별)** |  | - | 0.799 | 1 | 0.9889 |
| **Weekday(요일)** |  |  | - | **5.98E-15** | 0.1684 |
| **Month(구매월)** |  |  |  | - | **2.20E-16** |

위 표를 해석해보면, 나이와 성별, 계절은 상관이 있으며, 요일과 구매 월, 구매 월과 계절은 상관이 있습니다. 집단의 특성을 생각해보면, 구매 월과 계절은 비슷한 속성을 가지고 있으며, 두 집단의 p-value가 매우 작기때문에 구매 월 변수는 삭제하였습니다.

그러므로 최종적으로 사용될 독립변수는 나이, 성별, 구매건수, 평균기온, 일강수량, 평균풍속, 평균상대습도, PM10, 오존, 일산화탄소, 아황산가스, 10만건당 언급건수, 요일, 계절입니다. 설명의 편의성을 위해 앞으로는 구매건수 변수를 판매량이라고 지칭하겠습니다.

## 최종 변수 선택 및 분석 모형 설계

결측치 처리는 이미 진행하였기에 결측치는 신경쓰지 않도록 합니다. 범주형 변수들은 One Hot Encoding을 통하여 더미변수를 만들어 머신러닝 알고리즘에 사용할 수 있도록 변경하고 다양한 변수탐색 기법을 통해 각 품목별 가장 성능이 좋은 변수를 선별하여 모델에 적용해 보도록 하겠습니다.

변수선별에 사용한 기법은 선형 회귀법의 단계별 변수탐색법, Random Forest와 XGBoost의 importance이며, 예측모델을 생성할 품목이 4개이므로 각 품목별로 변수탐색을 진행하였습니다.

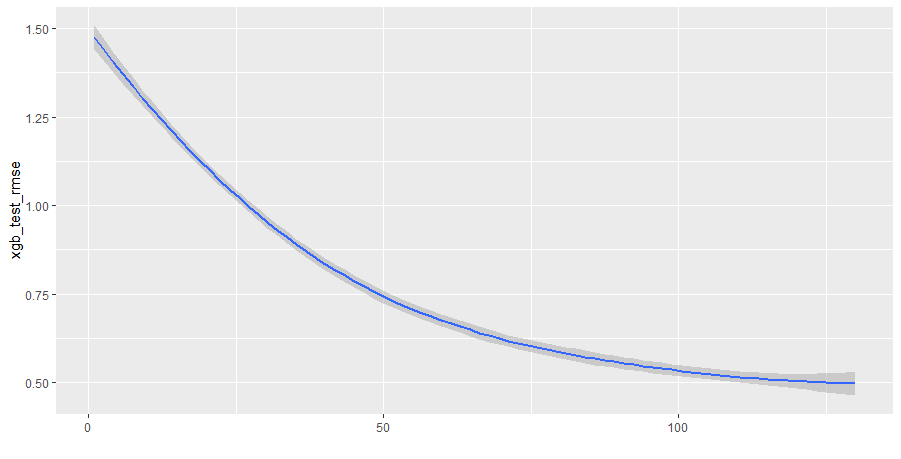
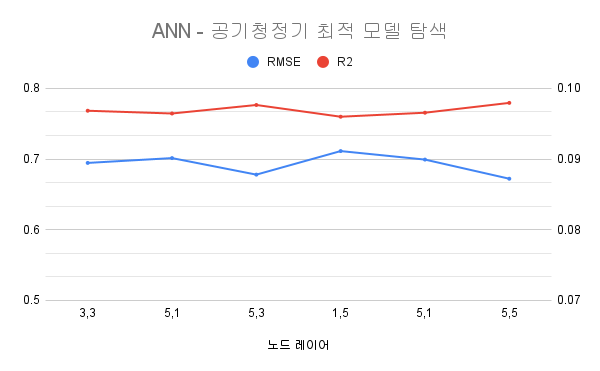
| XGBoost | Stepwise Linear Regression |
| --- | --- |
|  |  |
| RandomForest | |
|  | |

각 변수 선별과정에서 두 가지 이상 유의미하다고 나온 변수들로 추려 최종 모형을 선택하였습니다.

| 품 목 | 모 형 |
| --- | --- |
| 공기청정기 | 판매량 ~ 나이 + 성별 + 평균기온 + 일 강수량 + 평균풍속 + 평균상대습도 + PM10 + 일산화탄소 + 아황산가스 + SNS언급량 + weekday + season |
| 가습기 | 판매량 ~ 나이 + 성별 + 평균기온 + 평균풍속 + 평균상대습도 + PM10 + 오존 + 일산화탄소 + 아황산가스 + SNS언급량 + weekday + season |
| 에어컨 | 판매량 ~ 나이 + 성별 + 평균기온 + 일 강수량 + 평균풍속 + 평균상대습도 + PM10 + 오존 + +이산화질소 + 일산화탄소 + 아황산가스 + SNS언급량 + weekday + season |
| 제습기 | 판매량 ~ 나이 + 성별 + 평균기온 + 일 강수량 + 평균상대습도 + PM10 + 오존 +이산화질소 + 일산화탄소 + 아황산가스 + SNS언급량 + weekday + season |

## 예측 모델 성능 비교

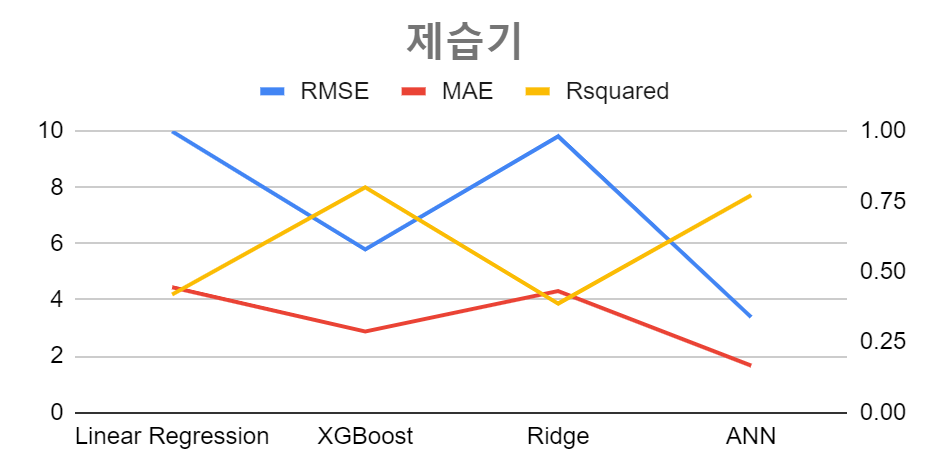
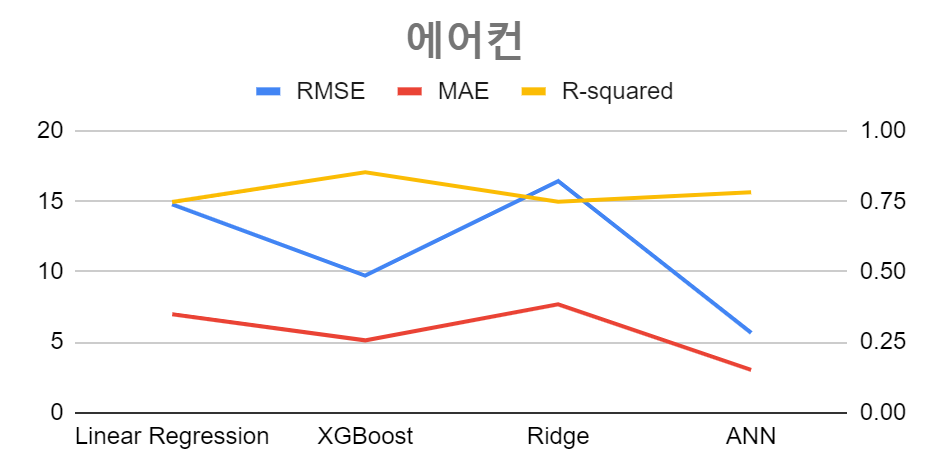
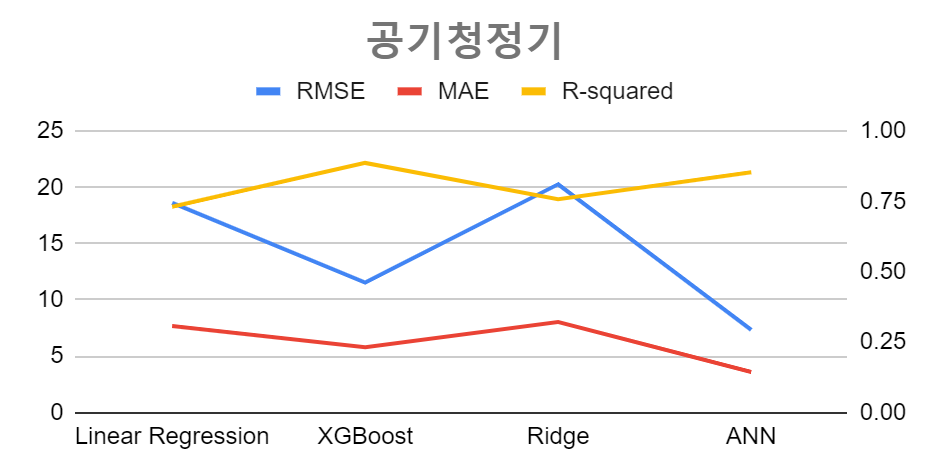
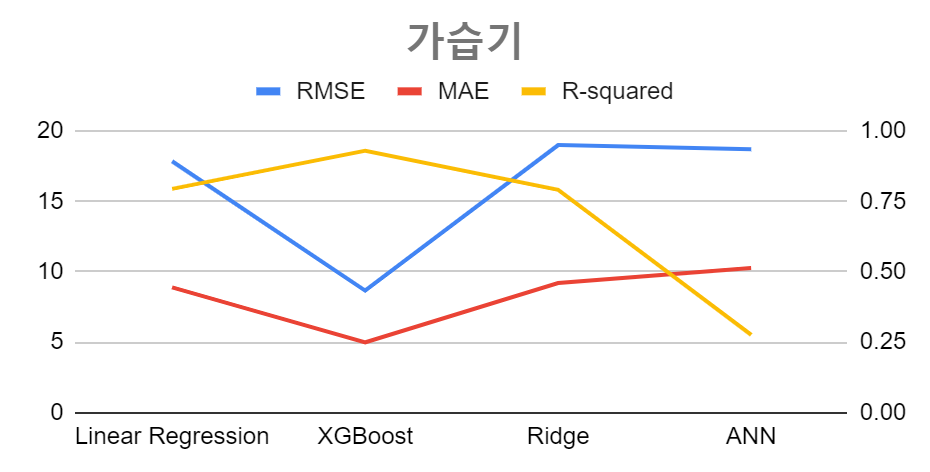
최종 모형을 선정하고 여러 모델을 학습시킨 후 예측 정확도를 확인했습니다. 연속형 수치의 예측이기 때문에 예측성능 확인 지표로는 RMSE, MAE, R-Square를 사용하였으며 트레이닝과 테스트 셋의 비율은 7:3으로 고정하였습니다.



각 모델 성능 비교 전에 튜닝이 가능한 XGBoost와 ANN은 하이퍼 튜닝을 통하여 가장 성능이 좋은 모델을 정하여 비교했습니다. XGB 튜닝 방식은 round별 Max.Depth랑 learning rate를 여러 조합으로 사용하여 가장 RMSE가 낮은 조합을 선택했고, ANN은 히든 레이어를 2개로 고정시킨 상태에서 노드수를 조정하여 최적의 모델을 탐색하였습니다. 최종 선정된 Hyperparameters는 아래와 같습니다.

|  | **XGBOOST** | | | **ANN** |
| --- | --- | --- | --- | --- |
|  | ROUND | MAX.DEPTH | LEARNING RATE | HIDDEN NODE |
| **가습기** | 50 | 6 | 0.24 | 5,5 |
| **공기청정기** | 50 | 5 | 0.28 | 5,5 |
| **에어컨** | 50 | 5 | 0.22 | 5,3 |
| **제습기** | 50 | 5 | 0.24 | 5,5 |

최종 선택된 hyper parameters를 사용하여 예측 모델을 만들어 성능비교 하였습니다.



| **공기청정기** | 구분 | RMSE | MAE | R-squared |
| --- | --- | --- | --- | --- |
| Linear Regression | 18.62321 | 7.686604 | 0.7311 |
| XGBoost | 11.5395889 | 5.7803879 | 0.8867911 |
| Ridge | 20.2695991 | 8.0348782 | 0.7578041 |
| ANN | 7.326714 | 3.594439 | 0.853645 |

| **가습기** | 구분 | RMSE | MAE | R-squared |
| --- | --- | --- | --- | --- |
| Linear Regression | 17.85284 | 8.888611 | 0.7945 |
| XGBoost | 8.654049 | 4.97334 | 0.93003 |
| Ridge | 19.0018578 | 9.1943096 | 0.7908842 |
| ANN | 18.71075 | 10.27014 | 0.2755635 |

| **에어컨** | 구분 | RMSE | MAE | R-squared |
| --- | --- | --- | --- | --- |
| Linear Regression | 14.78609 | 6.981016 | 0.7481 |
| XGBoost | 9.7250299 | 5.1220076 | 0.8538161 |
| Ridge | 16.4428324 | 7.6813986 | 0.7490489 |
| ANN | 5.651998 | 3.022738 | 0.7822558 |

| **제습기** | 구분 | RMSE | MAE | R-squared |
| --- | --- | --- | --- | --- |
| Linear Regression | 9.990518 | 4.450325 | 0.4195 |
| XGBoost | 5.7921421 | 2.8773874 | 0.8005077 |
| Ridge | 9.8168282 | 4.3146421 | 0.3867485 |
| ANN | 3.387199 | 1.659884 | 0.7722288 |

예측 결과를 보면 전체적으로 가장 설명력이 높은 모델은 XGBOOST 모델이고, 가장 RMSE 값이 작은 모델은 ANN 모델이었습니다. LM이나 Ridge모델의 성능은 상대적으로 낮게 나왔습니다. 품목 별로 보면 공기청정기는 XGBOOST에서 설명력이 가장 높고 ANN에서 RMSE값이 제일 낮았으나 XGBOOST에서도 충분히 낮은 값을 보이기 때문에 XGBOOST의 예측력이 가장 좋았습니다. 가습기와 에어컨의 경우 설명력과 RMSE값 모두 XGBOOST에서 가장 좋은 성능을 나왔으며, 제습기의 경우 XGBOOST에서 설명력이 가장 높고 ANN에서 RMSE값이 작게 나왔습니다.

# 결 론

## 프로젝트 의의

## 주어진 지표와 상관관계가 있음을 확인하였으나 기상에 큰 영향을 받을 것이라는 통념과 달리 기상정보와 상관관계는 크게 보이지 않았습니다.

## 반대로 SNS정보와 상관관계가 크게 나타남을 확인했고, 마케팅 활용에 있어 기상정보 이외에 소셜 데이터 활용한 분석또한 중요해질 것 입니다. 따라서 향후 가전제품 마케팅에 있어 현재 다음과 같은 과정을 제안합니다.

1. SNS 데이터를 통한 customer analysis, 자사의 현황 분석과 타사의 데이터 활용 여부를 파악하여 시장조사를 실시합니다.
2. 마케팅 대상이 될 SNS의 데이터를 분석하여 segmentation을 수행하고 표적시장이 될 시장(target SNS 혹은 커뮤니티)를 선정하여 앞서 분석한 고객 데이터를 기반으로 제품의 Positioning을 합니다.
3. 판촉 방법을 구상(SNS follower 활용 혹은 SNS 광고 등)하고 여기에 기상데이터를 활용하여 적절한 Price의 Product를 예측 모델을 통해 선정하여 효율적인 유통방식을 구상합니다.

즉, 기존의 온라인 경로를 통한 유통에서 SNS와 기상정보를 활용한 예측모델 (인공지능)을 활용하면 마케팅 전략 수립 과정에서의 효율성을 증가시키며 특히 고객 분석 단계와 유통 단계에서 특히 효율성을 증대시킬수 있는 방법일 것입니다.

## 모델 보완점

1. ANN 에서의 특정 품목(가습기) 에서의 설명력이 굉장히 낮게 나왔으나 원인을 파악하지 못하였습니다. 원인을 파악하여 해결한다면 더 높은 설명력을 기대해 볼 수 있을 것입니다.
2. 상품군에 대한 분석이므로 특정 상품에 대해서 적용하기 위해서는 각 상품에 맞게 추가적인 변수 튜닝이 필요할 것으로 보입니다.
3. SNS데이터의 구분이 10만건당 언급량(cnt) 뿐이기에 해당 상품군에 대한 정확한 여론(긍정,부정)이 조사되어있지 않아 해당 언급량이 긍정적인 영향을 주었는지 부정적인 영향을 주었는지에 대한 분석이 어려워 언급량이 높았음에도 판매가 이루어 지지 않은 제품들에 대한 해석이 부족했습니다.
4. 범주형 변수들 중 상관성이 있다고 나온 변수가 있었으나 마케팅 활용 가능성으로 인해 모두 제거하지 못하였고, 기술적 문제로 범주형 변수와 연속형 변수의 상관관계를 확인하지 못하였습니다. 이로인해 LM의 성능이 예상보다 높게 나왔다고 생각됩니다.

## 진행 소감

## 

## **<부록>**

## 

변수 선별 : lm, xgboost 돌려서 중요 변수 선정

⇒ 변수 선정 결과 그래프로 확인하기

\*\* lm 돌려서 나온 그래프 (결과)

\*\* xgboost 돌려서 나온 그래프 (결과)

<1차>

| **공기청정기** | **lm** | **xgboost** | **randomForest** |
| --- | --- | --- | --- |
| 성별 | \*\*\* | 11 | 99.96345 |
| 나이 | \*\*\* | 2,3,4,6,7 | 2658.90614 |
| 평균기온 | \*\*\* | 5 | 759.88982 |
| 강수량 |  | 14 | 68.84819 |
| 풍속 |  | 12 | 178.70081 |
| 습도 | \*\*\* | 8 | 268.73557 |
| PM10 | \*\*\* | 9 | 500.56592 |
| 오존 | \*\*\* | 10 | 260.76192 |
| 이산화질소 | \* | 13 | 223.97830 |
| 일산화탄소 | \*\* | 15 | 185.04725 |
| 아황산가스 | \*\*\* | 16 | 80.26850 |
| total\_cnt | \*\*\* | 1 | 2039.77435 |
| RMSE |  |  |  |
| MAE |  |  |  |
| R^2 |  |  |  |

| **가습기** | **lm** | **xgboost** | **randomForest** |
| --- | --- | --- | --- |
| 성별 | \*\*\* | 5, | 204.18382 |
| 나이 | \*\*\* | 3,4,6,7,11 | 2252.53310 |
| 평균기온 | \*\*\* | 2 | 1850.46580 |
| 강수량 |  | 16 | 77.79523 |
| 풍속 | \*\*\* | 13 | 158.08734 |
| 습도 |  | 8 | 305.92103 |
| PM10 | \*\*\* | 12 | 208.33335 |
| 오존 | \*\*\* | 9 | 705.91100 |
| 이산화질소 | \*\*\* | 10 | 207.56704 |
| 일산화탄소 | \*\*\* | 14 | 304.42142 |
| 아황산가스 | \* | 15 | 87.85355 |
| total\_cnt | \*\*\* | 1 | 2420.52884 |
| RMSE |  |  |  |
| MAE |  |  |  |
| R^2 |  |  |  |

| **제습기** | **lm** | **xgboost** | **randomForest** |
| --- | --- | --- | --- |
| 성별 | \*\*\* | 10,13 | 92.42599 |
| 나이 | \*\*\* | 4,5,6,15 | 649.20721 |
| 평균기온 | \*\*\* | 1 | 654.74405 |
| 강수량 | \*\* | 12 | 144.04835 |
| 풍속 |  | 9 | 181.50858 |
| 습도 | \*\*\* | 2 | 435.29330 |
| PM10 | \*\*\* | 7 | 206.42810 |
| 오존 | \* | 8 | 262.13525 |
| 이산화질소 |  | 11 | 164.85125 |
| 일산화탄소 | \*\*\* | 14 | 119.34254 |
| 아황산가스 | \* | 16 | 56.09878 |
| total\_cnt | \*\*\* | 3 | 495.79221 |
| RMSE |  |  |  |
| MAE |  |  |  |
| R^2 |  |  |  |

| **에어컨** | **lm** | **xgboost** | **randomForest** |
| --- | --- | --- | --- |
| 성별 | \*\*\* | 11 | 204.18382 |
| 나이 | \*\*\* | 2,3,5,9,10 | 2252.53310 |
| 평균기온 | \*\*\* | 4 | 1850.46580 |
| 강수량 | \* | 16 | 77.79523 |
| 풍속 | \*\*\* | 12 | 158.08734 |
| 습도 | \*\*\* | 7 | 305.92103 |
| PM10 |  | 6 | 208.33335 |
| 오존 |  | 8 | 705.91100 |
| 이산화질소 | \*\*\* | 13 | 207.56704 |
| 일산화탄소 | \*\*\* | 15 | 304.42142 |
| 아황산가스 | \*\*\* | 14 | 87.85355 |
| total\_cnt | \*\*\* | 1 | 2420.52884 |
| RMSE |  |  |  |
| MAE |  |  |  |
| R^2 |  |  |  |

가설 설정 : 선별된 변수가 판매량에 영향을 미칠 것이다, 아니다 를 가설로 세우고

t, f test 통해서 통계적으로 유의미한 변수인지 확인

모델 설계: 각각 품목별로 모델 설계. (y = wx1+ ~)

**<2차>**

| **공기청정기** | **lm** | **xgboost** | **Ridge** |
| --- | --- | --- | --- |
| 나이 | \*\*\* |  |  |
| 평균기온 | \*\*\* |  |  |
| 습도 | \*\*\* |  |  |
| PM10 | \*\*\* |  |  |
| total\_cnt | \*\*\* |  |  |
| season1 |  |  |  |
| season2 | \*\*\* |  |  |
| season3 | \*\*\* |  |  |
| season4 | \*\*\* |  |  |
| weekday월요일 | \*\*\* |  |  |
| weekday화요일 | \* |  |  |
| weekday수요일 | \*\*\* |  |  |
| weekday목요일 | \*\* |  |  |
| weekday금요일 |  |  |  |
| weekday토요일 |  |  |  |
| weekday일요일 | \*\* |  |  |
| RMSE | 39.69684 |  | 0.8370764 |
| MAE | 37.12935 |  | 0.6913175 |
| R^2 | 0.4105 |  | 0.4117071 |

| **가습기** | **lm** | **xgboost** | **Ridge** |
| --- | --- | --- | --- |
| 성별 | \*\*\* |  |  |
| 나이 | \*\*\* |  |  |
| 평균기온 | \*\*\* |  |  |
| 습도 | \*\*\* |  |  |
| 오존 | \*\*\* |  |  |
| 이산화질소 | \* |  |  |
| 일산화탄소 | \*\*\* |  |  |
| total\_cnt | \*\*\* |  |  |
| season1 |  |  |  |
| season2 |  |  |  |
| season3 | \*\*\* |  |  |
| season4 | \*\*\* |  |  |
| weekday월요일 | \*\*\* |  |  |
| weekday화요일 | \*\*\* |  |  |
| weekday수요일 | \*\* |  |  |
| weekday목요일 | \*\* |  |  |
| weekday금요일 |  |  |  |
| weekday토요일 | \*\*\* |  |  |
| weekday일요일 | \*\*\* |  |  |
| RMSE | 38.76346 | 0.3947035 -  0.3838442 |  |
| MAE | 36.19382 | 0.3051030 - 0.2955869 |  |
| R^2 | 0.6623 | 0.8936165 - 0.8993582 |  |

| **제습기** | **lm** | **xgboost** | **Ridge** |
| --- | --- | --- | --- |
| 성별 | \*\*\* |  |  |
| 나이 | \*\*\* |  |  |
| 평균기온 | \*\*\* |  |  |
| 습도 | \*\*\* |  |  |
| PM10 | \*\*\* |  |  |
| total\_cnt | \*\*\* |  |  |
| season1 |  |  |  |
| season2 |  |  |  |
| season3 | \*\*\* |  |  |
| season4 | \*\*\* |  |  |
| weekday월요일 | \*\*\* |  |  |
| weekday화요일 | \*\*\* |  |  |
| weekday수요일 | \*\*\* |  |  |
| weekday목요일 | \*\* |  |  |
| weekday금요일 |  |  |  |
| weekday토요일 | \*\*\* |  |  |
| weekday일요일 | \*\*\* |  |  |
| RMSE | 38.48313 |  | 0.7129088 |
| MAE | 35.97348 |  | 0.5680962 |
| R^2 | 0.6494 |  | 0.2664314 |

| **에어컨** | **lm** | **xgboost** | **Ridge** |
| --- | --- | --- | --- |
| 나이 | \*\*\* |  |  |
| 평균기온 | \*\*\* |  |  |
| 습도 | \*\*\* |  |  |
| 오존 | \*\*\* |  |  |
| total\_cnt | \*\*\* |  |  |
| season1 |  |  |  |
| season2 | \*\*\* |  |  |
| season3 | \*\*\* |  |  |
| season4 |  |  |  |
| weekday월요일 | \* |  |  |
| weekday화요일 |  |  |  |
| weekday수요일 |  |  |  |
| weekday목요일 |  |  |  |
| weekday금요일 |  |  |  |
| weekday토요일 | \*\* |  |  |
| weekday일요일 | \*\* |  |  |
| RMSE | 39.38569 |  | 0.5956524 |
| MAE | 37.09195 |  | 0.4726452 |
| R^2 | 0.5809 |  | 0.7378464 |

모델링 : 에러를 사용해서 어떤 모델이 더 적합한지 비교

\*\* predict \_y , y 그래프

\*\* 모델 별 에러 정리한 표

예측값

|  | 변수뽑기 | tmp | d | xgboost | maxdepth | eta |
| --- | --- | --- | --- | --- | --- | --- |
| 에어컨 | 5, 7, 8, 13, 14, 20, 21, 22, 24 | 1,8,9 | 2:6 | :21 | 4 | 0.23 |
| 가습기 | 5, 6, 7, 8, 13, 16, 17, 18,20,21, 22, 24 | 1,2,11,12 | 3:9 | :25 | 5 | 0.2 |
| 제습기 |  |  |  |  |  |  |
| 공기청정기 |  |  |  |  |  |  |