

시계열 분석 모델을 통한 매장 식료품 판매 예측

KAGGLE PREDICTION COMPETITION

아이티윌

서동현 유재현
이승원 문혜선

프로젝트 일정

프로젝트 전체 일정표 입니다

역할 분담

	담당업무	공동업무
서동현 (조장)	프로젝트 총괄 담당모델 : Light GBM	
유재현 (팀원)	이론 스터디 주관 및 발표 담당모델 : SARIMAX	- 시각화 - 모델 설계 및 구현
이승원 (팀원)	PPT 제작 리눅스 / 윈도우 GPU 작업환경 구현	
문혜선 (팀원)	데이터 시각화 담당모델 : XGBoost	

목차

01. 개요

매장 매출 예측의 필요성

데이터 탐색

02. 진행

데이터 분석

시계열 분석

모델 생성

03. 결과

모델 선정

결과 테이블

결과 지표

진행

- 1. 분석 개요 및 데이터 설명**
- 2. 환경구축**
- 3. 데이터 EDA 및 전처리**
- 4. 모델링**
 - 시계열 분석
 - 신경망 모델
 - 앙상블 모델
- 5. 모델 결과 및 케글 등수**

매장 매출 예측의 필요성

매장 매출에 있어 예측이 필요한 이유에 대한 설명입니다

예측 전

- 과잉 재고 발주 및 보관 **비용 발생**
- 유통기한 이전 미소진 시 **폐기**
- 잦은 품절 시 **고객 불만 발생**

주관적인 판단에 의존한
비효율적 매장 관리

예측 후

- 효율적인 재고 관리로 **보관비용 절감**
- 적정한 순환을 통한 식료품 **신선도 유지**
- 품절 방지를 통한 **고객 만족도 향상**

데이터 기반 다양한 변수를 고려한
자동화된 예측으로 **효율성 증대**

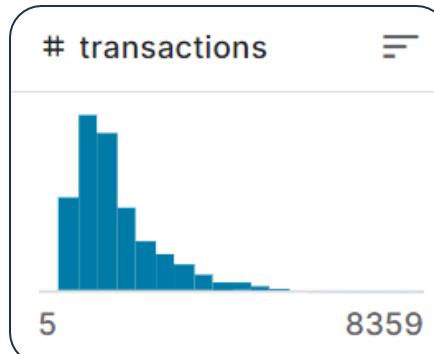
kaggle 데이터 설명

Kaggle에서 제공된 데이터 종류와 형태에 대한 설명입니다

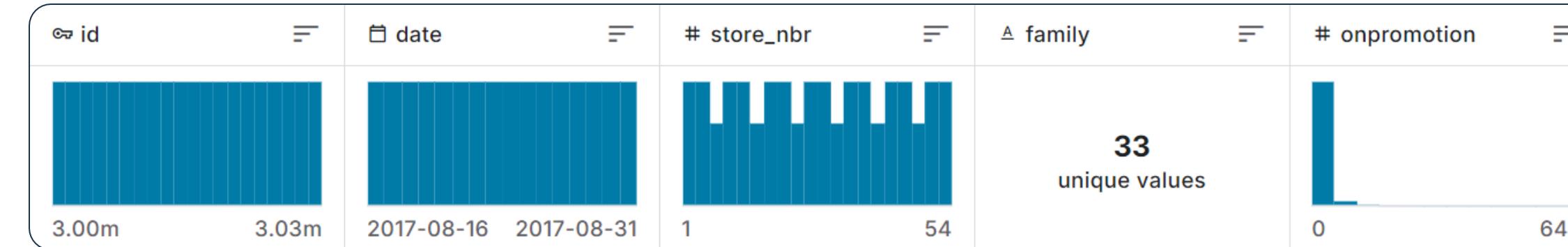
참고 문구

- holidays_events.csv 파일의 transferred 컬럼은 이벤트가 다른 날짜로 이동되었는지 여부를 나타낸다.. 예를 들어, 특정 휴일이 주말과 겹쳐 다음 평일로 이동된 경우 transferred 값이 1이 될 수 있다.
- oil.csv 파일의 유가 데이터는 에콰도르의 경제 상황에 영향을 미치는 중요한 요소이다.
- Stores.csv 파일의 cluster 컬럼은 유사한 특성을 가진 매장들을 그룹화한 것으로, 모델링 시 유용하게 활용될 수 있다.

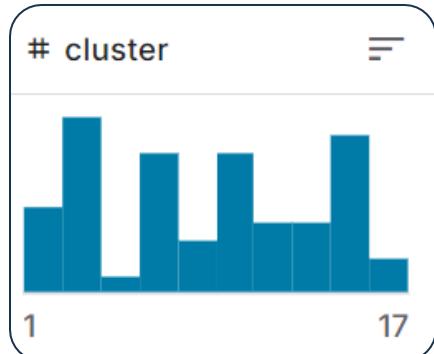
transcation



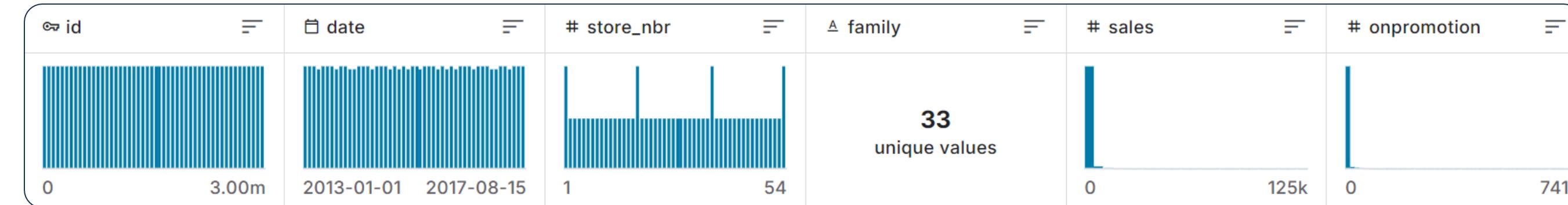
test



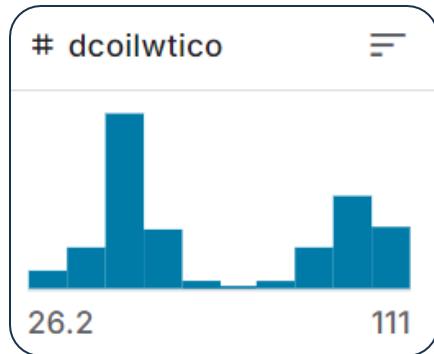
stores



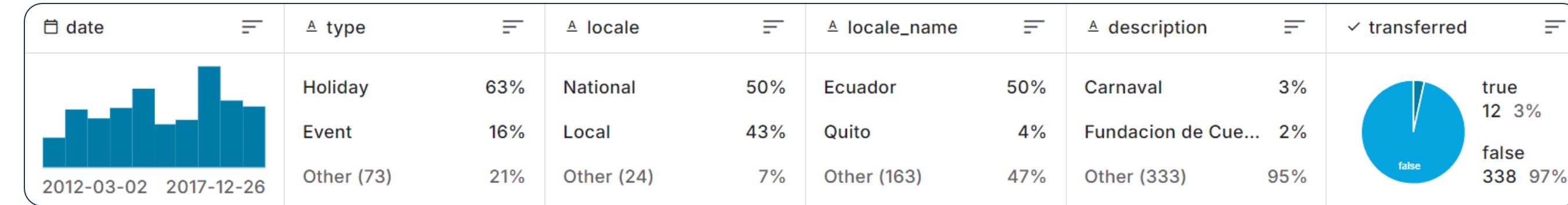
train



oil



holiday_events



진행

1. 분석 개요 및 데이터 설명
- 2. 환경구축**
3. 데이터 EDA 및 전처리
4. 모델링
 - 시계열 분석
 - 신경망 모델
 - 앙상블 모델
5. 모델 결과 및 케글 등수

분석 환경 구축

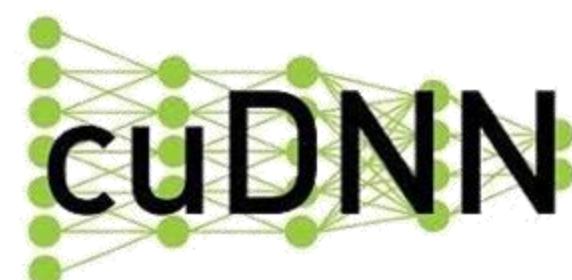
분석을 위한 GPU 구축 환경입니다.

하드웨어

- GPU: NVIDIA GTX 1080
- CPU: Intel Core i5

소프트웨어

- 운영 체제: Linux CentOs 7 / Windows 10
- 프로그래밍 언어: Python 3.8
- 딥러닝 프레임워크: PyTorch
- GPU 드라이버 및 라이브러리
 - NVIDIA GPU Driver 550.120
 - CUDA 11.7
 - cuDNN 8.5



진행

1. 분석 개요 및 데이터 설명

2. 환경구축

3. 데이터 EDA 및 전처리

4. 모델링

- 시계열 분석

- 신경망 모델

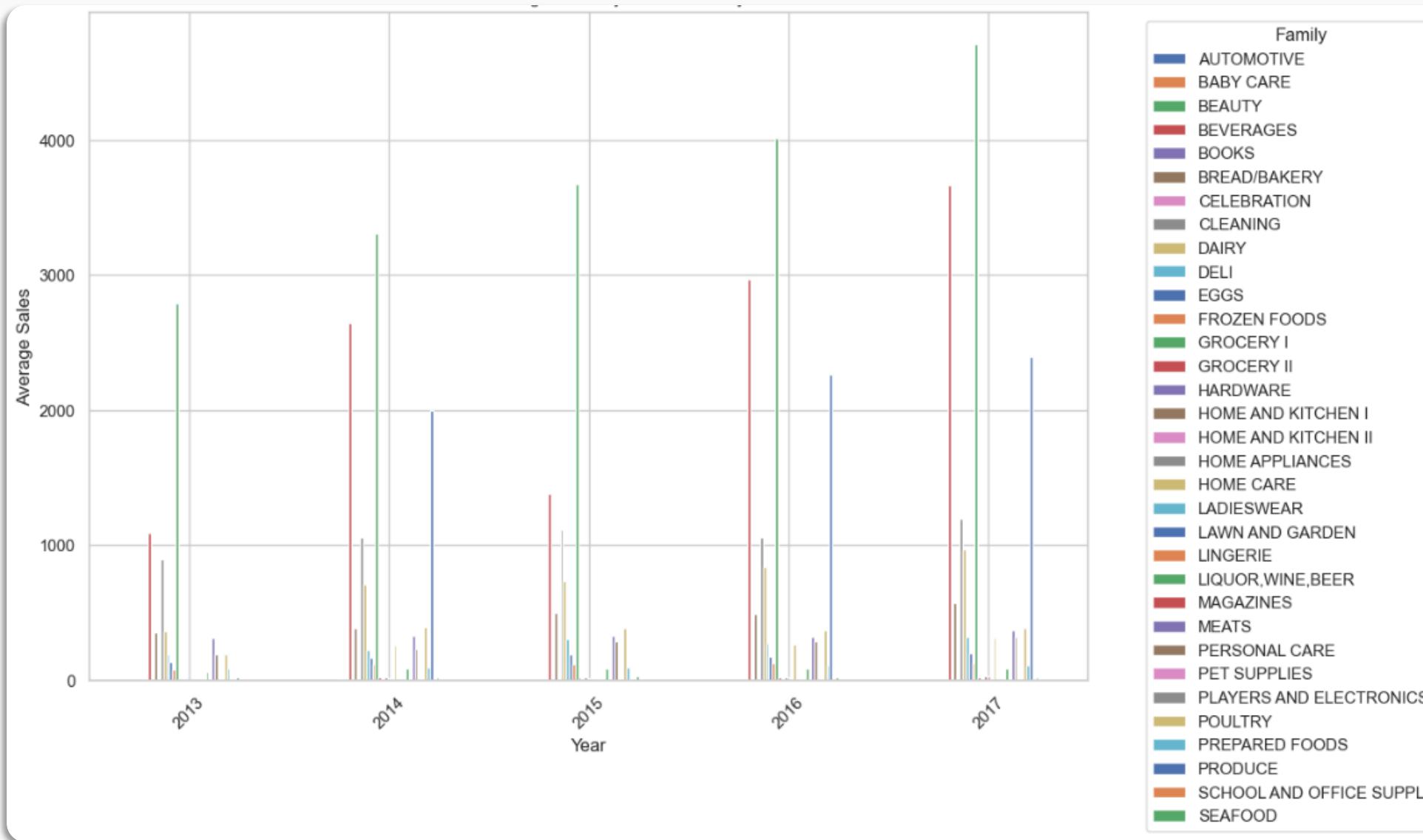
- 앙상블 모델

5. 모델 결과 및 케글 등수

데이터 분석

지진 발생 [2016년 4월]을 기준으로 판매량 분석 페이지입니다.

2013년부터 2017년도 까지 매장 제품군의 판매량 전체 비교 그래프



그래프 분석

생필품 관련 제품군:

[Cleaning, Grocery, Frozen Foods](#) 등의 생필품 관련 제품군은 재난 상황에서 수요가 증가 했을 가능성이 있습니다. 특히, 이러한 제품군의 판매 증가는 지진 직후 필요 물품을 구매하려는 행동과 관련이 있을 수 것으로 분석 됩니다.

비필수 소비재:

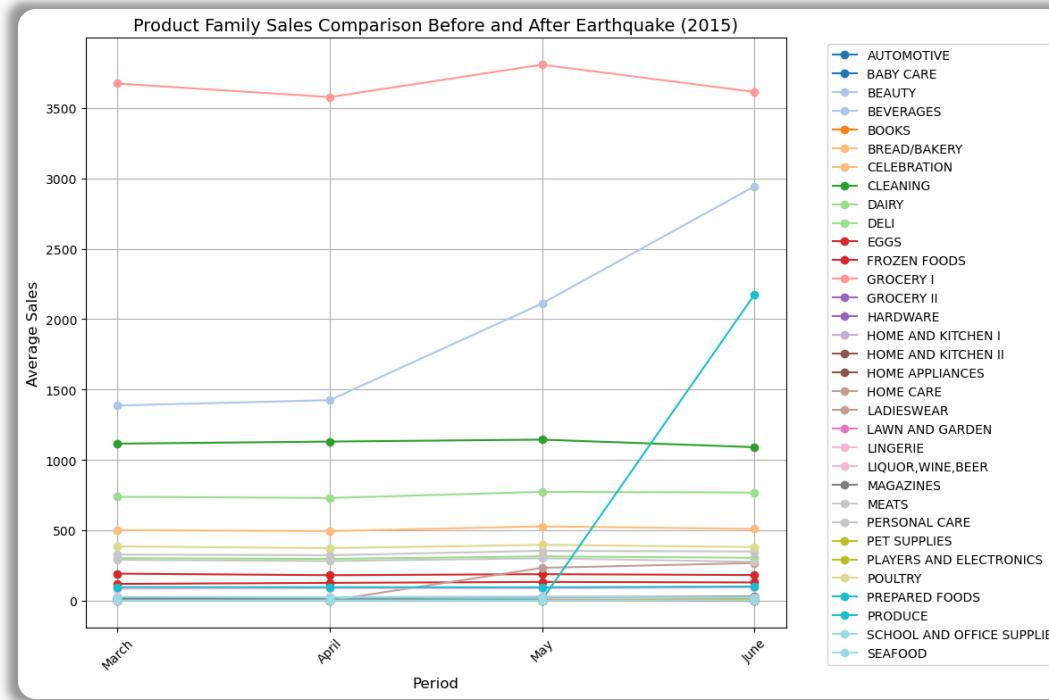
[Books, Lawn and Garden, Celebration](#) 등의 제품군은 지진 이후 수요가 감소 했을 가능성이 있으며, 이는 경제적 불확실성과 긴급하지 않은 소비를 줄이려는 경향과 관련이 있을 것으로 분석 됩니다.

데이터 분석

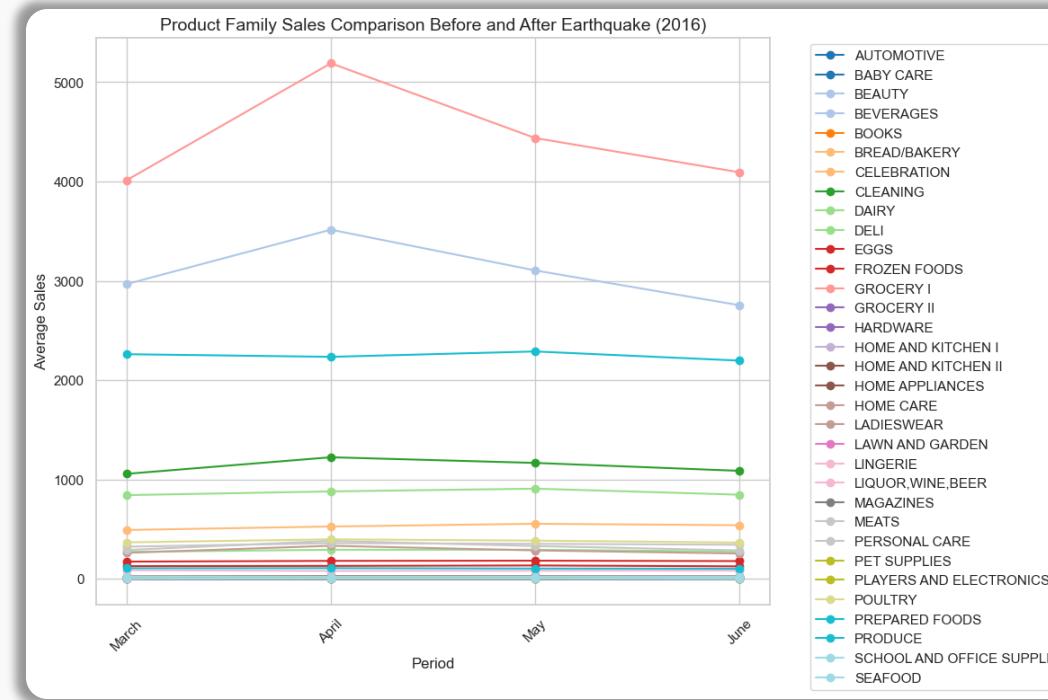
지진 발생 [2016년 4월]을 기준으로 판매량 분석 페이지입니다.

지진 전, 지진 발생, 지진 발생 후의 연도별 비교 그래프

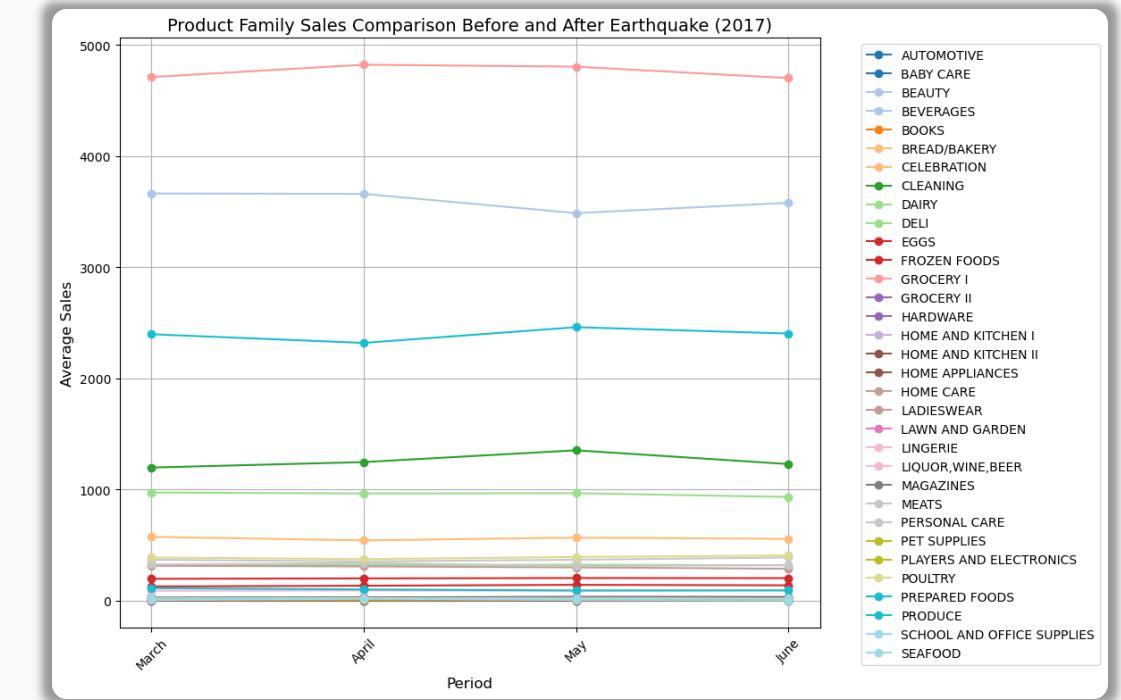
2015년 지진 전 그래프



2016년 지진 발생 그래프



2017년 지진 발생 후 그래프



그래프 분석

2015년 그래프에서 안정적인 판매량을 보여주며, 대부분의 제품군이 일정한 트렌드를 유지하고 있는 것으로 분석 됩니다.

2016년 그래프 (지진 발생 시점)에서는 일부 제품군에서 판매량이 급감하거나 급증하는 패턴이 나타납니다.

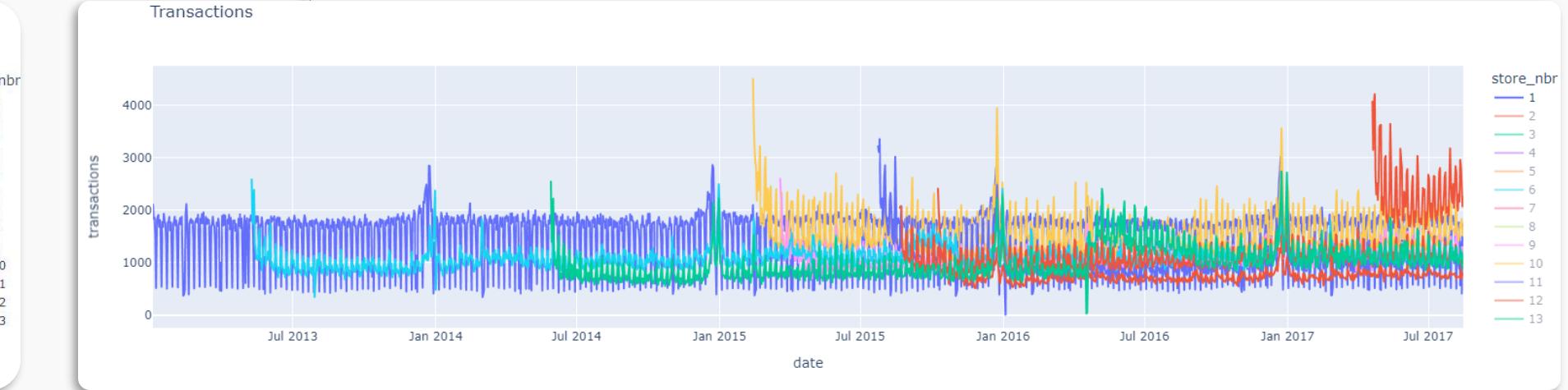
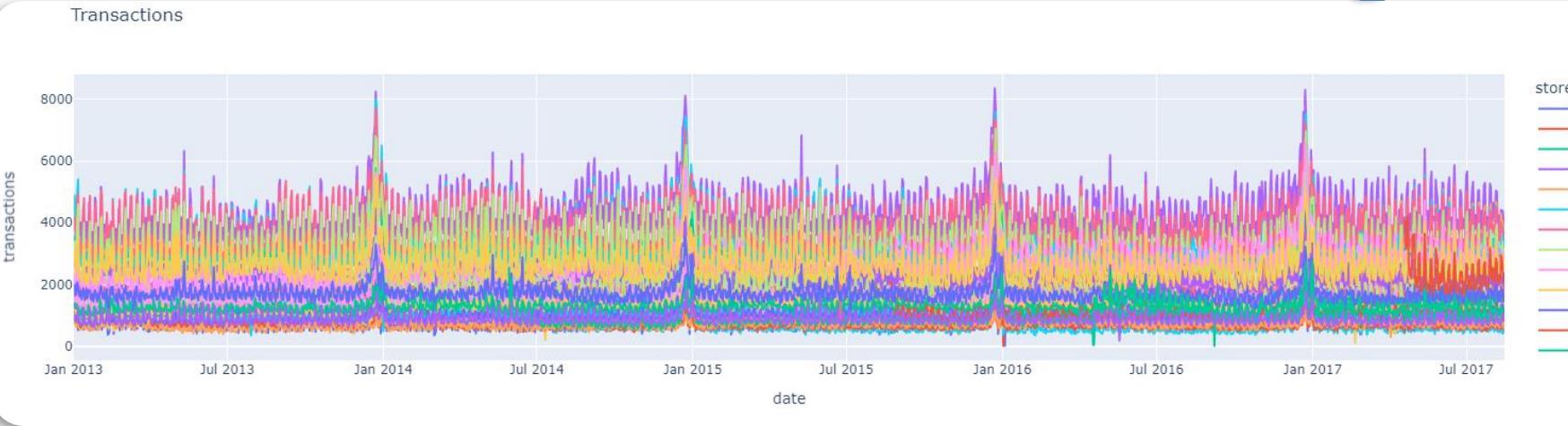
특히 필수 소비재(예: 음식, 물품 등)의 판매량이 급증한 것으로 보이며 이는 재난 대비로 인해 관련 제품 구매가 증가했을 가능성으로 보여집니다.

2017년 그래프 (지진 이후)에서는 2016년 대비 전반적인 판매량이 안정화되는 경향을 보이며 일부 제품군에서는 지진 이전보다 판매량이 감소하는 추세로 분석됩니다.

데이터 분석

제공된 데이터들을 분석한 결과입니다.

일자별 거래 건수 그래프



[가게별 개업 일자의 차이를 확인]

오일 거래량 결측치 처리 전후 비교 그래프



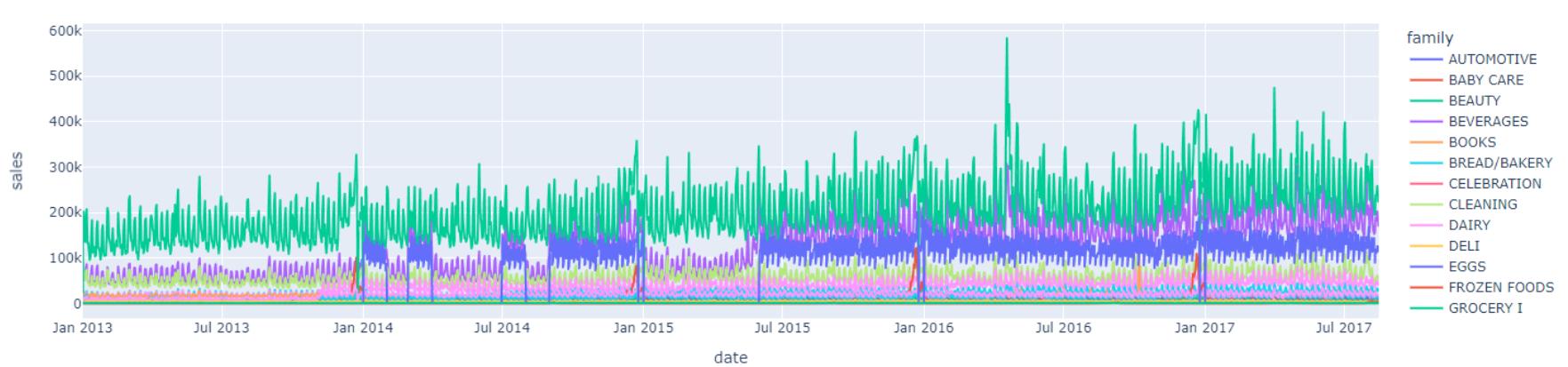
[오일데이터의 경우 사이 사이가 끊어져있는 것을 발견하여 연결해주는 처리]

데이터 분석

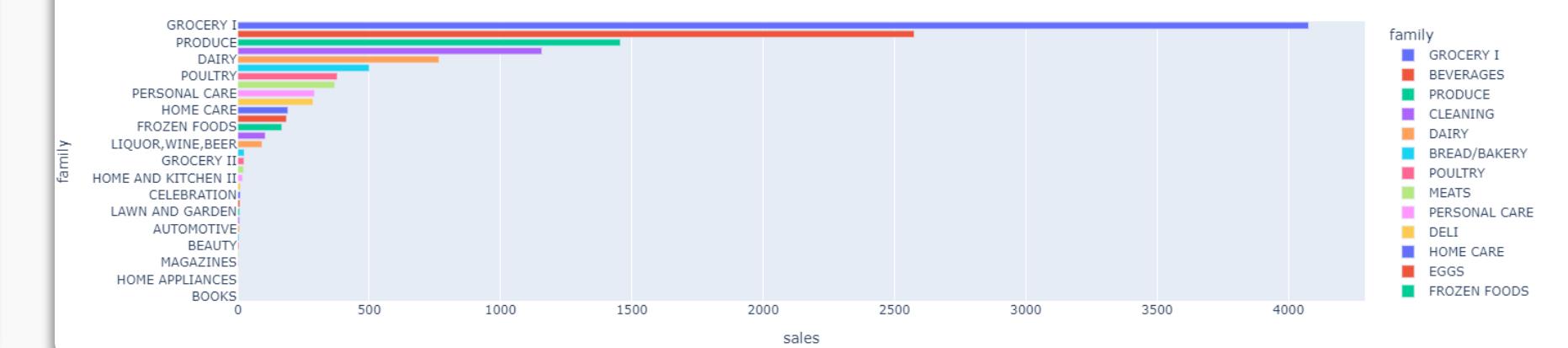
제공된 데이터들을 분석한 결과입니다.

제품군별 sales 데이터 그래프

Daily total sales of the family



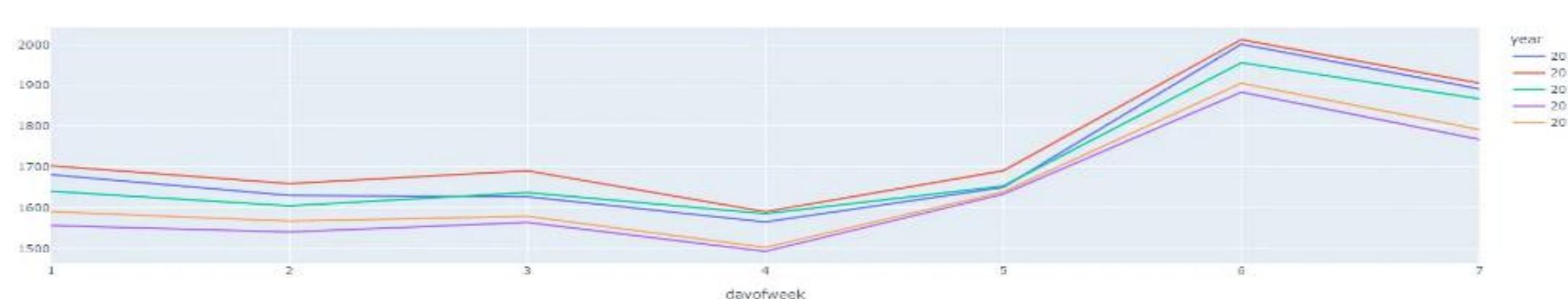
Which product family preferred more?



[Grocery1, Produce, Dairy 순으로 판매량이 높은 것을 확인]

요일별 거래건수 그래프

transactions

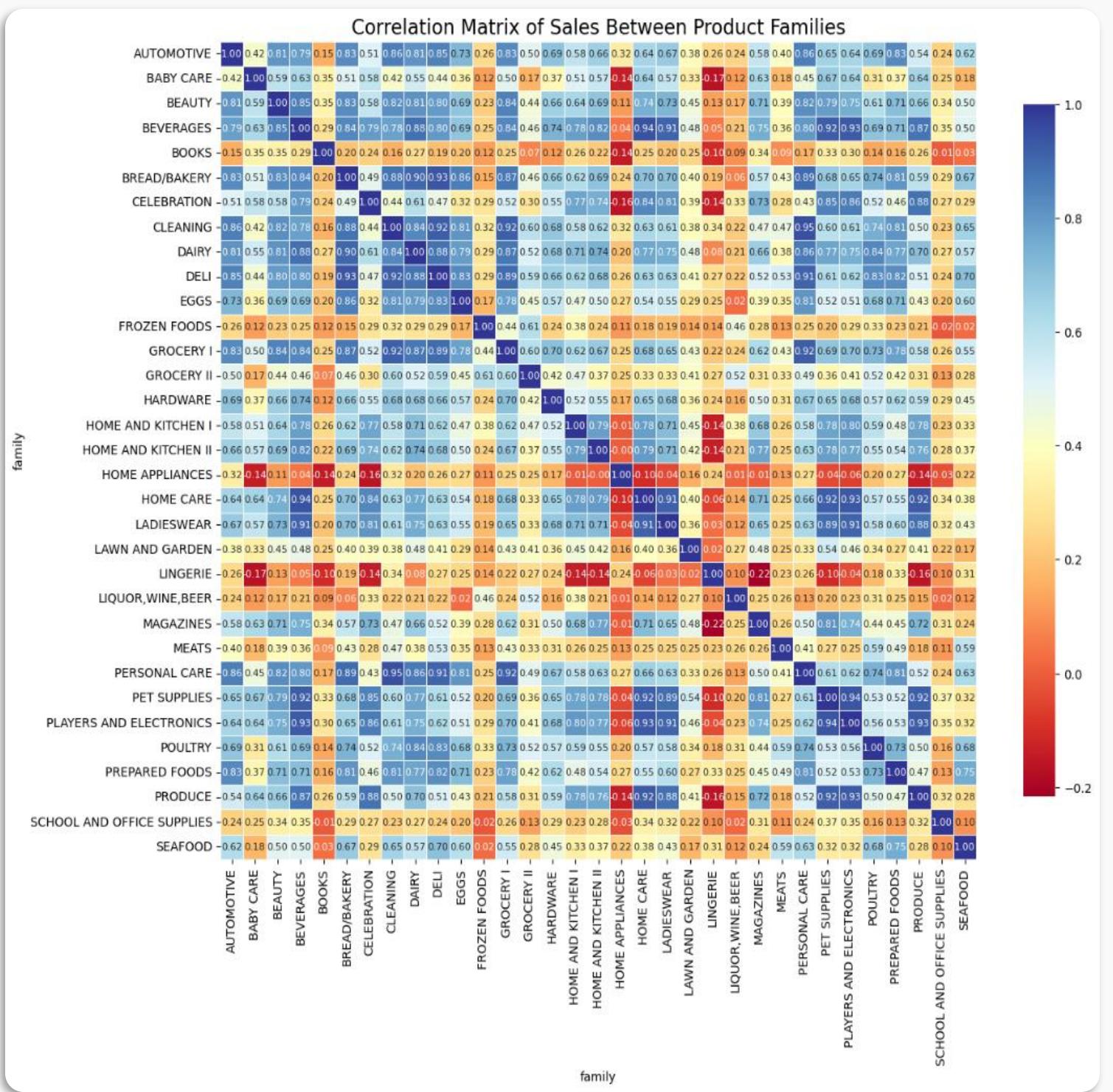


[주로 주말에 거래량이 급등하는 것을 확인]

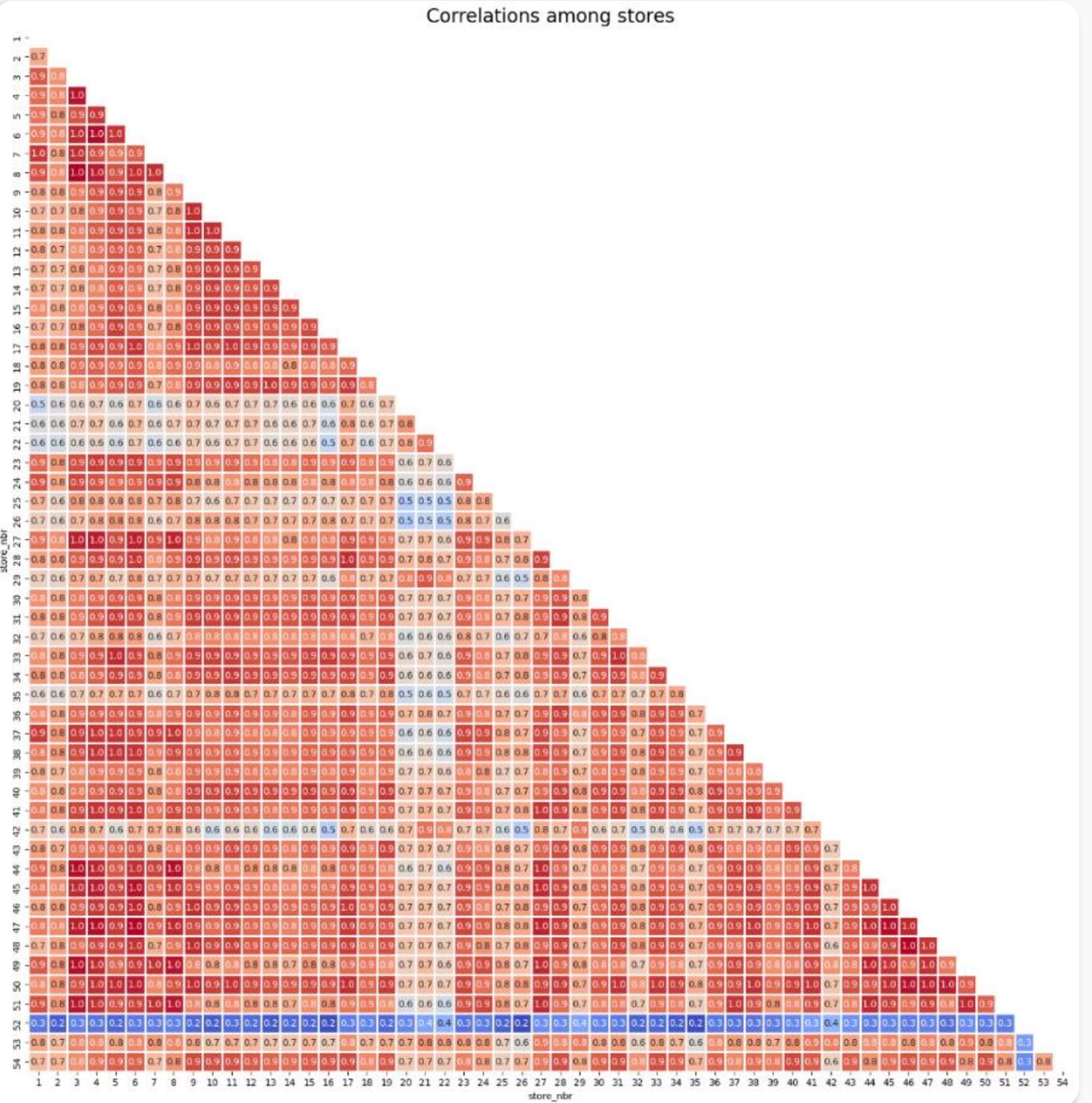
데이터 분석

제공된 데이터들을 분석한 결과입니다.

제품간 매출 상관관계



매장간 상관관계



데이터 전처리

제공된 데이터들 분석을 바탕으로 한 데이터 전처리 과정입니다

Train.csv

- 없는 상품/ 판매하지 않는 상품 train데이터에서 제거
- 판매하지 않는 제품군에 대한 분석 및 trian 데이터 수정
 - 그래프로 출력하지 않고 오픈 전 가게 데이터 삭제
 - 판매액이 0인 품목의 경우 판매하지 않는 것으로 취급
 - train 데이터에서 제거
- 2017년 8월 15~31일까지의 데이터안에 미리 위 제품군을 0으로 대체
- holiday와 train데이터 결합 & holiday의 종류에 따라서 컬럼을 생성
 - 요일 관련 데이터 컬럼 추가

Test.csv

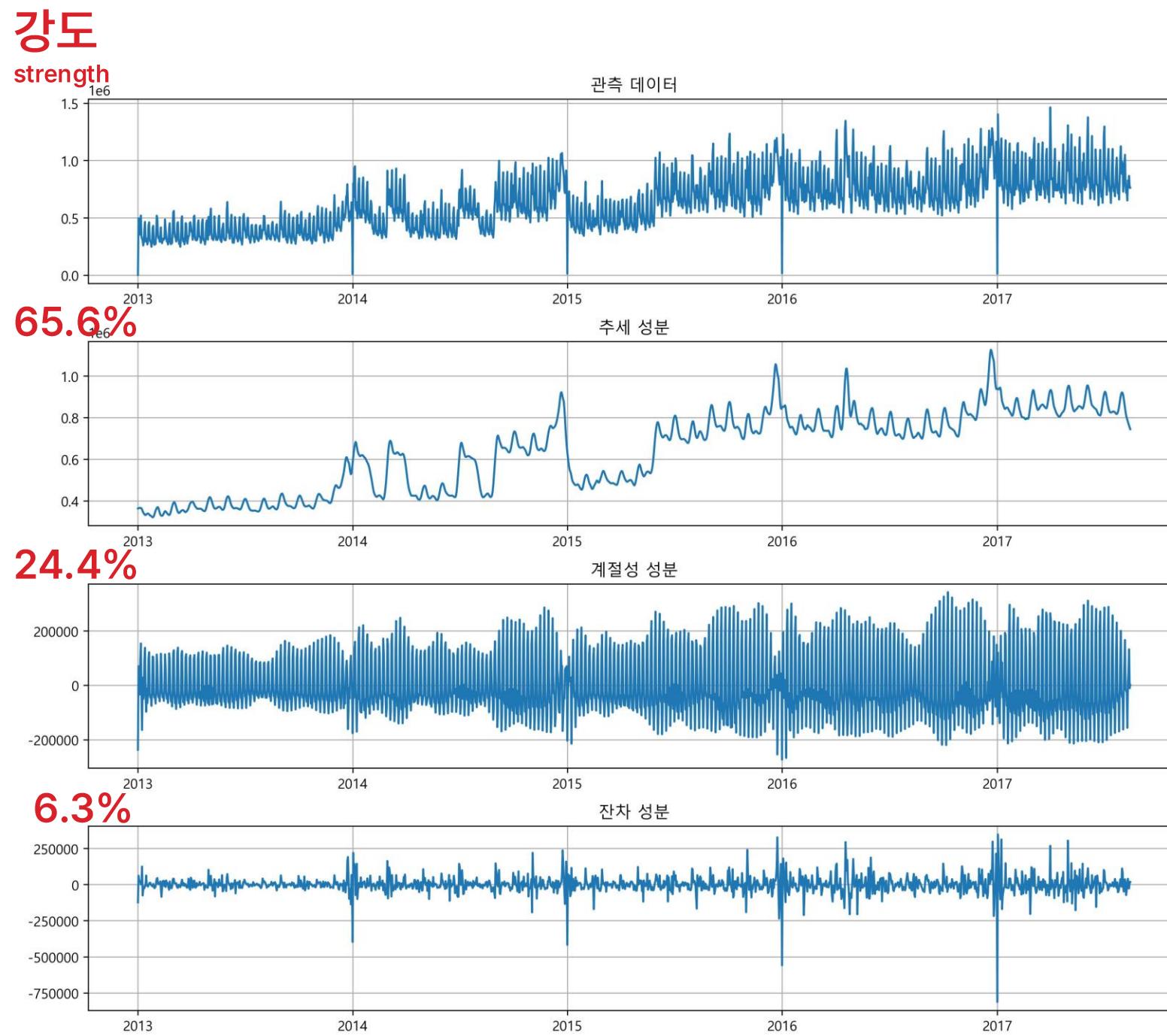
- train 데이터 이후 15일 까지의 판매액을 예측
 - Kaggle에 제출

진행

1. 분석 개요 및 데이터 설명
2. 환경구축
3. 데이터 EDA 및 전처리
- 4. 모델링**
 - 시계열 분석
 - 신경망 모델
 - 앙상블 모델
5. 모델 결과 및 케글 등수

시계열 분석

시계열 데이터의 구성요소별 탐색 설명입니다



시계열 데이터의 구성 요소

시계열 데이터는 장기적인 변화 양상인 "추세"와 시간 주기에 따라 규칙적으로 나타나는 패턴인 "계절성"과 설명되지 않는 변동인 "잔차"로 구성되어 있습니다

관측값 observed
원본 시계열 데이터



추세 trend
장기적으로 완만한 상승세를 확인할 수 있음



계절성 seasonal
주말과 평일의 매출 차이가 뚜렷하게 나타나는 것을 확인할 수 있음



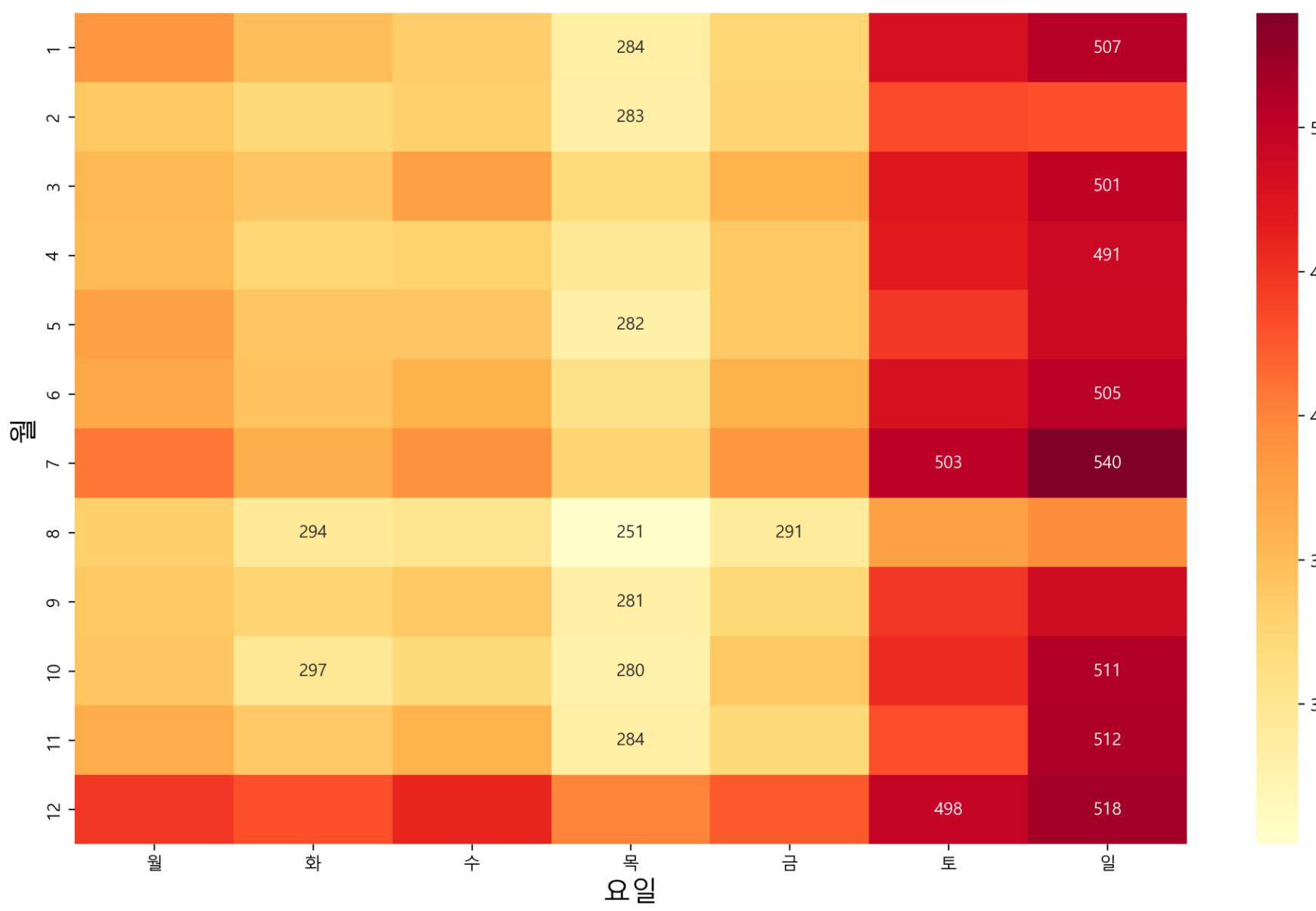
잔차 residual
잔차의 강도가 작은 것은 매출 패턴이 매우 예측 가능하다는 것을 의미

시계열 분석

각도에서 계절성을 탐색하는 과정입니다

계절성 탐색 : 시각화

월별-요일별 평균 매출액 히트맵



월 단위 monthly

12월이 전반적으로 가장 높은 매출을 보이며, 이는 연말 시즌의 소비 증가를 반영함을 확인할 수 있다.
8월이 대체로 가장 낮은 매출을 보이며, 이는 휴가 또는 특정 휴일이나 이벤트와 연관이 있을 가능성을 탐색해볼 수 있다.
월별 특징이 뚜렷하여 월 단위 계절성을 고려할 필요가 있다고 판단할 수 있다.

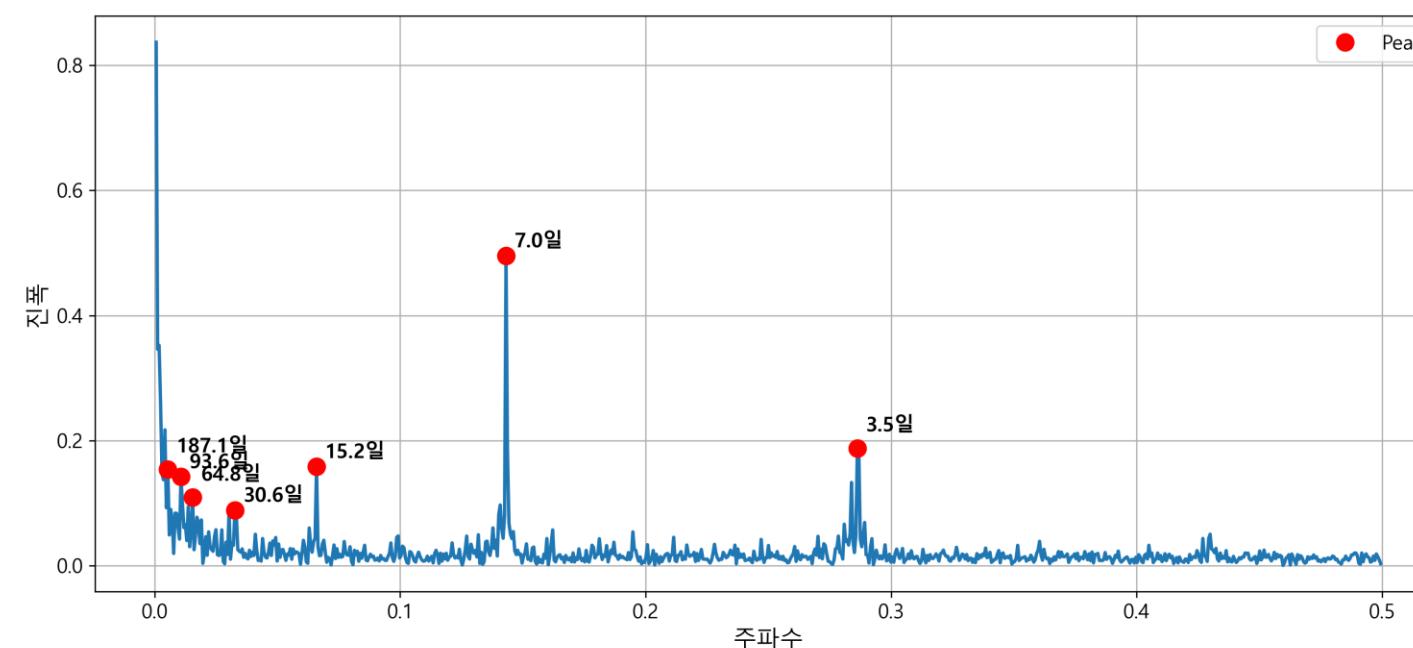
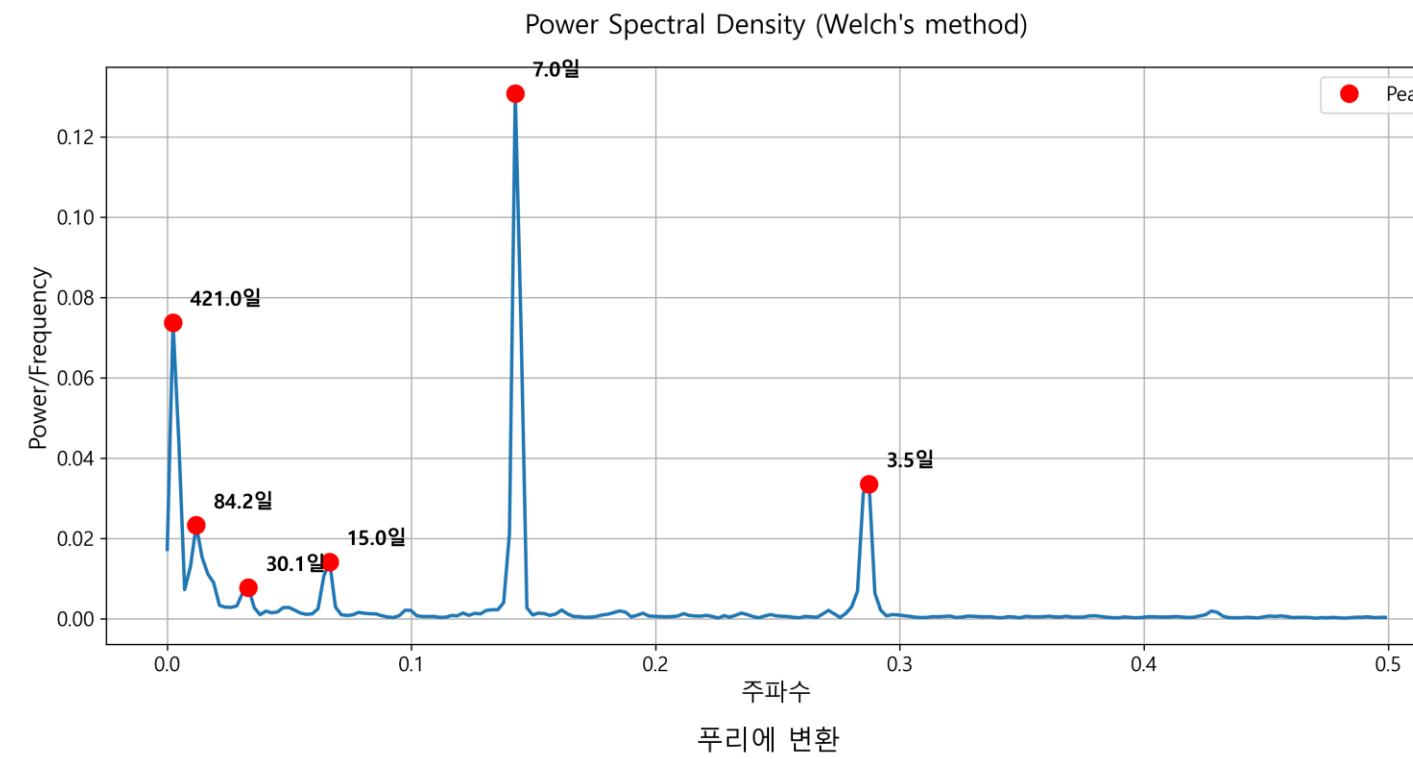
요일 단위 weekly

일요일이 대부분의 경우 가장 높은 평균 매출을 보이고 있다.
일요일에 이어 토요일이 2번째로 높은 매출로 주말 쇼핑의 선호도를 엿볼 수 있다.
목요일이 가장 낮은 평균 매출을 기록하고 있다.
평일과 주말의 차이가 명확하여 분석과 예측시에 중점을 두고 적용할 패턴이라고 판단할 수 있다.

시계열 분석

각도에서 계절성을 탐색하는 과정입니다

계절성 탐색 : 주파수



Power Spectral Density (Welch's method)

시계열 데이터를 주파수 영역으로 변환하여 신호의 에너지를 분포로 나타냅니다. Welch의 방법은 데이터 분할과 평균화를 통해 노이즈의 영향을 최소화합니다.

푸리에 변환 fourier transform

시계열 데이터를 구성하는 주기적 신호를 주파수 성분으로 분해합니다. 이를 통해 주파수와 시간 간의 관계를 탐색하여 숨겨진 주기성을 확인하는데 활용합니다.

계절성 후보

<주간 계절성> 두 기법 모두에서 7일, 3.5일이 명확하고 강함

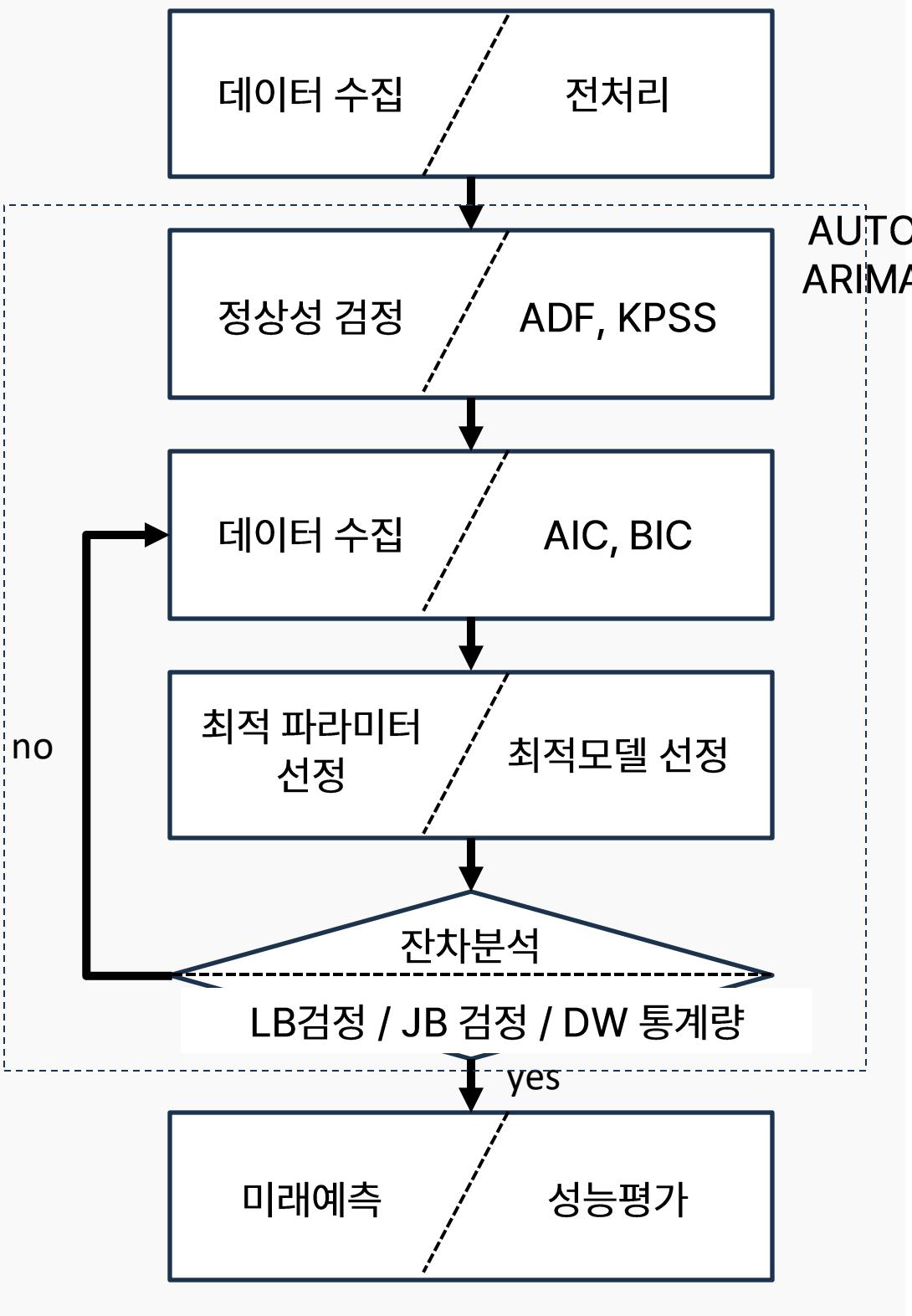
<월간 계절성> 30일, 15일 등도 유의미한 세기를 보여주고 있음

<특정 계절성> 84일, 94일 등으로 3개월 간의 계절 단위의 주기성도 고민해볼 수 있음. 또한 187일(6개월), 421일(14개월) 주기의 특정 패턴도 참고 가능.

SARIMAX 모델

데이터에 따라 파라미터(p, d, q)를 조정하여 최적의 ARIMA 모델을 생성하는 과정을 설명입니다

SARIMAX 플로우차트



SARIMAX

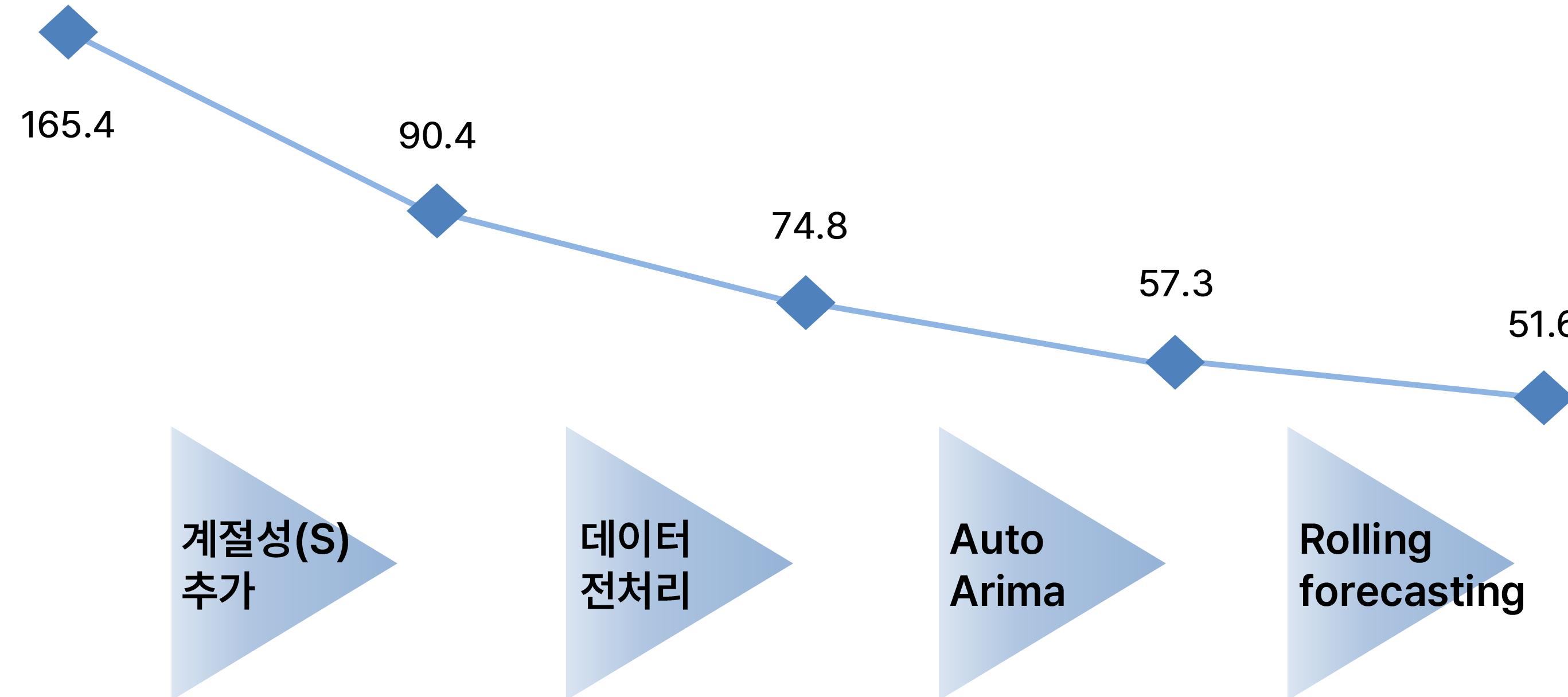
- AR (자기회귀, AutoRegressive): 과거의 데이터 값을 이용해 현재 값을 예측
- I (차분, Integrated): 평균이 다른 데이터를 차분하여 정상화
- MA (이동평균, Moving Average): 과거의 예측 오차를 활용해 현재 값을 예측
- S (계절성, Seasonality): 일정한 주기로 패턴이 반복되는 특성
- X (외생변수, eXogenous variables): 시계열 외부요소. 광고 휴일 이벤트 등

SARIMAX

ARIMA 모델의 주요 파라미터

- p: 자기회귀(AR) 모델의 차수로, 이전 데이터 포인트가 현재 데이터에 얼마나 영향을 미치는지를 나타냄
- d: 차분의 횟수로, 데이터를 정상성(stationary) 상태로 만들기 위해 얼마나 차분을 적용해야 하는지를 나타냄
- q: 이동 평균(MA) 모델의 차수로, 과거의 예측 오차가 현재 데이터에 미치는 영향을 설명
- P: 계절적 자기회귀(SAR) 모델의 차수로, 계절적 패턴에서 자기회귀 부분의 차수를 나타냄
- D: 계절적 차분 횟수로, 계절적 패턴을 정상성 상태로 만들기 위해 몇 번의 차분을 해야 하는지를 나타냄
- Q: 계절적 이동 평균(SMA) 모델의 차수로, 계절적 패턴에서 예측 오차의 영향을 설명
- s: 계절성 주기로, 계절적 패턴이 몇 개의 주기를 가지는지를 나타냄(예: 12개월, 7일 등)
- x: 외생변수로, 모델에 추가로 포함된 외부 변수들로 예측에 영향을 줄 수 있는 요인들을 나타냄

SARIMAX 모델 DEVELOP

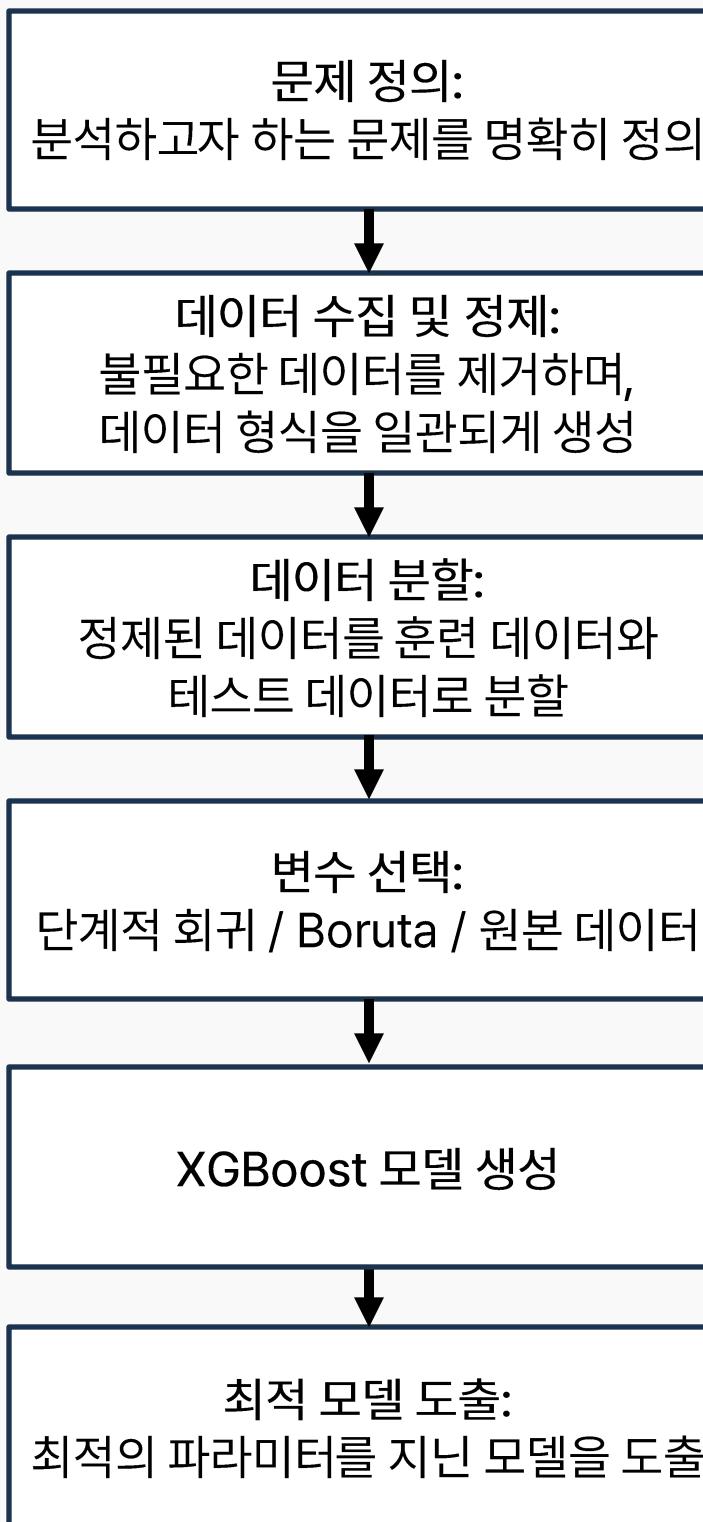


MAE	165.41	90.37	74.76	57.29	51.58
RMSE	612.53	309.90	251.93	132.43	136.42
R ²	0.795	0.938	0.957	0.96	0.96

XGBoost 모델

이전 트리의 오차를 보정하는 방식으로 새로운 트리를 추가하여 성능을 점진적으로 개선하는 모델 XGBoost 모델 설명입니다

XGBoost 플로우차트



(eXtreme Gradient Boosting)

XGBoost

XGBoost의 약자 = eXtreme Gradient Boosting
여러 개의 결정 트리를 결합하여 예측하는 방법 사용
이전 모델인 "Gradient Boosting"은 오류를 수정하는 방식으로 학습하며
XGBoost는 더 빠르고 정확하게 만들기 위해 개선된 알고리즘

XGBoost 모델의 주요 파라미터

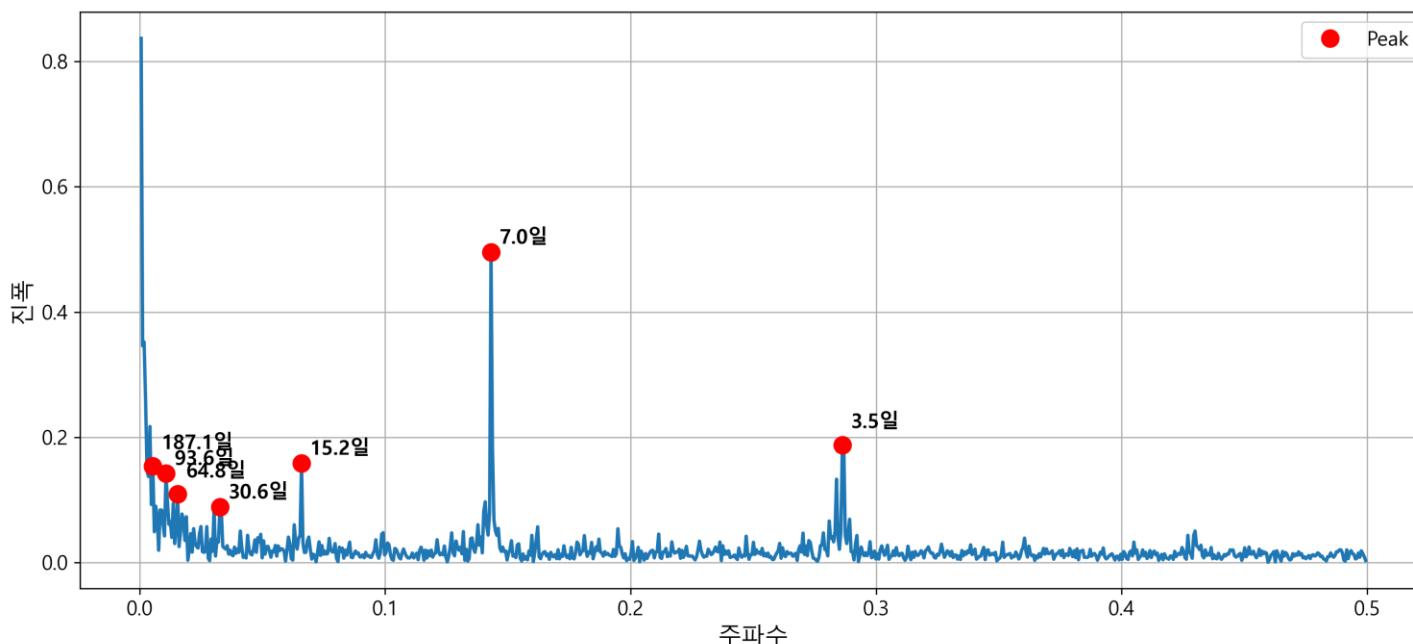
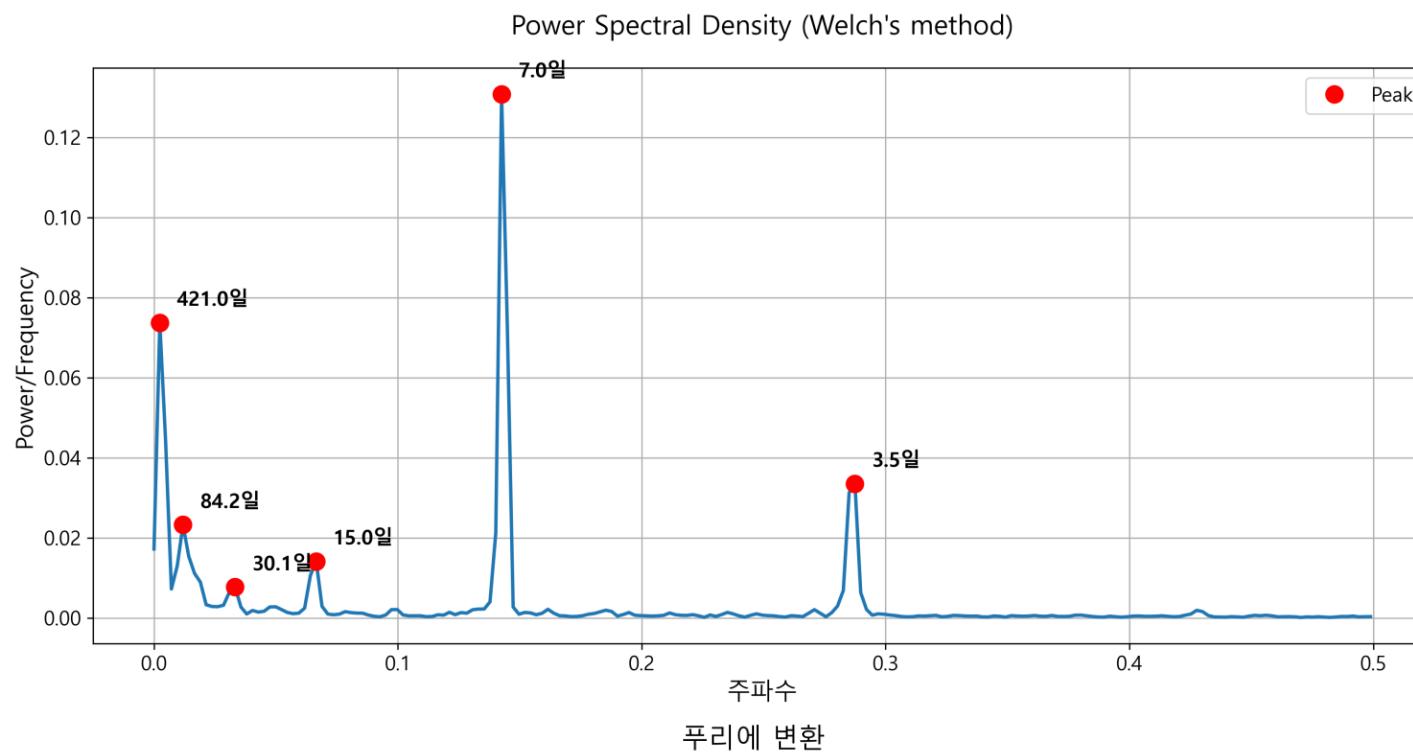
- learning_rate (학습률): 트리 간의 학습 진행 속도를 조정하는 매개변수
- n_estimators (추정기 개수): 생성할 트리의 개수를 결정
- max_depth (최대 깊이): 각 결정 트리의 최대 깊이를 설정
- subsample (샘플링 비율): 각 트리를 학습할 때 사용할 데이터 샘플의 비율을 설정
- colsample_bytree (특성 샘플링 비율): 각 트리를 학습할 때 사용할 특성(피처)의 비율을 설정
- gamma (감마): 트리 분할 시의 최소 손실 감소를 의미
- Objective (손실함수) : 최적화할 목표 함수 지정
- min_child_weight (오버피팅 방지): 자식 노드가 분할되기 위한 최소 가중치 합. 큰 값을 설정하면 과적합을 방지할 수 있음
- Verbosity (정보 수준): 학습 과정의 로그 출력력을 조정

XGBoost 모델

XGBoost 모델은 Boosting을 기반으로 하는 머신러닝 모델입니다.

파생변수 생성

- 시계열 분석을 통해 파악한 계절성(Seasonal)을 확인한 후 주기 패턴을 활용해 SMA, EWM, Lag features 등 파생변수 생성



파라미터 생성

```
# XGBoost 모델
params = {
    'objective': 'reg:squarederror',
    'eval_metric': 'rmse',
    'eta': 0.1,
    'max_depth': 6,
    'min_child_weight': 3,
    'subsample': 0.8,
    'colsample_bytree': 0.8,
    'lambda': 1,
    'alpha': 0.1,
    'verbosity': 0
}
```

objective: 'reg:squarederror' : 학습 목적 함수를 설정합니다.

eval_metric: 'rmse' : 평가 지표를 설정합니다.

eta : 학습률을 의미합니다.

max_depth : 트리의 최대 깊이를 설정합니다.

min_child_weight : 자식 노드를 분할하기 위한 최소 가중치 합입니다.

Subsample : 각 트리 학습에 사용할 샘플링 비율을 설정합니다.

colsample_bytree : 각 트리의 학습 시 사용할 특성(피처)의 비율을 설정합니다.

lambda : L2 정규화 (Ridge Regularization) 항의 가중치입니다.

alpha : L1 정규화 (Lasso Regularization) 항의 가중치입니다.

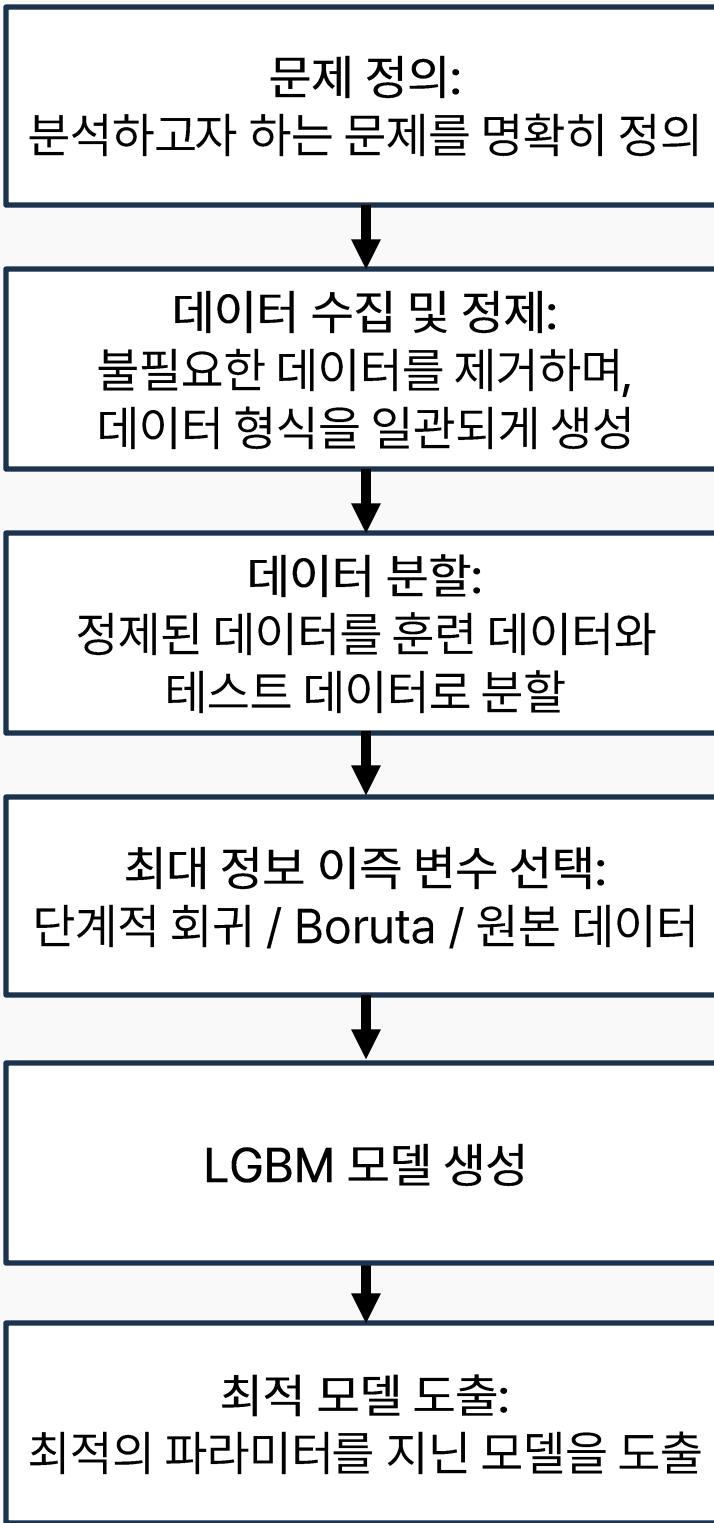
verbosity : 0 : 출력 로그 수준을 설정합니다.

이 파라미터들은 과적합을 방지하면서 예측 성능을 극대화하는 데 집중된 설정입니다.

LGBM 모델

랜덤 포레스트 기반으로, 트리 깊이와 학습률 등의 파라미터를 조정해 빠르고 효율적인 성능을 냅니다

Light GBM 플로우차트

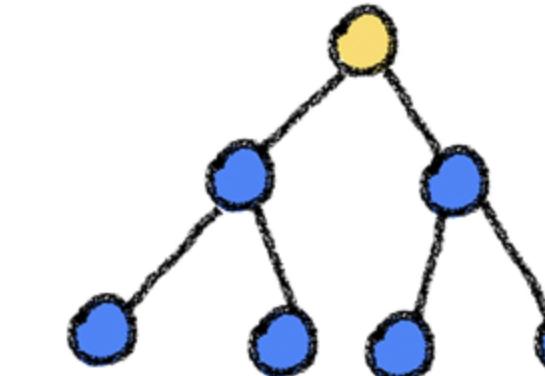


(Light Gradient Boosting Machine)

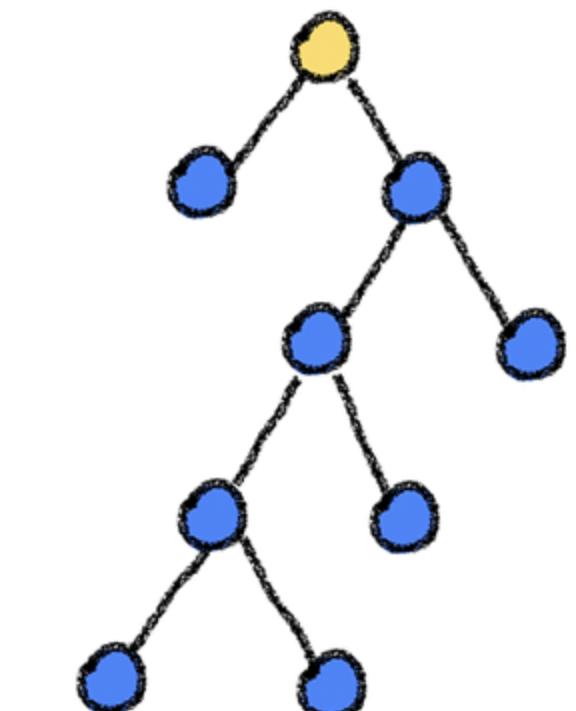


Gradient Boosting Machine(GBM) 기반의 알고리즘 기반 하여 예측 실패에 가중치를 부여해 순차적으로 트리를 생성 리프 중심 트리 분할 방식을 사용하여, 최대 손실 값을 가진 리프 노드를 지속적으로 분할함으로써 깊이 있는 트리를 형성하는 특징을 지님 일반적인 균형 트리 분할은 대칭적으로 트리의 균형을 맞추어 깊이가 깊어지지 않게 하지만, LightGBM은 깊이가 깊어지는 방식으로 학습하여 과적합 우려가 있지만 예측 오류 손실을 최소화

< 균형 트리 분할 >



< 리프 중심 트리 분할 >



LGBM 모델의 주요 파라미터

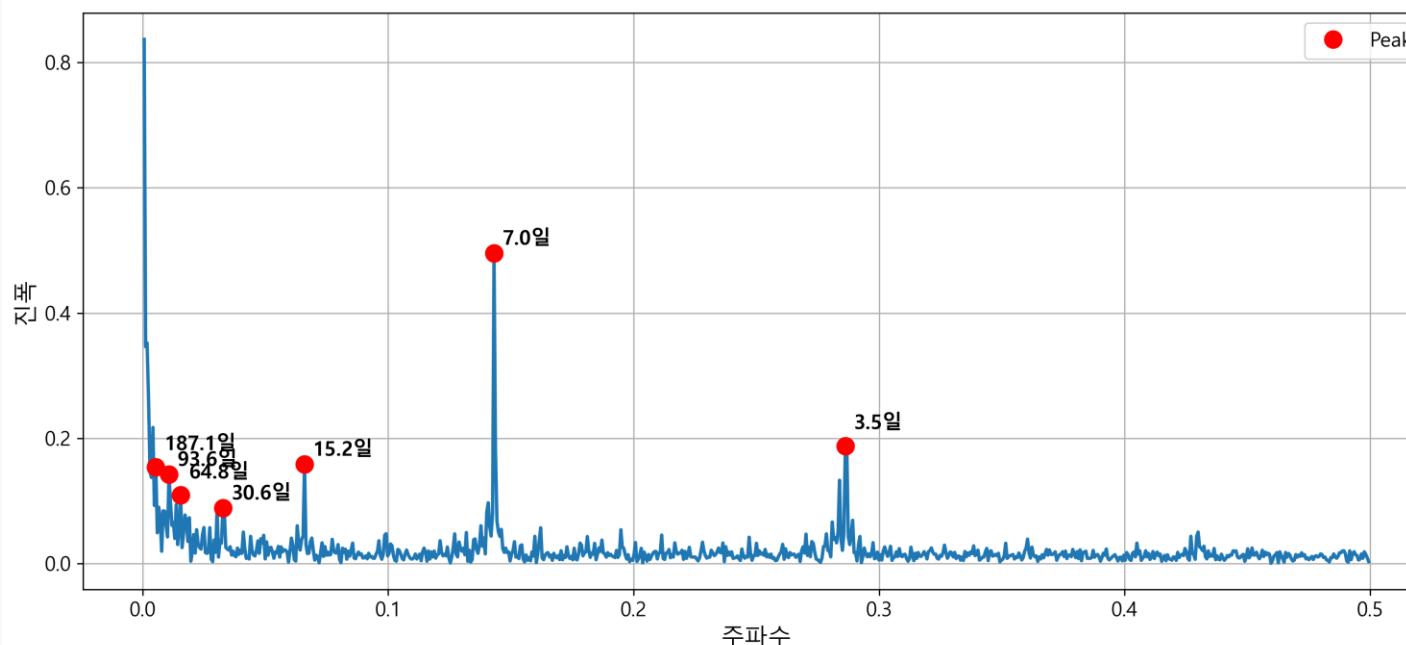
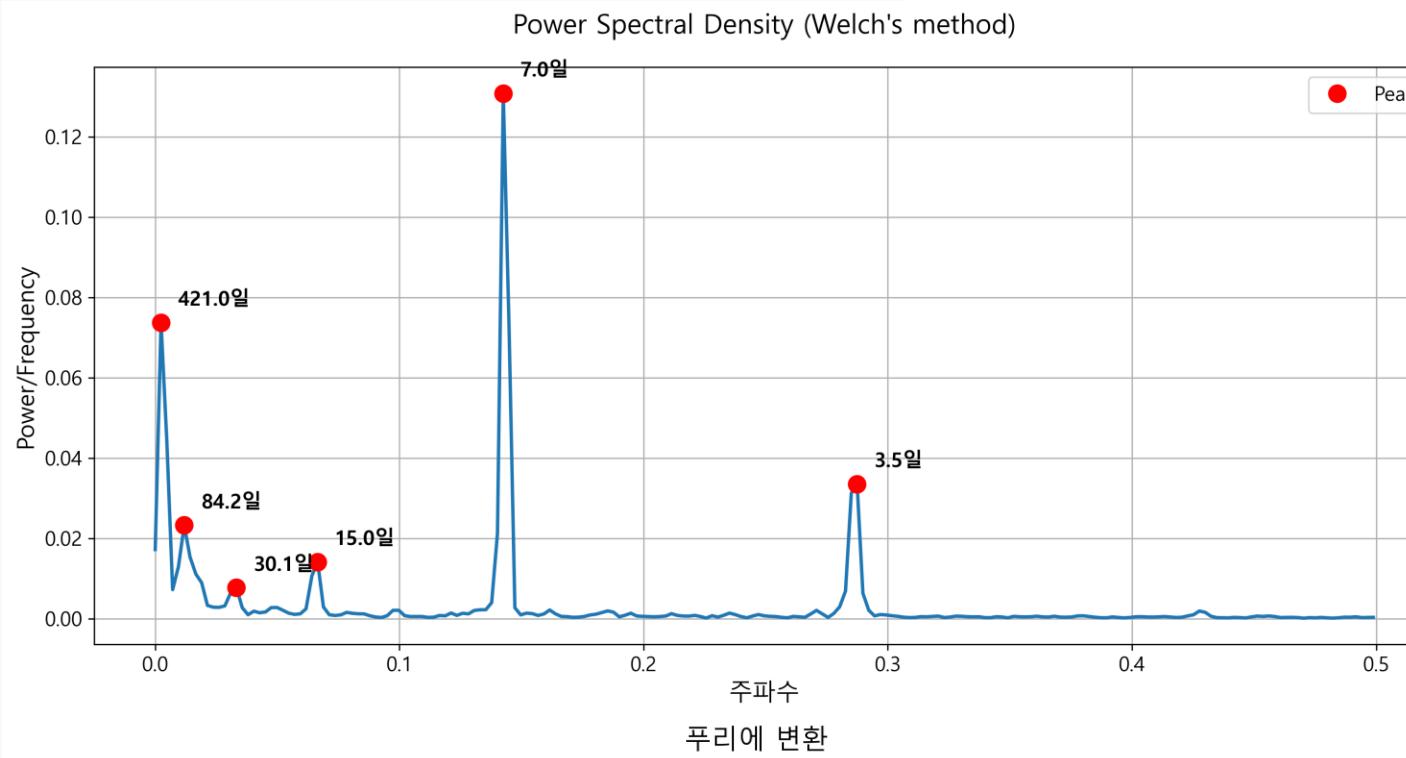
- num_leaves (리프 수): 트리의 리프 노드 최대 수. 모델 복잡도와 성능에 영향을 미침
- max_depth (최대 깊이): 트리의 최대 깊이. 과적합을 방지하기 위해 제한하는 역할
- learning_rate (학습률): 각 트리가 학습하는 속도. 낮추면 더 많은 트리로 학습하여 성능 향상 가능
- n_estimators (트리 개수): 생성할 트리의 수. 많을수록 복잡하지만 과적합 위험이 있음
- min_data_in_leaf (리프당 최소 데이터 수): 리프 노드에 있어야 할 최소 데이터 수. 큰 값은 과적합 방지
- bagging_fraction (샘플링 비율): 각 트리를 학습할 때 사용할 데이터의 비율. 샘플링을 통해 일반화 성능 향상
- feature_fraction (피처 샘플링 비율): 각 트리에서 사용할 피처의 비율. 데이터의 다양성을 높여 성능 개선

LGBM 모델

랜덤 포레스트 기반으로, 트리 깊이와 학습률 등의 파라미터를 조정해 빠르고 효율적인 성능을 냅니다

파생변수 생성

- XGBoost에서 활용한 SMA, EWM, Lag features
파생변수를 활용



파라미터 생성

```
params = {
    'objective': 'regression',
    'metric': 'mae',
    'boosting_type': 'gbdt',
    'lambda_l2': 0.3,
    'verbose': -1
}
```

objective: 모델의 학습 목표를 지정

regression: 회귀 문제를 해결하기 위한 설정

metric: 모델의 성능을 평가할 지표를 지정

mae(Mean Absolute Error): 평균 절대 오차를 사용하여 성능을 평가

boosting_type: 부스팅 알고리즘의 종류를 설정

gbdt: 기본 그래디언트 부스팅 방식

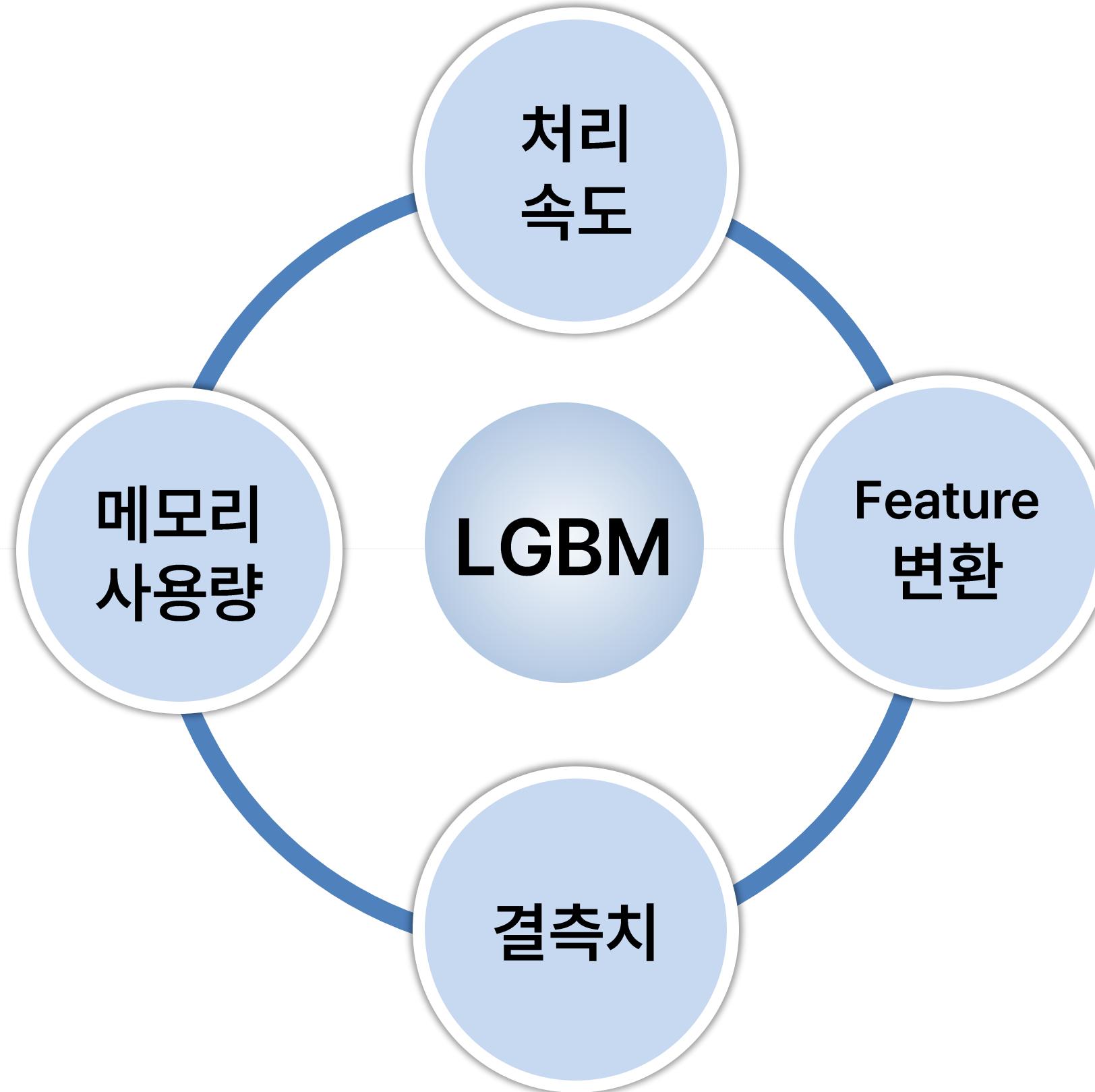
lambda_l2: 정규화 강도를 조정하는 하이퍼파라미터 [값이 클수록 과적합을 방지]

verbose: 출력 로그의 상세 수준을 설정

-1: 로그 출력을 최소화

LGBM 장점 및 활용성

LGBM 모델 장점에 대한 설명입니다



1. 처리속도

리프 중심 트리 분할 방식을 채택하여 데이터 처리의 효율성을 극대화합니다. 이로 인해 학습 속도가 빠를 뿐만 아니라, 예측 수행 시간도 단축됩니다. 특히 대규모 데이터셋을 다룰 때 이점이 큅니다

2. 적은 메모리 사용량

메모리 효율성이 뛰어나, 많은 데이터를 처리할 때에도 상대적으로 적은 메모리를 소모합니다. 특히 제한된 메모리 환경에서 유리합니다.

3. Feature 변환

원-핫 인코딩과 같은 전처리 과정 없이도 카테고리형 feature를 자동으로 변환할 수 있는 기능을 제공합니다. 이 기능은 데이터 전처리의 복잡성을 줄이고, 모델링 시간을 단축시키며, 최적의 노드 분할을 통해 성능을 높이는 데 기여합니다.

4. 결측치 처리 능력

결측치가 있는 데이터를 별도로 처리할 필요 없이, 자연스럽게 모델 학습에 반영할 수 있습니다. 이로 인해 데이터 전처리 과정에서의 추가 작업을 줄이고, 결측치가 있는 데이터셋에서도 유연하게 성능을 발휘할 수 있습니다.

진행

1. 분석 개요 및 데이터 설명
2. 환경구축
3. 데이터 EDA 및 전처리
4. 모델링
 - 시계열 분석
 - 신경망 모델
 - 앙상블 모델
5. 모델 결과 및 케글 등수

모델 선정

최적의 값을 도출한 모델 선정에 대한 설명입니다

모델이름	XGBoost	Light GBM
Kaggle 등수	765/801	68/724
Kaggle score	3.47144	0.42749

다른 모델들과의 비교에서 RMSLE(제곱 평균 오차)와 Kaggle의 등수를 기준으로 우수한 수치를 기록

LightGBM 모델이 해당 지표면에서 XGBOOST보다 나은 수치를 기록

또한, RMSLE는 예측값과 실제값 간의 비율적인 차이를 고려하며 큰 오차에 민감하게 반응하지 않는 지표인데,

이 역시 LightGBM모델이 보다 더 안정적인 성능을 보였습니다.

이러한 이유로 **LightGBM 모델은 작은 오차와 큰 오차 모두를 효과적으로 제어하며, 실제 데이터 예측에 가장 적합한 모델로 선정되었습니다.**

결과 테이블

Kaggle 제출 결과에 대한 설명입니다

kaggle Your Work

This is the **private view** of your content. To see what others see, visit [Your Profile](#)

+ Create

Overview Collections Code Datasets Models Competitions Discussions Bookmarks

Your Competitions (1)

Search Your Work

All Filters Privacy Role Status

0 selected

Store Sales - Time Series Forecasting
Use machine learning to predict grocery sales
Getting Started · 724 Teams · Ongoing

68/724*

...

감사합니다.