SMARCLE 2021 데이콘 스터디

4장 상점 신용카드 매출 예측 경진대회

이은지, 조동현, 강호연, 김지은

SMARCLE

CONTENTS

SMARCLE

1 문제 정의

01 문제 정의

1.1 경진대회 소개

봄, 예측 시 새학기, 황사,
가정의 달 등의 변수가 매출에
영향을 줌.

01 문제 정의

1.2 평가척도

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i'|$$

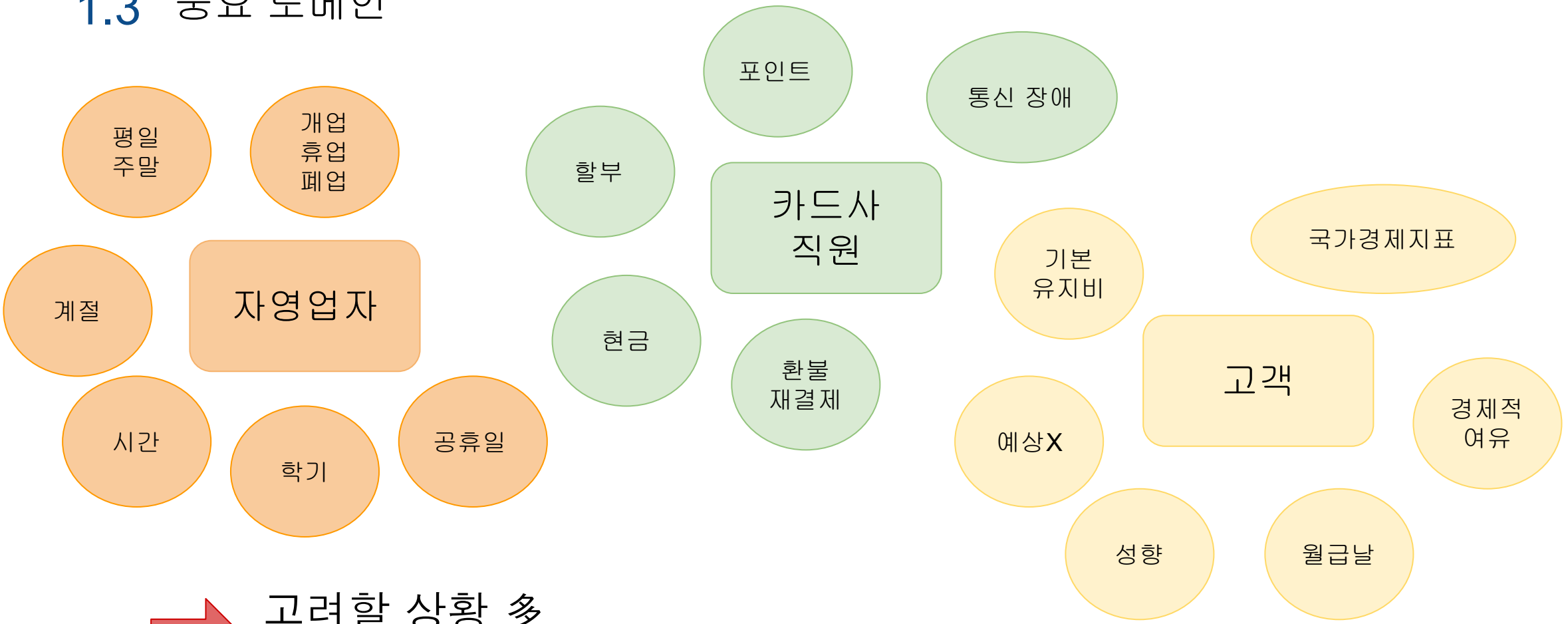
실제값과 예측값의 차이에 절댓값을 취하기 때문에, 오차의 크기 그대로 반영
따라서 오차에 따른 손실이 선형적으로 올라가는 상황에서 쓰기 적합

ex) A상점에서 200만원, B상점에서 300만원의 매출 예측, 실제로 A상점에서 150만원, B상점에서 180만원 매출이 나타난다면 손해보는 금액은
(200-150)+(300-180)

SMARCLE

01 문제 정의

1.3 중요 도메인



고려할 상황 多
각 상점의 매출 특성이 독립된 특징을 지님

01 문제 정의

1.4 문제 해결을 위한 접근 방식

시계열 모델

AR : 과거와 현재 자신과의 관계
MA : 과거와 현재 자신의 오차와의 관계
ARMA : 이전항의 상태와 오차에서 현재의 상태를 추론
ARIMA : 현재와 추세간의 관계
불규칙적인 시계열 데이터를 예측하기 위하여 고안

2 데이터 전처리

02 데이터 전처리

2.1 노이즈 제거

2.2 다운 샘플링

시계열 데이터에서 시간 간격을 넓게 재조정해 샘플 수를 줄이는 것

시간 간격이 좁을 시, 예측 구간이 커져 불확실성 증가

2.1 날짜 지정 범위 생성 & 시리즈 객체 변환

jupyter notebook->

3 탐색적 데이터 분석

03 탐색적 데이터 분석

3.1 상점별 매출 특성

1. 계절성이 있는 상점
2. 추세가 있는 상점
3. 휴업 중인 상점

3.2 시계열 데이터의 정상성

03 탐색적 데이터 분석

3.1 상점별 매출 특성

1. 계절성이 있는 상점

- 봄, 여름, 가을, 겨울 중 **특정 계절**을 중심으로 장사하거나 장사하지 않는 특징을 지닌 상점
- 상점아이디 257번 데이터를 시리즈 객체로 데이터 출력
- 257번 상점은 11~3월 매출이 0
- 상점아이디 2096번을 시리즈 객체로 데이터 출력+시계열 그래프
- 2096번 상점은 겨울 시즌(1~3월)에 낮은 매출액, 4월 매출액 급상승

⇨ **2019년 3~5월 기간에 매출의 급상승을 예상할 수 있음**

03 탐색적 데이터 분석

3.1 상점별 매출 특성

2. 추세가 있는 상점

- 매출이 꾸준하게 증가 or 꾸준히 감소하는 상점
- ex) SNS마케팅, 유튜브 광고 등을 통해 매출이 증가하는 상품 취급 상점
- 335번의 시계열 그래프를 통해 2016년 6월~2017년 10월 매출액 감소추세, 2017년 11월~2019년 2월 매출액 증가추세 알 수 있음
- 510번 상점은 2017년 9월 이후에는 꾸준히 감소

⇨ **2019년 3~5월의 매출이 증가 추세임을 알 수 있음**

03 탐색적 데이터 분석

3.1 상점별 매출 특성

3. 휴업 중인 상점

- 몇개월 간 매출이 발생하지 않은 상점
- ex) 2019년 1,2월 매출이 0, 그 전부터 매출액이 존재하지 않는 상점
- 111번의 데이터와 시계열 그래프를 통해 2018년 10월~2019년 2월까지 매출 발생하지 않았음을 알 수 있음
- 279번 상점은 2019년 2월 매출이 재감소하면서 휴업 가능성이 보임

⇨ **111번 상점은 2019년 3~5월의 매출이 발생하지 않을 것으로 예측**

SMARCLE

03 탐색적 데이터 분석

3.2 시계열 데이터의 정상성

정상성이란?

- 추세나 계절성이 없는 시계열 데이터
- 데이터가 시간의 변동에 따라 평균과 분산이 일정

.

시계열 데이터의 정상성을 판단하기 위해 **ADF-Test**사용

ADF-Test

통계학에서 시행하는 가설 검정 절차 따름

- **귀무 가설** : 시계열 자료가 정상 시계열이 아니다.
- **대립 가설** : 시계열 자료가 정상성을 만족한다.

차분(differencing)을 통해 비정상 시계열을 평균이 일정한 정상 시계열로 변환하는 작업

4 모델 구축과 검증

SMARCLE

04 모델 구축과 검증

4.0 시계열 분석 모델

4.1 자기회귀누적이동평균 모델

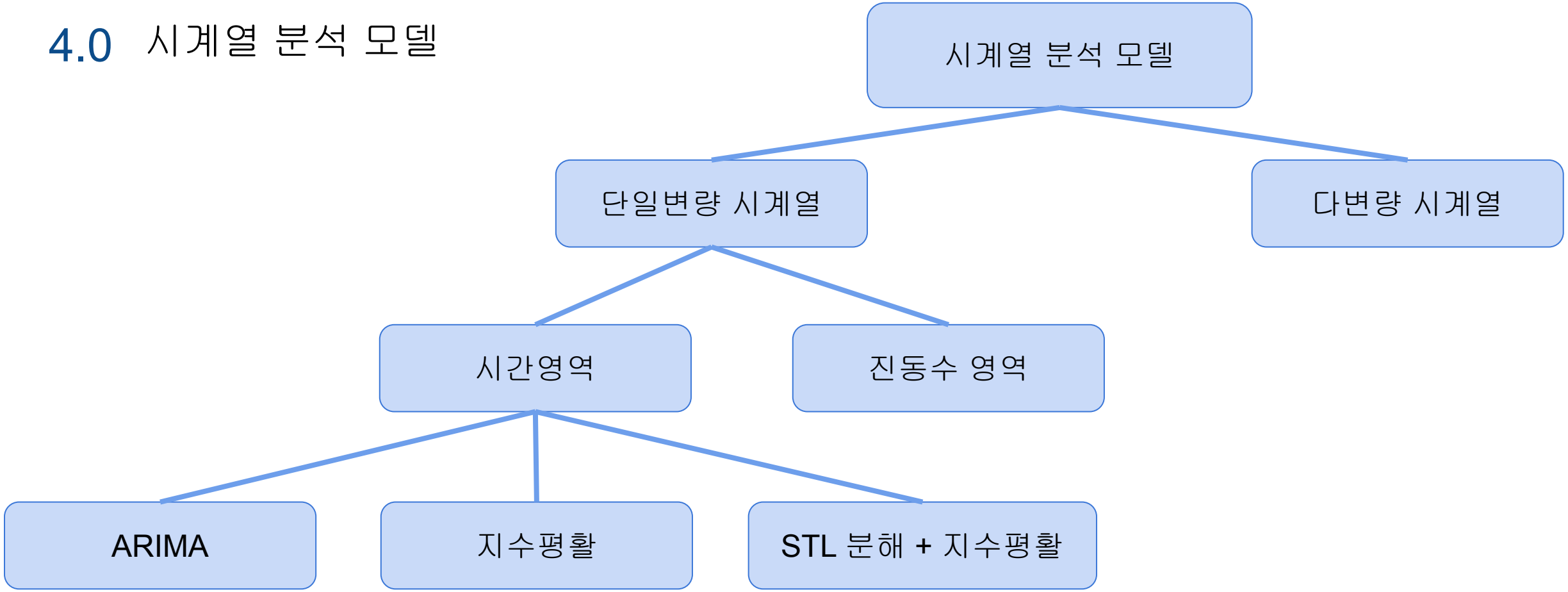
4.2 지수평활법

4.3 STL 분해를 적용한 지수평활법

4.4 코드

04 모델 구축과 검증

4.0 시계열 분석 모델



04 모델 구축과 검증

4.1 자기회귀누적이동평균 모델

ARIMA

AR – Auto-regressive

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

MA – Moving Average

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

I – Integrated

ARIMA(p,d,q) – Auto-regreessive Integrated Moving Average

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

SMARCLE

04 모델 구축과 검증

4.2 지수평활법

1. 단순 지수평활법
2. 홀트의 선형추세 기법

04 모델 구축과 검증

4.2 지수평활법

1. 단순 지수평활법

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots$$

알파 값 -> SSE 값을 최소화

$$\text{SSE} = \sum_{t=1}^{T}(y_t - \hat{y}_{t|t-1})^2 = \sum_{t=1}^{T} e_t^2$$

04 모델 구축과 검증

4.2 지수평활법

2. 홀트의 선형추세 기법

예측식	$\hat{y}_{t+h\|t} = l_t + hb_t$
수준식	$l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1})$
추세식	$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1}$

04 모델 구축과 검증

4.2 지수평활법

AIC – Akaike's Information Criterion(아카이케 정보 기준)

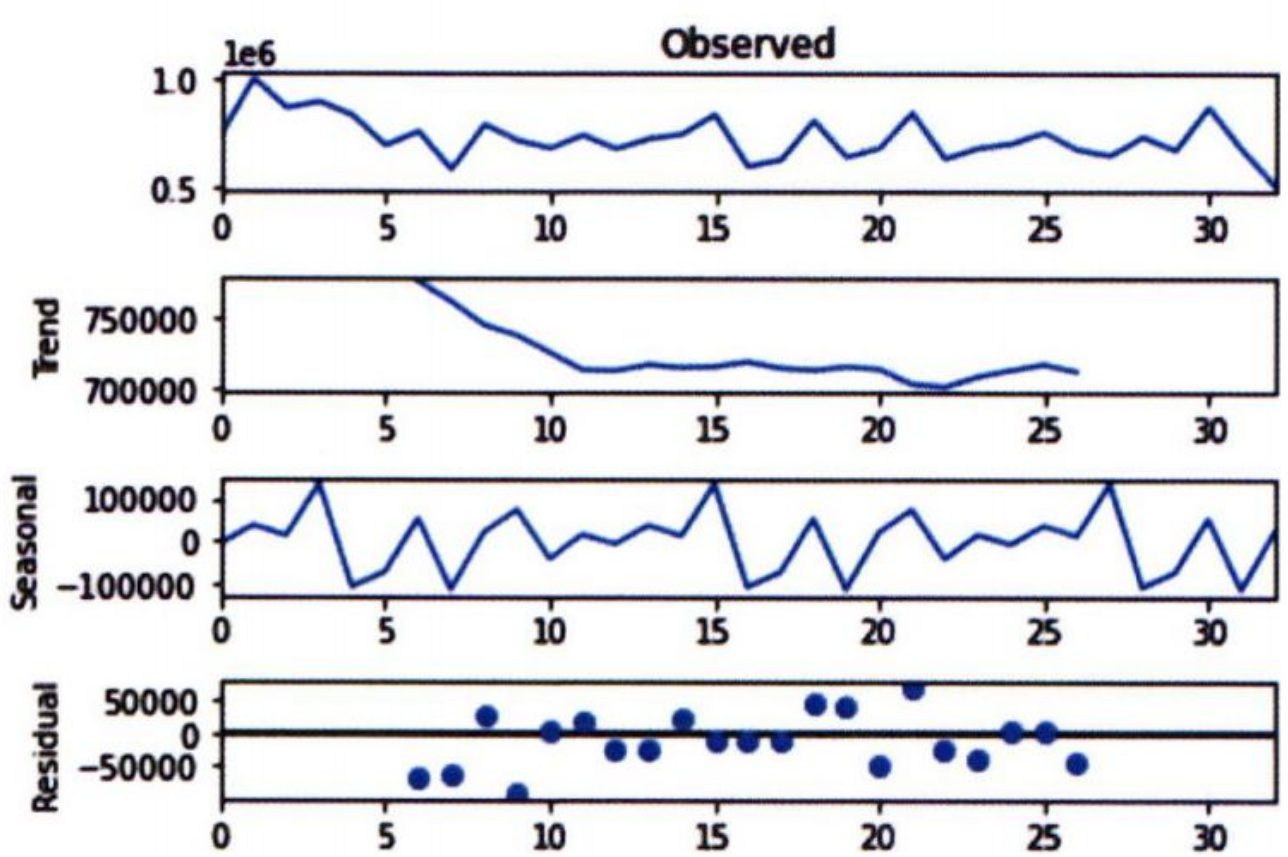
$$AIC = -2\ln(L) + 2k$$

L – Likelihood, 모델의 적합도
k – 파라미터 개수

04 모델 구축과 검증

4.3 STL 분해를 적용한 지수평활법

STL – Seasonal and Trend decomposition using Loess
Loess – Local regression



04 모델 구축과 검증

4.4 코드

5 성능 향상을 위한 방법

05 성능 향상을 위한 방법

정리

1. 데이터 전처리

```
# 매출 변동 계수를 구하는 함수
def coefficient_variation(df, i):
    cv_data = df.groupby(['store_id']).amount.std()/df.groupby(['store_id']).amount.mean()
    cv = cv_data[i]
    return cv

# 매출액 변동 계수가 0.3 미만인 경우만 log를 씌움
if cv < 0.3:
    train_log = ts(log(store['amount']), start=c(start_year,start_month), frequency=12)
  # 앙상블 예측
    forecast_log = hybridModel(train_log)
    final_pred.append(np.sum(pandas2ri.ri2py(exp(forecast_log)).values))
# 매출액 변동 계수가 0.3 이상인 경우
else:
    train = ts(store['amount'], start=c(start_year,start_month), frequency=12)
  # 앙상블 예측
    forecast = hybridModel(train)
    final_pred.append(np.sum(pandas2ri.ri2py(forecast).values))
```

매출액의 작은 변동을 안정화해 더 큰 트렌드를 파악하기 위해

매출액 변동계수(표준편차를 평균으로 나눈 지표)를 고려한 로그 정규화

2. 앙상블

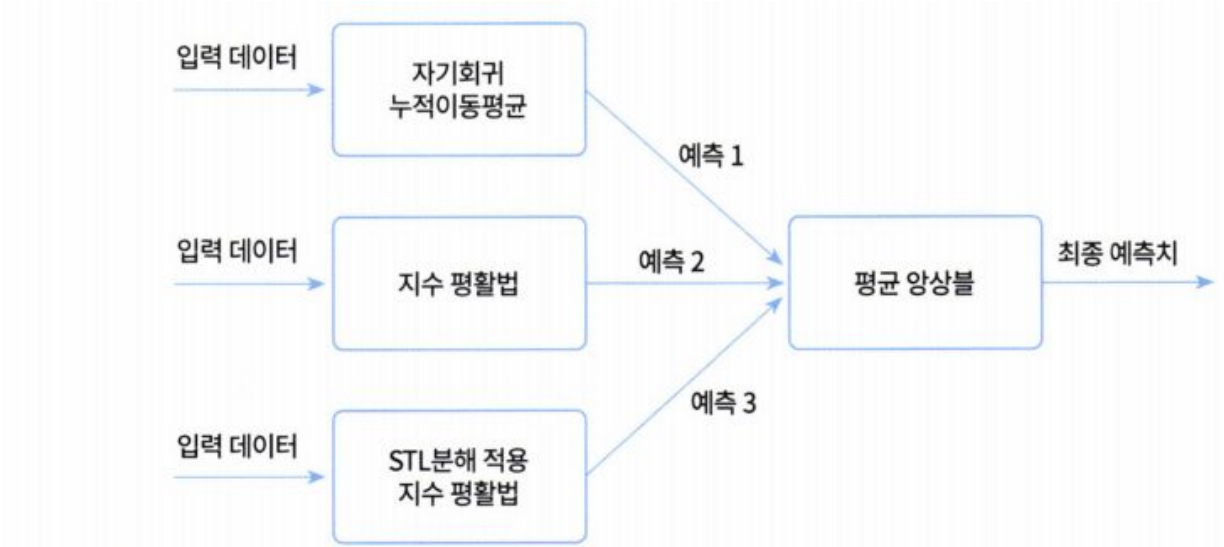


그림 4.17 평균 앙상블 방법

R시계열 패키지 forecastHybrid를 통한 앙상블

자기회귀누적이동평균모델, 지수평활법, STL분해를 적용한 지수평활법으로 3개의 예측치 생성
평균값을 구해 최종 매출액 계산

과적합 방지. 성능향상

SMARCLE

06 정리

⭐ 코랩에서 실행되지 않는 코드가 많아 책에 나온것처럼 주피터 노트북 사용함

⭐ 대회에 참가하기 전에 경진대회와 관련 사전 도메인 공부 추천

⭐ 관련 논문 리뷰, 관련 분야 전문가 인터뷰를 통해 사전 도메인을 알면
데이터 전처리와 파생 변수 생성, 모델 적용 과정 훨씬 수월

⭐ 데이콘에 있는 베이스라인 코드와 다른 참가들의 코드(영상)를 확인하여 책 이외의
다른 방법들로도 코드를 작성해보기!

SMARCLE

Thank you

SMARCLE