19회 ADP 기출 분석

2022. 1. 23

정창균

1. 기계학습 - Credit Data : 고객이 이탈여부 분류 문제 (50점)
독립변수 : 성별, 나이, 카드등급, 소득 등

- **1-1. 데이터 전처리 및 시각화 (5점)**
  - 데이터 확인 -> 이상치/결측치 처리 -> 정규화(MinMaxScaler), 표준화(StandardScaler)
    * 시험이니 범주화, 파생변수는 간단한 것만

- **1-2. 훈련,검증 데이터 분할 7:3 및 Confusion Matrix 만들기 (15점)**
  - from sklearn.model_selection import train_test_split
    * 재현성, 층화추출 옵션 사용 random_state = 2022, stratify = y
  - from sklearnmetrics imort confusion_matrix

- **1-3. 분류분석 3개를 앙상블하여 Credit_test를 예측하기(30점)**
  - randomforest, xgboost, logisticregression, decisiontree 중 시험당일 느리거나 오류나는것 뺀 나머지 3개 사용 할듯함

1-3. 분류분석 3개를 앙상블하여 Credit_test를 예측하기(30점)

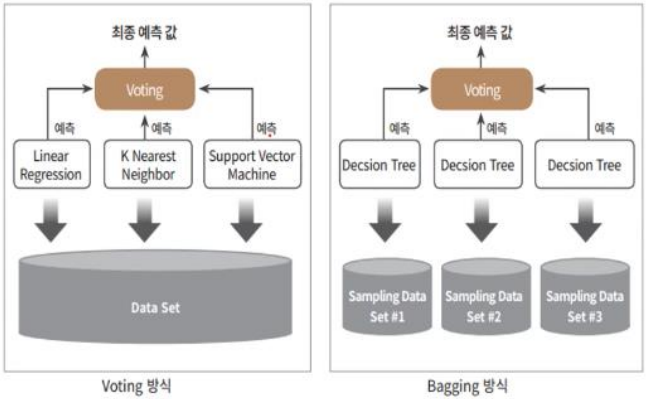
- 분류분석 3개를 앙상블 하라?

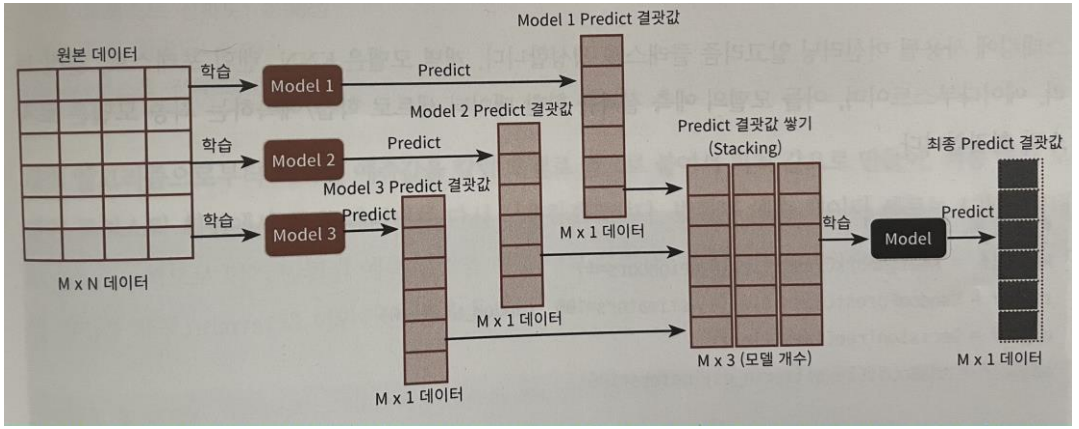
- 1. 배깅, 부스팅, 랜덤포래스트 가 앙상블?

- 2. 모델1, 모델2, 모델3 result값을 비율별로 합한 값이 앙상블

  예) (모델1 x 0.5) + (모델2 x 0.2) + (모델3 x 0.3)

- 3. 스태킹이 앙상블?





2. 통계학습 - Traffic EPS 시계열 분석 - 20년치 데이터, 분기데이터 (1년에 4개) (50점)

- **2-1. 시계열 데이터의 정규성과 이분산성을 분석하기 위해 시각화 하고 설명(10점)**
- 정상성(Stationary) : 시점에 관계없이 평균과 분산이 일정한 상태
- 이분산성 : 등분산성이 결여된 경우

- **2-2 시계열데이터 정규성이 아니라면, 고정시계열이 있는지 확인하고 이를 처리(15점)**
- decompose를 통해 요소별로 구분? (트렌드, 계절성, 순환성, 잡음)
- differencing(차분) 또는 transformation(log변환) 등을 활용

- **2-3. SARIMA 분석을 실시, 여러 파라미터를 적용해보고 가장 성능 좋은 것을 제시(15점)**
- autoarima를 사용하자

- **2-4. 모델의 잔차와 잡음에 대해 시각화 하고 분석(10점)**

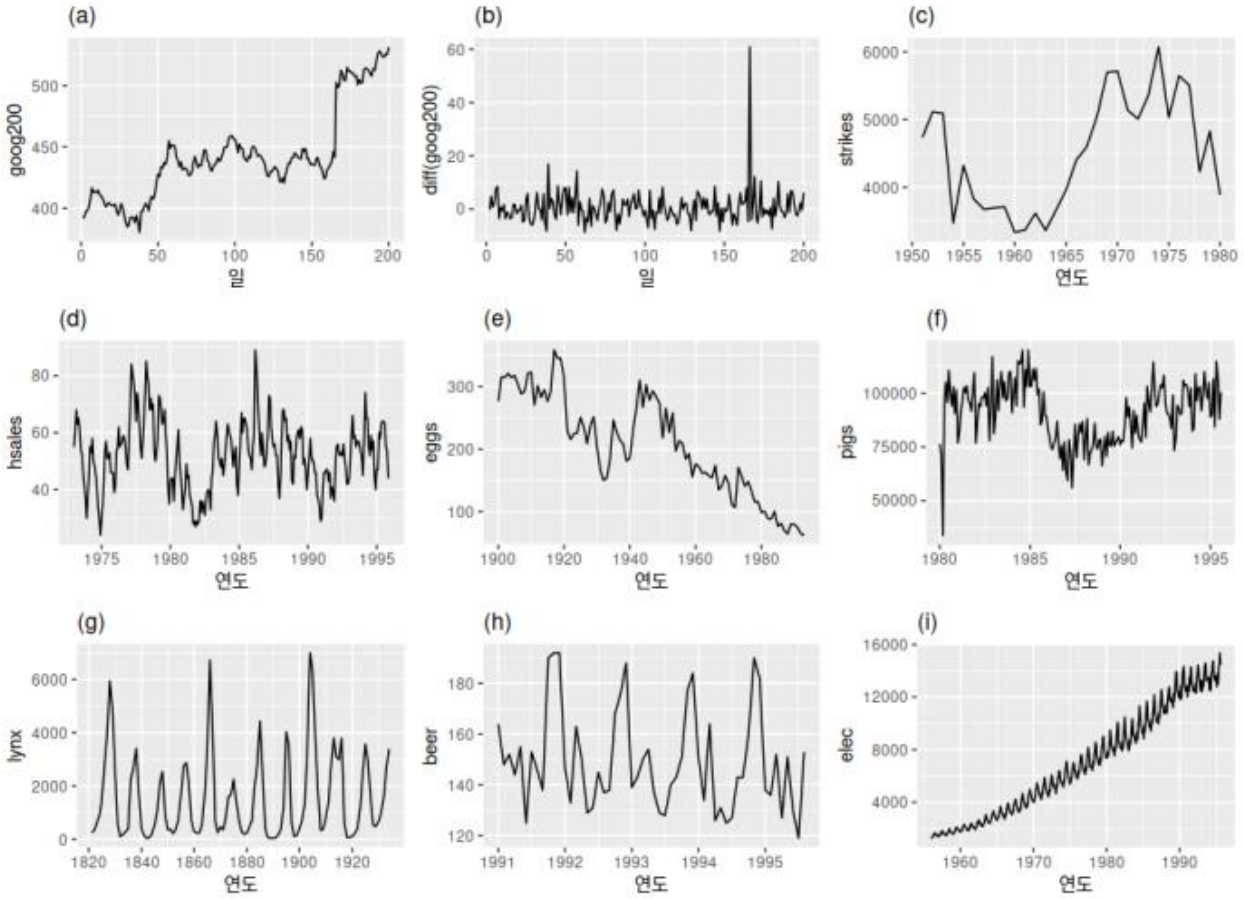
Traffic eps data ?

- 교통량 데이터이고 인구증가와 분기별 주기성이 있다고 가정

2-1. 시계열 데이터의 정규성과 이분산성을 분석하기 위해 시각화 하고 설명(10점)

- 정상성(Stationary) : 시점에 관계없이 평균과 분산이 일정한 상태
- 이분산성 : 등분산성이 결여된 경우

• 정상성(stationarity)을 나타내는 시계열은 시계열의 특징이 해당 시계열이 관측된 시간에 무관합니다. 따라서, 추세나 계절성이 있는 시계열은 정상성을 나타내는 시계열이 아닙니다 — 추세와 계절성은 서로 다른 시간에 시계열의 값에 영향을 줄 것이기 때문



2-1. 시계열 데이터의 정규성과 이분산성을 분석하기 위해 시각화 하고 설명(10점)
- 정상성(Stationary) : 시점에 관계없이 평균과 분산이 일정한 상태
- 이분산성 : 등분산성이 결여된 경우

- 시계열에서 이분산성은 잔차 분석에서 쓰이는 말로 오차값(잔차)들간의 상관성이 있는 경우로 시계열 분석이 완전하지 않음을 의미함

  * 회귀모형에 대한 가정에 등분산성 기억할것(선형성, 독립성, 등분산성, 비상관성, 정상성)

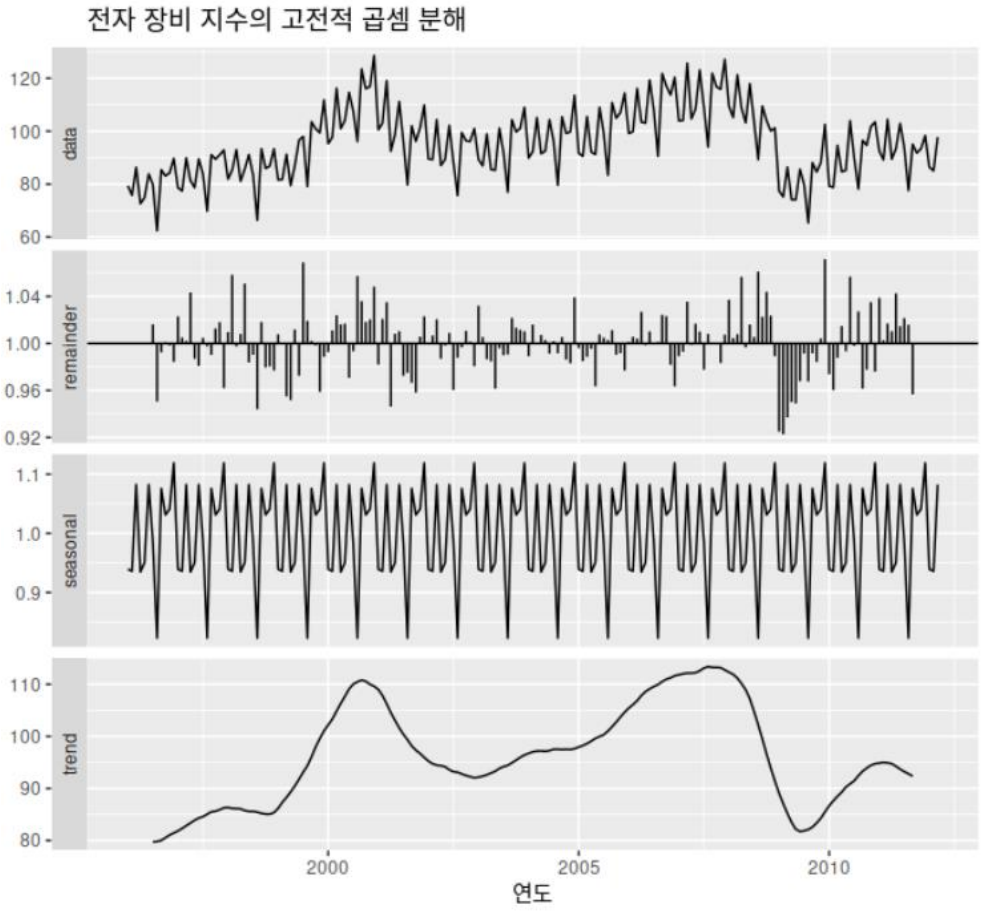
- 잔차(residual)는 어떤 모델이 데이터의 정보를 적절하게 잡아냈는지 여부를 확인할 때 유용합니다. 좋은 예측 기법은 다음과 같은 특징을 갖는 잔차(residual)를 낼 것입니다.

  1. 잔차(residual)에 상관 관계가 없습니다. 잔차 사이에 상관관계(correlation)가 있다면, 잔차에 예측값을 계산할 때 사용해야하는 정보가 남아 있는 것입니다.

  2. 잔차의 평균이 0입니다. 잔차의 평균이 0이 아니라면, 예측값이 편향(bias)될 것입니다.

2-2 시계열데이터 정규성이 아니라면, 고정시계열이 있는지 확인하고 이를 처리(15점)
decompose를 통해 요소별로 구분? (트렌드, 계절성, 순환성, 잡음)
differencing(차분) 또는 transformation(log변환) 등을 활용

- STL은 다양한 상황에서 사용할 수 있는 강력한 시계열 분해 기법입니다. STL은 "Seasonal and Trend decomposition using Loess(Loess를 사용한 계절성과 추세 분해)"의 약자입니다. 여기에서 Loess는 비선형 관계를 추정하기 위한 기법입니다.



2-3. SARIMA 분석을 실시, 여러 파라미터를 적용해보고 가장 성능 좋은 것을 제시 (15점)

autoarima를 사용하자

8.9 계절성 ARIMA 모델들

지금까지, 비-계절성 데이터와 비-계절성 ARIMA 모델에만 관심을 두었습니다. 하지만, ARIMA 모델로 다양한 종류의 계절성 데이터를 모델링 할 수도 있습니다.

계절성 ARIMA 모델은 지금까지 살펴본 ARIMA 모델에 추가적인 계절성 항을 포함하여 구성됩니다. 계절성 ARIMA 모델을 다음과 같이 쓸 수 있습니다:

$$\text{ARIMA}\ \underbrace{(p,d,q)}_{\text{모델의 비 계절성 부분}}\ \underbrace{(P,D,Q)_m}_{\text{모델의 계절성 부분}}$$

Table 8.2: 다양한 ARIMA 모델을 H02 월별 처방전 판매량 데이터에 적용한 결과에 대한 RMSE 값.

모델	RMSE
$\text{ARIMA}(3,0,1)(0,1,2)_{12}$	0.0622
$\text{ARIMA}(3,0,1)(1,1,1)_{12}$	0.0630
$\text{ARIMA}(2,1,3)(0,1,1)_{12}$	0.0634
$\text{ARIMA}(2,1,1)(0,1,2)_{12}$	0.0634
$\text{ARIMA}(2,1,2)(0,1,2)_{12}$	0.0635
$\text{ARIMA}(3,0,3)(0,1,1)_{12}$	0.0637
$\text{ARIMA}(3,0,1)(0,1,1)_{12}$	0.0644
$\text{ARIMA}(3,0,2)(0,1,1)_{12}$	0.0644
$\text{ARIMA}(3,0,2)(2,1,0)_{12}$	0.0645
$\text{ARIMA}(3,0,1)(2,1,0)_{12}$	0.0646
$\text{ARIMA}(4,0,2)(0,1,1)_{12}$	0.0648
$\text{ARIMA}(4,0,3)(0,1,1)_{12}$	0.0648
$\text{ARIMA}(3,0,0)(2,1,0)_{12}$	0.0661
$\text{ARIMA}(3,0,1)(1,1,0)_{12}$	0.0679

RMSE 값을 기준으로 볼 때 직접 고른 모델과 `auto.arima()` 둘 다 상위 4개의 모델에 포함됩니다.