시계열 모형을 사용해
암트랙 기차 월별 승객 수
예측 모형 만들기

과목 : 시계열 분석

교수 : 양기성 교수님

학과 : 정보통계보험수리학과

이름 : 20171421 김성연

제출일자 : 2022년 05월 08일

목 차

1. 연구 목적

현대사회의 교통수단으로써 기차의 위상은 자동차, 비행기에 비해 훨씬 떨어진다. 하지만 환경문제가 중요한 요소로 떠오르면서 기차의 위상은 조금씩 높아지고 있다. 승객 1명을 1㎞ 운송하는데 드는 에너지는 기차가 자동차, 비행기의 10분의 1 수준에 불과하다. 단순히 연료비가 적게 드는 것뿐만 아니라 한 사람을 이동시킬 때 발생하는 탄소 배출량이 적다는 의미다. 기차는 친 환경 교통수단으로 앞으로도 잘 생존할 것으로 보인다.

결국 기차라는 교통수단을 잘 활용해야 한다. 교통수단을 잘 활용하려면 효율적인 기차운행계획이 필요하고, 새로운 철도 수요를 잘 예측해야 한다. 이를 위해선 기본적으로 미래 승객 수를 예측하는 좋은 모형이 필요하다. 미래 승객수를 예측한다면 기차운행계획에 좋은 자료가 되고, 수요가 일반적인 상태와 차이가 있는지에 대한 파악도 가능하다. 이외에도 많은 중장기적 플랜도 세울 수 있다. 어떤 모형이 승객 수 예측에 뛰어난 지 살펴보려고 한다.

2. 연구 방법 소개

미국에서 와이오밍과 사우스다코타를 제외한 모든 주를 대상으로 영업하는 여객철도 공사인 암트랙에서 제공한 1991년 1월 ~ 2004년 3월까지의 월별 승객 수 데이터를 사용한다.

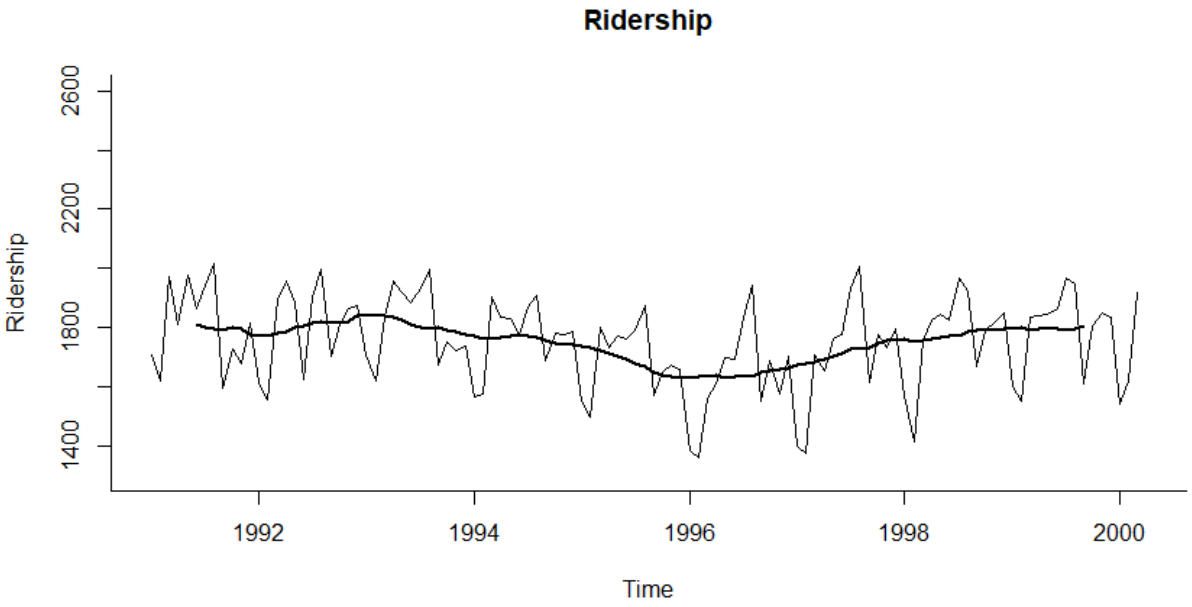
1991년 1월부터 2000년 3월까지의 데이터를 Train 데이터 셋으로, 2000년 4월부터 2002년 3월까지 24개월 데이터를 Validation 데이터 셋으로, 그리고 2002년 4월부터 2004년 3월까지 24개월 데이터를 Test 데이터 셋으로 구분한다. Train 데이터 셋으로 모형을 학습시킨 뒤 Validation 데이터 셋으로 모형성능을 비교한다. 가장 성능이 뛰어난 모형을 선택한 뒤 Train 데이터 셋과 Validation 데이터 셋을 사용해 선택한 모형을 다시 학습한 뒤 Test 데이터 셋으로 모형 성능을 평가한다.

모형을 사용하기 전 자료에 추세와 계절성을 시각적으로 확인하는 탐색적 자료 분석을 시행한다. 이후 총 6개의 시계열 모형을 사용해 자료를 분석하고 최종 모형을 결정한다. 1) Naive forecast with roll-forward partitioning 모형, 2) Trailing Moving Average with 12 months of window width 모형, 3) Holt-Winter's DES with additive trend, multiplicative seasonality, and multiplicative error 모형, 4) Holt-Winter's DES with additive trend, additive seasonality, and additive error 모형, 5) Regression model with quadratic trend and seasonal dummy variables 모형, 6) Regression model with quadratic trend and trigonometric functions 모형.

모형 비교에 사용하는 지표는 많이 사용하는 RMSE이다. 왜냐하면 승객 수를 과다예측한 경우 재정적 손실이 있으며 과소예측을 한 경우 고객의 서비스 만족도가 떨어지기 때문이다. 여러 지표를 사용하면 모형 비교에 혼란이 올 수 있기 때문에 모형 평가 지표로 RMSE 만을 사용하기로 한다.

3. 모형 적합

3-1. 탐색적 자료 분석



1991년 1월부터 2000년 3월까지에 암트랙 월별 승객 수 데이터를 이용한 시계열 그래프이다. 모형 구축을 Train 데이터 셋으로 하기 때문에 Train 데이터만 사용했다. 가운데 있는 굵은 선은 12-윈도우 중심이동평균 선이다.

우선 계절성이 눈에 띄게 관찰된다. 월별 데이터이기 때문에 계절성의 주기는 12로 추측되는데 이 그래프에서도 확인된다. 이런 이유로 중심이동평균의 윈도우를 12로 정했다. 또 12-윈도우 중심이동평균 선을 관찰해보니 1996년까지는 승객 수가 꾸준히 감소하다가 반등하여 2000년까지 증가하는 모양세다. 추세가 눈에 띄게 관찰된다.

3-2. Naive forecast with roll-forward partitioning 모형

시계열의 가장 최근 값을 예측 값으로 사용하는 단순(Naive) 예측 모형이다. 이 모형은 사용하는데 간편하면서도 최근의 정보가 미래를 예측하는데 관련성이 높다는 점을 반영했다. 단순한 모형이기 때문에 다른 복잡한 모형의 기준이 되는 Baseline 모델 역할을 기대한다.

이번 연구에서는 탐색적 자료 분석에서 계절성을 분명하게 관측했기 때문에 예측 값으로 직전 년도 동일 달의 데이터를 사용했다. 직전 년도 동일 달의 데이터를 사용하는 것은 Train과 Validation 데이터 셋 구분을 계속 바꾸는 roll-forward partitioning 방법이다. Train과 Validation 데이터 셋 구분을 명확하게 고정하는 다른 모형보다 유리한 측면이 있는 점을 감안해야 한다.

3-3. Trailing Moving Average with 12 months of window width 모형

시간적으로 연속하는 일정한 윈도우 동안의 관측치들의 평균을 예측 값으로 사용하는 이동평균 예측 모형이다. 탐색적 자료 분석에서 계절성 주기가 12로 추측되므로 계절성을 제거하기 위해 이동평균 윈도우를 12로 한다. 예측을 위해서는 미래 값을 알지 못하므로 이전 윈도우의 평균치를 계산한 후행이동평균 방식을 쓴다.

다만 이동평균 예측모형은 승객이 많은 시즌에는 과소예측을 하게 되고, 승객이 적은 시즌에는 과다예측을 하는 단점이 있다. 또 상승하는 추세에서는 과소예측을 하게 되고, 하락하는 추세에서는 과다예측을 하는 경향이 있다. 탐색적 자료 분석에서 추세와 계절성을 관찰했기 때문에 이 모형은 성능이 좋지 않을 것으로 추측된다.

3-4. Holt-Winter's DES with additive trend, multiplicative seasonality, and multiplicative error 모형

모형의 추세와 계절성이 전역적이라고 가정하지 않고 시간에 따라 변한다고 가정한 이중 지수평활을 기반으로 한 Holt-Winter의 지수 평활기법을 사용한 예측 모형이다. 이중 지수평활을 기반으로 하기 때문에 수준, 추세, 계절성 3가지 요인이 매 시점 업데이트 된다. 3가지 요인이 모든 과거의 값을 가중평균 하는데 가중치는 과거로 갈수록 지속적으로 감소한다. 최근 정보에 더 많은 가중치를 부여하지만 오래된 정보를 완전히 배제하지는 않는다는 아이디어에 근거한다.

요인 업데이트 방식이 추세는 가법, 계절성은 승법이며 오차도 승법으로 계산한다. 승법방식은 시간이 지남에 따라 값이 변할 때 사용하는 방식이고 가법방식은 값이 시간과 상관없이 일정할 때 사용한다.

3-5. Holt-Winter's DES with additive trend, additive seasonality, and additive error 모형

윗 모형과 대부분 동일한 모형이나 계절성 요인 업데이트 방식이 가법이고 오차도 가법으로 계산한다. 탐색적 자료 분석에서 요인 업데이트 방식이 가법인지 승법인지 여부는 뚜렷한 경우일때만 시각적으로 확인이 가능하기 때문에 두 모형을 모두 적합한 뒤 비교하기로 한다.

3-6. Regression model with quadratic trend and seasonal dummy variables 모형

통계에서 주로 사용하는 회귀모형을 이용하여 예측 값을 구하는 모형이다. 반응변수로 추세와 계절성을 사용한다. 탐색적 자료 분석을 보면 승객 수가 감소했다가 1996년을 기점으로 증가하기 때문에 일차선형추세보다 이차 추세항을 추가하는 것이 좋아 보인다. 계절성은 범주형 변수이므로 상수항을 포함해 12개의 더미변수로 사용한다.

다만 추세는 외부 충격에 따라 변하는데 본 모형은 이 부분을 고려하지 않았다. 회귀 모형의 예측력은 관측된 범위의 X값에서만 유효한데, 시계열 모형은 관측한 범위를 벗어나는 외삽(extrapolation)이 일어나기 때문이다. 탐색적 자료 분석에서 이차 추세 변수가 양수로 추정되기 때문에 월 승객수가 지속적으로 증가하지 않는다면 시간이 지날 수록 예측력은 떨어질 것으로 추측된다.

3-7. Regression model with quadratic trend and trigonometric functions 모형

윗 모형과 유사하나 계절성 부분을 더미변수를 사용하지 않고 사인과 코사인 함수를 이용한 모형이다. 계절성 패턴이 한 계절에서 다음 계절로 부드럽게 바뀔 때 유리하다. 계절성 패턴이 부드럽게 바뀐다면 윗 모형과 달리 변수 개수가 대폭 감소하므로 모델이 상대적으로 유연(flexible)해서 예측력이 높아진다. 다만 계절성 패턴의 부드러움은 주관적이므로 탐색적 자료분석에서 파악하기 힘들다. 그래서 두 모형을 모두 적합한 뒤 비교하기로 한다.

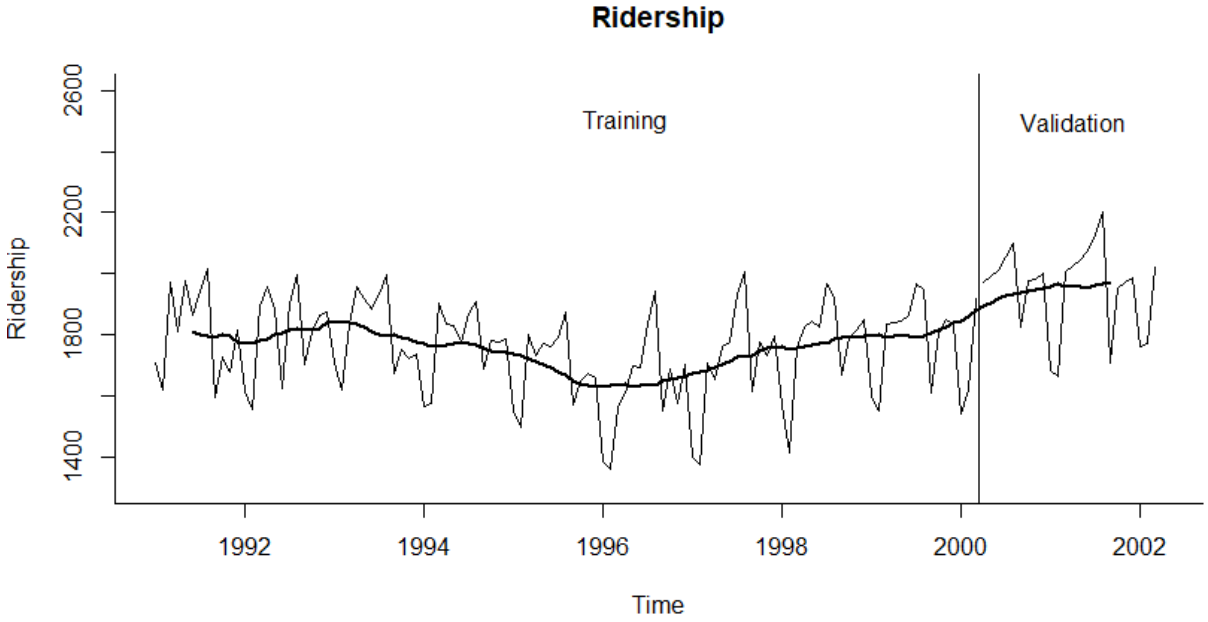
4. 최종 모형 선택

	심플한 모형	
모형	단순(Naive) 예측	이동 평균 예측
RMSE	111.48	206.07
	Holt-Winter's DES 모형	
모형	가법 추세, 승법 계절성/오차	가법 추세/계절성/오차
RMSE	109.40	115.07
	회귀 기반 2차 추세 모형	
모형	계절성을 더미변수로	계절성을 삼각함수로
RMSE	74.83	132.14

※ 단순 예측은 다른 모형과 달리 roll-forward 분할을 하여 단순한 RMSE 수치비교에서 유리할 수 있습니다. 그러므로 모형 성능 평가의 기준점으로만 사용해야합니다.

Validation 데이터 셋의 RMSE을 최소로 하는 Regression model with quadratic trend and seasonal dummy variables 모형을 최종 모형으로 선택한다. 동일 조건에서 계절성을 삼각함수로 한 것 보다 더미변수로 할 때 예측력이 많이 좋아졌다. 계절성 패턴이 부드럽지 않다고 확실히 말할 수 있다. 또 Holt-Winter's 모형에서 승법 계절성/오차 가정 모형이 예측력 부분에서 우수한 퍼모먼스

를 보인다. 계절성이 시간에 지남에 따라 값이 바뀐다고 할 수 있다. 그리고 이동 평균 예측 모형은 예상한 대로 성능이 많이 좋지 못하다.



앞서 살펴본 Train 기간에 추가로 Validation 기간까지 시각화 했다. 추세를 파악하기 위해 12-윈도우 이동평균선도 그렸다. Validation 기간에서 월 승객수가 지속적으로 증가하는 것을 알 수 있다. 외삽(extrapolation) 부분의 예측력은 보장할 수 없긴 하나 X 범위 밖 부분도 모형과 일치한다면 예측력이 좋을 수도 있고, 이번 모형에서 확인했다. 다만 실제 모형 성능은 Test 데이터 셋에서 확인하는데, Test 기간까지 예측력이 좋을 지는 다시 살펴봐야 한다.

5. 최종 모형 성능 평가

성능지표	RMSE	MAE	MPE	MAPE	MASE
Train 데이터	66.03	51.68	-0.15	2.98	0.64
Test 데이터	150.85	141.35	-7.16	7.16	1.76

※ Regression model with quadratic trend and seasonal dummy variables 모형을 사용했으며 Train와 Validation 기간인 1991년 1월부터 2002년 3월까지 데이터를 모형 학습에 사용했고 Test 기간인 2002년 4월부터 2004년 3월까지에 데이터를 모형 예측에 사용했다.

동일한 형태의 모형을 Train 기간을 학습 데이터로 사용하여 Validation 기간을 예측 할 때 RMSE 값은 74.83인 반면, Train와 Validation 기간을 학습 데이터로 사용하고 Test 기간을 예측 할 때

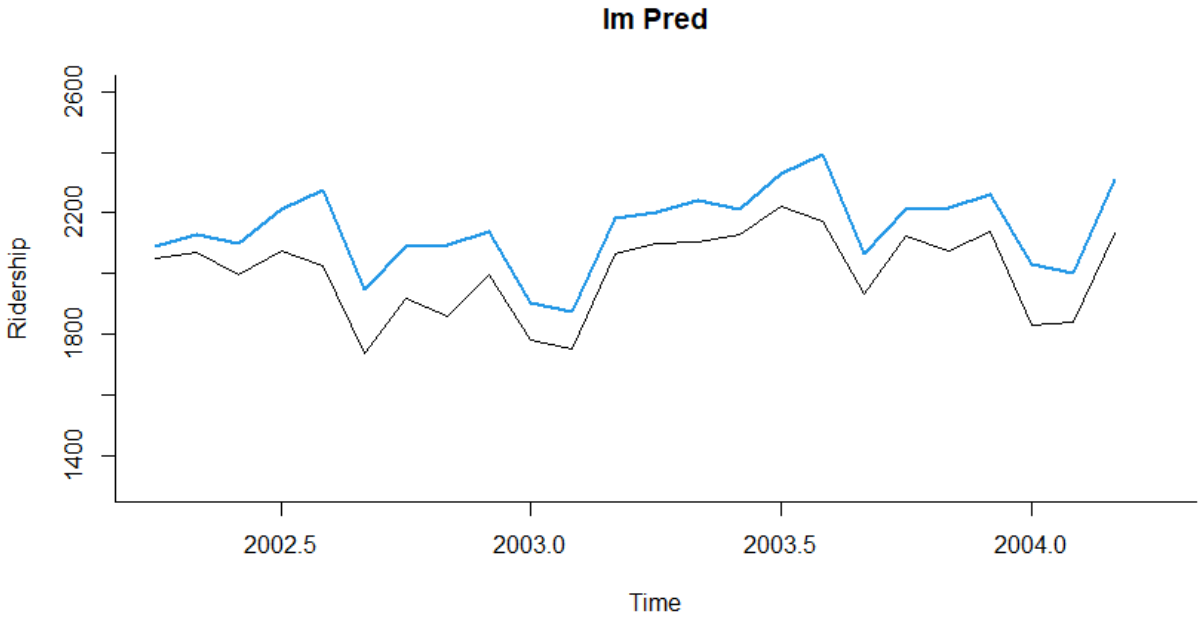
RMSE 값은 150.85로 눈에 띄게 높아졌다. 또 최종 모형을 학습한 자료와 비교하는 것과 테스트 자료와 비교했을 때에 모든 예측평가지표에서 큰 폭의 성능 저하를 겪었다. 물론 학습한 자료에 모형이 더 잘 적합하는 과 적합 문제는 빈번하게 발생하는 문제이다. 그리고 Train와 Validation 기간 데이터에 비해 Test 기간 데이터가 형태가 많이 다를 경우 모든 모형에서 오차가 커질 수도 있다.

다만 Test 기간에 모형 성능 저하 문제가 다른 모형에서도 일어나는지 관찰해볼 필요는 있다. 두 번째로 좋은 퍼포먼스를 보인 Holt-Winter's DES with additive trend, additive seasonality, and additive error 모형을 사용해본다.

성능지표	RMSE	MAE	MPE	MAPE	MASE
Train 데이터	56.56	44.63	0.10	2.53	0.56
Test 데이터	86.58	73.73	0.42	3.71	0.92

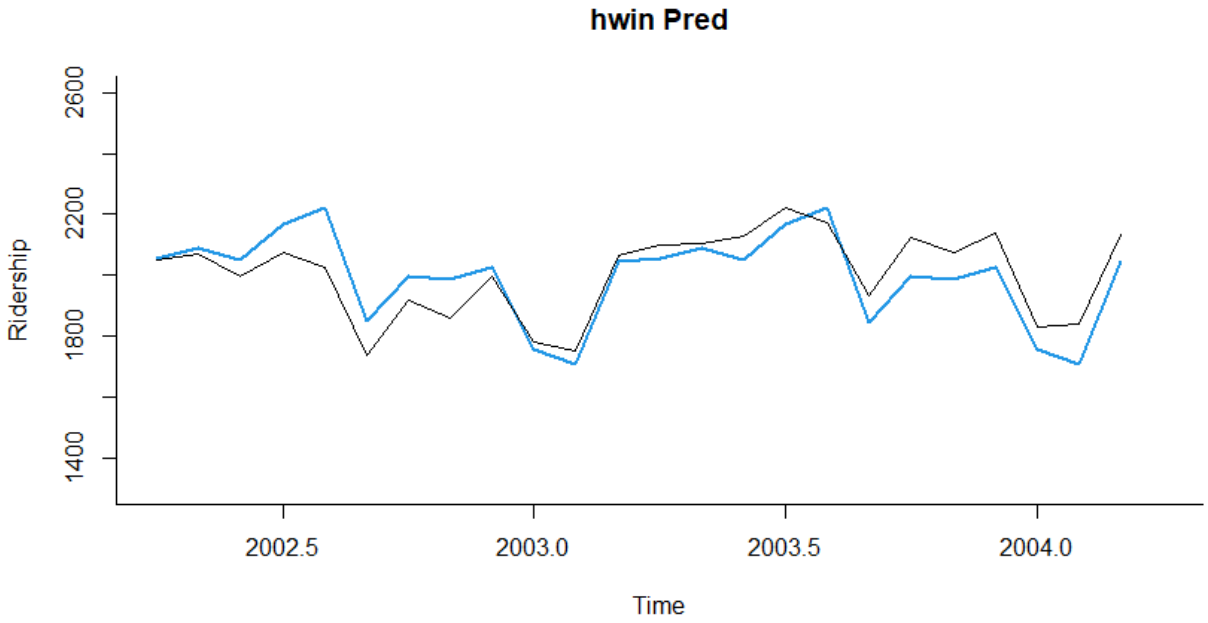
※ Holt-Winter's DES with additive trend, additive seasonality, and additive error 모형을 사용했으며 Train와 Validation 기간인 1991년 1월부터 2002년 3월까지 데이터를 모형 학습에 사용했고 Test 기간인 2002년 4월부터 2004년 3월까지에 데이터를 모형 예측에 사용했다.

회귀 기반 모형보다 Holt-Winter 모형이 RMSE 값이 86.58로 Test 기간에서 더 좋은 퍼포먼스를 보인다. 다만 이 모형도 학습한 자료에 모형이 더 잘 적합하는 과 적합 문제는 해결해야 하는 숙제이다. 왜 이런 결과가 나오는지 그래프로 확인해본다.



※ Regression model with quadratic trend and seasonal dummy variables 모형을 사용했으며 Train와 Validation 기간인 1991년 1월부터 2002년 3월까지 데이터를 모형 학습에 사용했고 Test 기간인 2002년 4월부터 2004년 3월까지에 데이터를 모형 예측에 사용했다. 파란색은 모형으로 구한 예측 값이고 검은색은 실제 값이다.

실제 값에 비해 모형이 과다예측을 한다. 실제로는 월 승객수가 계속 상승하지 않기 때문에 모형이 과다예측을 한 것으로 보인다. Validation 기간에서 좋은 성능을 보인 것이 우연이라고도 생각할 수 있다.



※ Holt-Winter's DES with additive trend, additive seasonality, and additive error 모형을 사용했으며 Train와 Validation 기간인 1991년 1월부터 2002년 3월까지 데이터를 모형 학습에 사용했고 Test 기간인 2002년 4월부터 2004년 3월까지에 데이터를 모형 예측에 사용했다. 파란색은 모형으로 구한 예측 값이고 검은색은 실제 값이다.

앞 모형보다 예측 값이 실제 값을 잘 따라간다. Test 기간에서는 확실히 Holt-Winter 모형이 좋다. global time trend가 잘 성립하지 않는, 추세가 급격히 변하는 모형에서는 회귀 기반 모형의 성능이 떨어진다. 다만 Holt-Winter 모형이 항상 우수하다는 근거 또한 부족하다. 추세에 영향을 주는 외부 변수를 찾고 두 모형을 복합적으로 이용한다면 예측력이 더 좋아질 것이다.

참고자료

'철도의 가치' 기사

https://www.edaily.co.kr/news/read?newsId=01374326628947568&mediaCodeNo=257

'R기반의 데이터마이닝을 이용한 시계열 예측 분석 및 활용' 책

(저자 : Galit Shmueli, Kenneth C. Lichtendahl Jr.

역자 : 신택수, 홍태호

출판사 : 청람)