|  |
| --- |
| **고성능 인공지능 알고리즘 활용**  **2025년도 KTX 수송수요 예측 연구** |

**- 자문보고서 -**

2024. 06.



**제 출 문**

**한국철도공사 연구원장 귀하**

**본 보고서를**

**「** **고성능 인공지능 알고리즘 활용 2025년도 KTX 수송수요 예측 연구 」**

**자문 최종보고서로 제출합니다.**

2024. 06.

**인천대학교 글로벌정경대학 교수 김 경 원**

목차

1. 연구 개요

1-1. 연구의 배경 및 목적

1-2. 연구의 대상 및 방향

2. 수요예측 데이터

2-1. 기술적 분석

2-2. 데이터 전처리

3. 예측 방법론

3-1. 머신러닝 기반 예측

3-2. 딥러닝 기반 예측

4. 수송수요 예측 결과

4-1. 2024년도 예측 성능 검증

4-2. 2025년도 예측 결과

5. 결론

5-1. KTX 수요예측 결론

5-2. 향후 연구 방향

# 연구 개요

## 1.1연구의 배경 및 목적

### 1.1.1 연구의 배경

○ KTX 수송수요(승차인원수)를 정확하게 예측하기 위해 고성능의 예측 모델의 개발이 필요함

- 최근 각광받는 인공지능 알고리즘을 활용하여 수송수요 예측 수행

- 다양한 형태의 인공지능 알고리즘을 활용하여 시계열 데이터인 수송수요에 적용하여 KTX 노선별 예측 수행

### 1.1.2 연구의 목적

○ 2025년도 KTX 수송계획 및 열차운행계획 수립을 위해 필요한 2024년도 단기 및 2025년도 중기 수송수요예측을 위한 고성능 인공지능 모델을 활용한 수송수요를 예측 방법을 구축하는 것이 주 목적으로 함

○ 기존 예측 방법론의 고도화를 달성하기 위해서, 시계열 데이터인 KTX 노선별 수송수요 대상 머신러닝 및 딥러닝 기반의 전통적인 알고리즘부터 최신 알고리즘을 활용하여 예측 분석을 진행함

- KTX 수요에 영향을 줄 수 있는 다양한 파생변수들을 생성하여 모델에 반영함으로써 향후 더욱 복합적인 변수들의 반영을 쉽게 하여 성능 향상에 쉽게 기여할 수 있도록 개발

- 머신러닝과 딥러닝은 과거부터 현재까지 수요예측 성능 향상에 기여할 수 있도록 빠르게 진화하였기 때문에 이를 반영하여 개발

- KTX 노선별 주중, 주말, 그리고 전체 수송수요 예측성능을 검증하고 가장 성능이 높은 모델을 중심으로 2025년도 수송수요를 예측하도록 개발

## 1.2연구의 대상 및 방향

### 1.2.1 연구의 대상

○ KTX 수송수요 예측을 위해 총 5개의 KTX 노선을 대상으로 함

- 경부선, 경전선, 동해선, 전라선, 호남선

- 각 노선은 수송수요의 차이가 있으므로 노선 별로 구분하여 예측

○ 수송수요 예측은 주말, 주중, 그리고 전체를 대상으로 함

- 주말의 기준은 “금토일”이며, 주중의 기준은 “월화수목”으로 함

- 각 요일에 따라서 수송수요의 차이가 있기 때문에 구분하여 예측

○ 수송수요 예측의 시간 빈도는 “월(Month)”을 기준으로 함

- 2025년도 수송수요 예측을 위해 일별과 월별 데이터를 활용할 수 있는데, 본 연구에서는 2025년도 1월부터 12월까지의 수요가 필요하기 때문에 입력과 출력에 사용되는 데이터의 시간은 월을 사용함

○ 각 노선별 주말/주중/전체로 구분하여 총 15개의 월별 KTX 수송수요 예측

|  |  |  |  |
| --- | --- | --- | --- |
| **시간빈도** | **노선** | **대상** | **총 15개 월별 KTX 수송수요** |
| **월별** | **경부선** | **전체** | 월별 경부선 전체 수요 |
| **주말** | 월별 경부선 주말 수요 |
| **주중** | 월별 경부선 주중 수요 |
| **경전선** | **전체** | 월별 경전선 전체 수요 |
| **주말** | 월별 경전선 주말 수요 |
| **주중** | 월별 경전선 주중 수요 |
| **동해선** | **전체** | 월별 동해선 전체 수요 |
| **주말** | 월별 동해선 주말 수요 |
| **주중** | 월별 동해선 주중 수요 |
| **전라선** | **전체** | 월별 전라선 전체 수요 |
| **주말** | 월별 전라선 주말 수요 |
| **주중** | 월별 전라선 주중 수요 |
| **호남선** | **전체** | 월별 호남선 전체 수요 |
| **주말** | 월별 호남선 주말 수요 |
| **주중** | 월별 호남선 주중 수요 |

<표> 1‑1. 본 연구의 분석 대상인 총 15개의 월별 KTX 수송수요

### 1.2.2 연구의 방향

○ **[데이터 전처리]** KTX 수송수요 예측에 실제 추가 변수들이 도움이 되는지 여부를 확인하고 필요시 반영을 위해서 다양한 파생 또는 외부 변수 후보들을 준비함

- 파생변수(Derived Variables)란, 기존 정보나 변수를 활용하여 재계산하거나 조합하여 새롭게 추가 반영된 변수를 의미

- 과거 수송수요가 현재 또는 미래의 수송수요에 영향을 줄 수 있기 때문에 과거 1개월~12개월(1년) 전의 수송수요 파생변수를 생성함

- 시간정보를 포함하고 있는 시계열 데이터를 통해 각 월의 실제 날짜 수, 주말 수, 주중 수, 공휴일 수, 명절 수 등의 파생변수 생성함

- COVID-19 종료 이벤트를 반영하기 위해 예방접종 인원수, 격지자 수, 사망자 수, 정부대응 지수 등의 파생변수를 결합함

- 경제상황의 변동을 모델에 반영하기 위해서 한국의 주식 시장 지표와 소비자의 물가 수준과 관련이 있는 지표들을 결합함

- 그 외에 제공받은 수송수요를 포함한 공급좌석 정보, 열차 정보, 운행 정보 등을 재계산하여 파생변수들을 생성함

○ **[데이터 분리]** 수송수요 예측 모델링에 사용되는 데이터는 크게 학습(Training), 검증(Validate), 예측(Test) 3가지로 구분되어 활용됨

- 모델링에 사용되는 데이터를 학습(Training)이라고 하며, 모델링 완료 후 학습에 사용되지 않은 데이터로 예측 성능을 검증하는 데 이를 검증(Validate)이라고 함, 그리고 검증과정에서 성능이 높은 모델을 사용하여 필요한 기간의 수송수요를 예측하는 것을 예측(Test)라고 함

- 학습과 검증, 그리고 예측에 사용되는 데이터의 기간은 아래와 같음

- 학습(Training) 기간: 2015년 1월 ~ 2023년 3월

- 성능(Validate) 기간: 2023년 4월 ~ 2024년 3월

- 예측(Test) 기간: 2024년 4월 ~ 2025년 12월

○ **[모델링]** 수송수요 예측에 활용되는 모델은 머신러닝(Machine Learning) 및 딥러닝(Deep Learning) 기반으로 시계열 수송수요를 분석하여 성능을 비교함

- 머신러닝 알고리즘은 딥러닝 알고리즘보다 성능은 낮은 편으로 알려져 있지만 학습과 예측에 소요되는 시간이 매우 빠르고, 데이터의 품질이 좋은 경우 딥러닝에 견주어도 성능이 떨어지지 않는 알고리즘으로 수송수요 예측 성능 비교에 활용되는 베이스라인(Baseline) 모델로 활용됨

- 딥러닝은 데이터의 종류에 따라서 활용될 수 있는 알고리즘이 개발되고 진화되고 있으며, 시계열 데이터에 활용될 수 있는 알고리즘은 다층 퍼셉트론(Multi-layer Perceptron, MLP), 순환신경망(Recurrent Neural Network, RNN), 그리고 최근 생성형AI에 활용되고 있는 트랜스포머(Transformer)가 있음

- 월별 데이터의 특성상 데이터의 개수가 많지 않아서 트랜스포머를 제외한 딥러닝의 MLP 및 RNN 계열에서 최근까지 개발된 알고리즘들을 비교 모델로 활용함

|  |  |  |  |
| --- | --- | --- | --- |
| **모델링 종류** | **알고리즘 분야** | **이름** | **개발연도** |
| **Machine Learning** | **Bagging** | RandomForest | 2001 |
| **Boosting** | XGBoost | 2014 |
| LightGBM | 2016 |
| CatBoost | 2017 |
| **Deep Learning** | **MLP** | MLP | 2010 |
| NBEATS | 2019 |
| NHITS | 2021 |
| NBEATSx | 2021 |
| TiDE | 2023 |
| DeepNPTS | 2023 |
| **RNN** | RNN | 2014 |
| LSTM | 2014 |
| GRU | 2014 |
| DilatedRNN | 2017 |
| TCN | 2018 |

<표> 1‑2. 수송수요 예측에 사용되는 머신러닝 및 딥러닝 기반 알고리즘 종류

# 수요예측 데이터

## 2-1. 기술적 분석

○ KTX 수송수요의 통계량과 시각화를 통해 예측하고자 하는 승차인원수의 패턴을 확인 함으로써 데이터 준비 또는 전처리의 방향과 모델 적용의 방향을 확인함

○ **[노선별 승차인원수 통계량]** 장기간의 월별 데이터로 구성된 승차인원수를 이해하기 위해, 전체/주말/주중 승차인원수의 중심값인 평균과 퍼진 정도인 표준편차, 그리고 최대기록인 최대값을 확인함

- 전체 요일을 기준으로 경부선 > 호남선 > 전라선 > 경전선 > 동해선의 순서대로 승차인원수가 감소하고 있음

- 경부선의 경우 나머지 노선들의 총합(약 163만)보다 많은 약 241만명의 승객이 KTX를 이용하고 있음

- 동해선을 제외한 모든 노선에서 주중보다 주말에 승차인원수가 많음

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **주운행선** | **전체주중주말** | **월별 승차인원수** | | |
| **평균** | **표준편차** | **최대값** |
| **경부선** | **전체** | 2,413,616 | 1,269,865 | 6,921,600 |
| **주말** | 1,219,760 | 654,927 | 3,222,240 |
| **주중** | 1,193,856 | 633,191 | 3,699,360 |
| **경전선** | **전체** | 418,350 | 227,084 | 1,432,442 |
| **주말** | 211,947 | 117,683 | 673,228 |
| **주중** | 206,402 | 112,380 | 759,214 |
| **동해선** | **전체** | 113,837 | 234,642 | 1,218,256 |
| **주말** | 54,068 | 111,225 | 527,740 |
| **주중** | 59,769 | 124,245 | 690,516 |
| **전라선** | **전체** | 423,697 | 251,796 | 1,534,990 |
| **주말** | 217,081 | 129,543 | 708,674 |
| **주중** | 206,615 | 125,473 | 826,316 |
| **호남선** | **전체** | 683,922 | 354,597 | 2,176,084 |
| **주말** | 343,749 | 182,679 | 999,630 |
| **주중** | 340,172 | 177,723 | 1,176,454 |

<표> 2‑1. 노선별 전체/주중/주말 월별 승차인원수 통계량

- 승차인원수의 중요도로 보았을 때, 모델링에서도 전체 > 주말 > 주중의 순서로 정확한 예측치 필요할 것으로 예상됨

- 대부분의 노선에서 표준편차는 평균의 약 절반(50%) 정도의 높은 수치를 보이고 최대값은 평균의 약 3배(300%)를 나타낼 정도로 월별 승차인원수가 변동이 높음

- 승차인원수의 변동을 모델에 담아내기 위해, 시간정보에서 월별 속성과 일수, 공휴일수, 명절수 등 각 월이 담고 있는 패턴을 데이터 전처리 과정에 반영하는 것이 예측 성능을 높이는 방향일 수 있음

○ **[노선별 승차인원수 시각화]** 장기간의 월별 데이터로 구성된 승차인원수를 이해하기 위해, 노선별로 비교하여 어떠한 차이가 있는지 그리고 전체/주말/주중 요일 변화에 따른 차이도 시각화로 비교 확인함

- 각 노선별 승차인원수의 차이가 크기 때문에, 동등한 비교를 위해 각 노선별 승차인원수 최소 및 최대값을 모두 0과 1사이의 값으로 스케일링(Scaling)하여 비교함으로써 각 노선별 움직임을 용이하게 비교함

|  |
| --- |
|  |
| (가) 경부선의 승차인원수 |
|  |
| (나) 경전선의 승차인원수 |
|  |
| (다) 동해선의 승차인원수 |
|  |
| (라) 전라선의 승차인원수 |
|  |
| (마) 호남선의 승차인원수 |
| <표> 2‑2. 2015년 1월 ~ 2024년 3월 기간의 전체/주말/주중 별 승차인원수 비교 |

- 모든 노선에서 전체/주말/주중 요일에 따른 승차인원수의 증가와 감소와 같은 패턴은 다르게 나타나는 것을 알 수 있음

- 노선에 따라 차이가 있지만 주말과 주중의 증가와 감소 패턴은 반대로 나타나는 시점들이 다수 존재하기 때문에, 동일한 노선이라고 하더라도 전체/주말/주중의 승차인원수 예측에 가장 높은 성능을 보이는 알고리즘은 다를 수 있음

|  |
| --- |
|  |
| (가) 전체 요일의 승차인원수 |
|  |
| (나) 주말의 승차인원수 |
|  |
| (다) 주중의 승차인원수 |
| <표> 2‑3. 2015년 1월 ~ 2024년 3월 기간의 노선별 승차인원수 비교 |

- 각 노선별 승차인원수의 크기 차이는 있지만 전체/주말/주중 요일에 따라 증가와 감소와 같은 변화 패턴은 모든 노선이 매우 유사하게 움직이고 있음이 확인됨

- 특히 COVID-19을 겪고 있던 기간에선 증가와 감소의 패턴 뿐만 아니라 노선별 승차인원수의 크기 조차도 매우 유사함

- 따라서 동일한 요일범위(전체/주말/주중)라면 노선이 다르더라도 가장 예측 성능이 높게 나타나는 알고리즘은 모두 동일할 수 있으며, 모델링의 학습과정에 COVID-19와 관련된 변수들도 반영하여 예측 과정에선 COVID-19 효과를 제거하는 것처럼 전처리 하는 것이 예측 성능을 높이는 방법이 될 수 있음

## 2-2. 데이터 전처리

○ KTX 수송수요 예측에 도움이 될 수 있는 다양한 파생 변수 후보들을 생성함

- 과거 수송수요가 현재 또는 미래의 수송수요에 영향을 줄 수 있기 때문에 과거 1개월~12개월(1년) 전의 수송수요 파생변수를 생성함

- 시간정보를 포함하고 있는 시계열 데이터를 통해 각 월의 실제 날짜 수, 주말 수, 주중 수, 공휴일 수, 명절 수 등의 파생변수 생성함

- COVID-19 종료 이벤트를 반영하기 위해 예방접종 인원수, 격지자 수, 사망자 수, 정부대응 지수 등의 파생변수를 결합함

- 경제상황의 변동을 모델에 반영하기 위해서 한국의 주식 시장 지표와 소비자의 물가 수준과 관련이 있는 지표들을 결합함

- 그 외에 제공받은 수송수요를 포함한 공급좌석 정보, 열차 정보, 운행 정보 등을 재계산하여 파생변수들을 생성함

|  |  |  |  |
| --- | --- | --- | --- |
| **특성** | **변수명** | **예시** | **정의** |
| **요일구분** | **전체주중주말** | 전체 | 수요예측 대상 요일로 주말/주중/전체 |
| **열차종** | **주운행선** | 경부선 | KTX 주요 5개 노선 경부/경전/동해/전라/호남선 |
| **과거**  **승차인원수** | **승차인원수\_Lag1** | 3,464,111.00 | 1개월 전 승차인원수 |
| **승차인원수\_Lag2** | 3,318,669.00 | 2개월 전 승차인원수 |
| **승차인원수\_Lag3** | 3,647,548.00 | 3개월 전 승차인원수 |
| **승차인원수\_Lag4** | 3,547,176.00 | 4개월 전 승차인원수 |
| **승차인원수\_Lag5** | 3,643,417.00 | 5개월 전 승차인원수 |
| **승차인원수\_Lag6** | 3,291,100.00 | 6개월 전 승차인원수 |
| **승차인원수\_Lag7** | 3,352,224.00 | 7개월 전 승차인원수 |
| **승차인원수\_Lag8** | 3,279,768.00 | 8개월 전 승차인원수 |
| **승차인원수\_Lag9** | 3,274,361.00 | 9개월 전 승차인원수 |
| **승차인원수\_Lag10** | 6,921,600.00 | 10개월 전 승차인원수 |
| **승차인원수\_Lag11** | 3,281,372.00 | 11개월 전 승차인원수 |
| **승차인원수\_Lag12** | 3,152,500.00 | 1년 전 승차인원수 |
| **시간정보** | **운행년월** | 45,352.00 | 해당 열차가 운행된 연도와 월 |
| **일수** | 31.00 | 해당 월에 포함된 날짜의 수 |
| **주말수** | 15.00 | 해당 월에 포함된 주말(금토일) 수 |
| **주중수** | 16.00 | 해당 월에 포함된 주중(월화수목) 수 |
| **공휴일수** | 1.00 | 해당 월에 포함된 대체휴일을 포함한 공휴일의 수 |
| **명절수** | - | 해당 월에 포함된 대체휴일을 포함한 명절의 수 |
| **좌석 및 운행정보** | **공급차량수** | 62,658.00 | 운행될 수 있는 총 열차의 수 |
| **공급좌석합계수** | 3,301,402.00 | 열차 내에 판매할 수 있는 좌석의 수 |
| **승차수입금액** | 113,298,033,576.00 | 판매된 좌석의 총금액 |
| **승차인원수** | 3,473,501.00 | 판매된 좌석의 수 |
| **승차연인거리** | 8,239,441,412.00 | 승객이 타고간 거리 |
| **좌석거리** | 13,516,340,607.00 | 공급좌석이 이동가능한 거리 |
| **1인당수입율** | 1,006,017.52 | "승차수입금액"을 "승차인원수"로 나눈 값으로, 1인당 평균 수입미 |
| **공급대비승차율** | 32.33 | "승차인원수"를 "공급좌석합계수"로 나눈 값으로, 좌석 판매의 집중도 |
| **운행대비고객이동** | 52.84 | "좌석거리"를 "승차연인거리"로 나눈 값으로, 승객의 이용 집중도 |
| **관광** | - | 관광으로 편성된 열차 수 |
| **일반** | 97.00 | 일반으로 편성된 열차 수 |
| **일반/관광** | 408.00 | 일반/관광으로 편성된 열차 수 |
| **대수송** | - | 대수송으로 편성된 열차 수 |
| **임시** | - | 임시로 편성된 열차 수 |
| **확정** | 505.00 | 확정으로 편성된 열차 수 |
| **시발역** | 155.00 | 열차가 출발하는 시발역 종류의 수 |
| **종착역** | 155.00 | 열차가 도착하는 종착역 종류의 수 |
| **시발종착역** | 248.00 | 열차가 운행하는 "시발역+종착역" 노선의 수 |
| **열차운행횟수** | 3,791.00 | 운행을 한 총 열차의 수 |
| **외부환경** | **Stringency Index** | 1,463.82 | 코로나 진행정도 지수 |
| **Government Response Index** | 1,775.99 | 정부의 코로나 대응정도 지수 |
| **International Movement Restrictions** | 62.00 | 국가간 이동 제한정도 지수 |
| **Death People** | 82,667.00 | 코로나 사망자 수 |
| **Vaccinated People** | 1,246,537,550.00 | 백신접종을 시작한 인원수 |
| **Fully Vaccinated People** | 1,014,704,524.00 | 백신접종이 완료된 인원수 |
| **Containment People** | 1,918.90 | 격리된 인원수 |
| **Confirmed People** | 10,586,338.00 | 코로나 확진자 수 |

<표> 2‑4. 데이터 전처리 결과 모델의 입력으로 활용될 수 있는 독립변수명, 예시와 의미

- 예측 대상인 승차인원수는 종속변수(Y)이고 파생변수를 포함한 나머지 32개의 변수들은 독립변수(X)로 사용됨

- 전체/주말/주중 노선별 승차인원수 예측에 32개의 모든 독립변수가 사용되는 것은 아니며 여러 변수들의 조합을 활용하여 가장 예측 성능이 높은 독립변수의 조합이 입력되어 예측함

○ 모델링에 사용되는 학습(Training) 데이터, 모델링 완료 후 학습에 사용되지 않은 데이터로 예측 성능을 검증하는 검증(Validate) 데이터, 그리고 검증결과상 가장 성능이 높은 모델로 필요한 기간의 수송수요를 예측하는 예측(Test) 데이터로 구분함

|  |  |  |
| --- | --- | --- |
| **구분** | | **설명** |
| **입력변수** | | - 종속변수(Y): 승차인원수 - 독립변수(X): 파생편수를 포함하여 총 32개 |
| **분석기간** | **경부선** | 2015년 1월 ~ 2023년 3월 |
|  | **경전선** | 2015년 1월 ~ 2023년 3월 |
|  | **동해선** | 2015년 1월 ~ 2023년 3월 |
|  | **전라선** | 2015년 1월 ~ 2023년 3월 |
|  | **호남선** | 2015년 1월 ~ 2023년 3월 |
| **검증기간** | | 2023년 4월 ~ 2024년 3월 |
| **최종예측기간** | | 2024년 4월 ~ 2025년 12월 |

<표> 2‑5. 승차인원수 예측을 위한 데이터 전처리 정리

3. 예측 방법론

3-1. 머신러닝 기반 예측

3-2. 딥러닝 기반 예측

> \*\*Machine Learning:\*\*

> - RandomForest

> - XGBoost

> - LightGBM

> - CatBoost

> \*\*Deep Learning(MLP):\*\*

> - MLP (2010)

> - NBEATS (2019)

> - NHITS (2021)

> - NBEATSx (2021)

> - TiDE (2023)

> - DeepNPTS (2023)

> \*\*Deep Learning(RNN):\*\*

> - RNN (2014)

> - LSTM (2014)

> - GRU (2014)

> - DilatedRNN (2017)

> - TCN (2018)

> - DeepAR (2020) # 1D only

# 4. 수송수요 예측 결과

## 4-1. 2024년도 예측 성능 검증

## 4-2. 2025년도 예측 결과