JANT 한국항행학회논문지
Journal of Advanced Navigation Technology

항공 우주 해상

J. Adv. Navig. Technol. 28(4): 472-479, Aug. 2024

도심 항공 모빌리티 준비를 위한 승객 수요 예측 : 김포-제주 노선 사례 연구

Passenger Demand Forecasting for Urban Air Mobility Preparation: Gimpo-Jeju Route Case Study

김 정 훈* · **조 희 덕** · **최 선 미**
대한항공 항공기술연구원

Jung-hoon Kim* · **Hee-duk Cho** · **Seon-mi Choi**
Research and Development Center, Korean Air, Daejeon 34054, Korea

[요 약]

세계 전체인구의 절반이 도시에 거주하고 있고, 지속적인 도시화가 진행되고 있으며 2050년 경에는 도시인구가 전체 인구의 2/3를 초과할 것으로 예상하고 있다. 이러한 현상을 해소하기 위해 우리나라 정부에서는 새로운 도심 항공 모빌리티(UAM; urban air mobility) 산업 생태계 구축에 심혈을 기울이고 있다. 항공사 또한 UAM 산업 생태계에 속해 있으며 안전운항, 승객의 안전, 기재 운영 효율성, 정시성 등의 효율성 제고에 대한 준비를 하고 있다. 본 연구는 2019년부터 2023년까지 대한항공의 김포발 제주행 노선의 일일 승객 수에 대한 시계열 데이터를 활용하여 수요 예측을 수행한다. 이를 위해 SARIMA, Prophet, CatBoost, Random Forest와 같은 통계적 및 기계 학습 모델을 적용한다. 다양한 모델을 통해 승객 수요 패턴을 효과적으로 포착할 수 있는 방법을 평가하였고, 머신러닝 기반의 Random Forest 모델이 가장 뛰어난 예측 결과를 나타냈다. 연구 결과는 항공 산업에서 정확한 수요 예측을 위한 최적의 모델을 제시하여 운영 계획 및 자원 할당에 필요한 기초정보를 제공할 것이다.

[Abstract]

Half of the world's total population lives in cities, continuous urbanization is progressing, and the urban population is expected to exceed two-thirds of the total population by 2050. To resolve this phenomenon, the Korean government is focusing on building a new urban air mobility (UAM) industrial ecosystem. Airlines are also part of the UAM industry ecosystem and are preparing to improve efficiency in safe operations, passenger safety, aircraft operation efficiency, and punctuality. This study performs demand forecasting using time series data on the number of daily passengers on Korean Air's Gimpo to Jeju route from 2019 to 2023. For this purpose, statistical and machine learning models such as SARIMA, Prophet, CatBoost, and Random Forest are applied. Methods for effectively capturing passenger demand patterns were evaluated through various models, and the machine learning-based Random Forest model showed the best prediction results. The research results will present an optimal model for accurate demand forecasting in the aviation industry and provide basic information needed for operational planning and resource allocation.

Key word : Urban air mobility, Fleet operator, Demand forecasting, Machine learning.

http://dx.doi.org/10.12673/jant.2024.28.4.472

Received 2 July 2024; **Revised** 20 August 2024
Accepted (Publication) 23 August 2024 (30 August 2024)

***Corresponding Author; Jung-hoon Kim**

Tel: +82-42-868-6436
E-mail: kim.jungh@koreanair.com

Ⅰ. 서 론

21세기에 접어들며, 세계는 전례 없는 속도로 도시화를 경험하고 있다. 이러한 변화는 인간의 생활양식과 도시 기능에 깊은 영향을 미치고 있으며, 특히 도시 교통 문제는 중대한 사회적 도전 과제로 부상하고 있다. 유엔의 예측에 따르면, 2050년까지 세계 인구의 약 2/3가 도시에 거주할 것으로 예상되며 [1], 이는 도시 인프라에 대한 수요가 크게 증가할 것임을 시사한다. 대한민국에서는 서울, 부산, 인천 등의 대도시에 인구가 집중되고 있으며, 이로 인한 교통 혼잡은 도시의 지속 가능한 발전을 저해하는 주된 요인으로 작용하고 있다. 이에 따라, 효율적인 도시 교통 시스템 구축은 시급한 과제로 인식되고 있다. 3차원 공간을 활용하는 UAM(urban air mobility)은 도시 상공을 이용한 교통 수단으로, 기존 도로 교통 시스템의 한계를 넘어 교통 혼잡을 완화할 수 있는 혁신적인 솔루션으로 각광받고 있다.

현재 UAM은 전 세계적으로 개발 및 시험 운행 단계에 있으며, 상용화를 향해 나아가고 있는 국가들이 있다. 제주도는 전국 최초로 관광형 UAM 상용화를 목표로 하고 있고, 서울시는 2026년부터 응급의료 분야에서 UAM을 활용할 계획을 발표했다. 미국과 유럽을 중심으로 UAM 시장의 성장이 기대되며, 현대, GM 등 자동차 제조사와 에어버스, 보잉 등 항공기 제조사들이 eVTOL(electric vertical take-off and landing) 개발에 참여하고 있다. 그러나 아직 UAM이 완전히 상용화된 국가는 없으며, 각국은 기술 개발과 법적, 안전 문제 해결을 위한 준비를 진행 중이다 [2].

UAM의 주 고객은 도심 내의 교통혼잡을 피해 시간을 단축해야하는 상황이나, 정확한 시간에 목적지에 도착해야 하는 수요가 많을 것으로 예상된다. 이러한 UAM 산업에서는 정시성은 매우 중요한 요인이다. 정시성은 항공사나 다른 운송업체가 스케줄에 따라 정시에 운행을 하는 능력을 의미하며 [3], 이는 고객 만족도와 서비스 품질에 중요한 요소로 작용한다. 수요예측과 정시성은 밀접한 관계를 가지고 있다. 수요예측이 정확할수록, 항공사나 조업사, 버티포트 관계자는 자원을 효율적으로 배치하고 운영을 계획할 수 있다. 이는 정시성을 향상하는 데 도움이 된다. 반대로, 수요예측이 부정확하다면, 자원이 낭비되거나 부족해질 수 있으며, 이는 정시성을 저해하고 고객 만족도를 떨어뜨릴 수 있다. 따라서, 수요예측의 정확성은 정시성을 향상하는 데 중요한 역할을 한다. 이는 항공사나 여러 이해관계자가 안전 운항, 승객의 안전, 기재 운영의 효율성, 그리고 정시성을 제고하는 데 있어 중요한 역할을 한다.

이 연구는 대한항공의 김포-제주 국내 비행편 운항 데이터를 활용하여 수요예측을 하기 위한 여러 모델을 적용하여 미래 도시 교통 수요를 예측하는 모델을 개발함으로써 UAM 산업의 효율적인 계획과 운영을 위한 기초 자료를 제공하고자 한다. 현 시점 UAM은 개발 및 시험 운행단계로 실제 운항데이터는 존재하지 않으며, UAM 운항의 성격을 반영하여 짧은 비행시간과 하루 단위 많은 비행이 있는 김포 공항을 기준으로 김포-제주 노선 운항스케줄 데이터를 연구데이터로 활용하였다. UAM은 짧은 비행시간과 많은 편수의 비행을 특성으로 지니며, 운항 데이터 중 가장 비슷한 성격을 지니는 김포-제주 노선의 운항 데이터를 연구 데이터로 선정하였다. 2019년 1월 1일부터 2023년 12월 31일까지 대한항공에서 수집한 약 55000 편의 운항 데이터를 시계열 데이터로 전처리하였으며, 시계열데이터를 분석을 하기 위한 모델로 SARIMA (seasonal autoregressive integrated moving average) , Prophet, CatBoost (categorical boosting), Random Forest 모델을 선정하였다. SARIMA 모델은 시간의 흐름에 따른 계절적 패턴을 효과적으로 포착할 수 있는 통계적 모델로, 김포-제주 노선의 계절적 수요 변동을 예측하는 데 유용하다. Prophet 모델은 Facebook에서 개발한 시계열 예측 모델로, 강력한 추세 및 계절성 분석 기능을 갖추고 있으며, 공휴일 등의 특별한 이벤트를 고려한 예측이 가능하다. CatBoost 모델은 Yandex에서 개발한 기계 학습 알고리즘으로, 범주형 변수를 효율적으로 처리하며, 높은 예측 성능을 자랑한다. 마지막으로, Random Forest 모델은 다수의 결정 트리를 사용한 앙상블 학습 방법으로, 데이터의 복잡한 패턴을 잘 학습하여 높은 예측력을 제공한다. 이 모델들을 통해 본 연구는 김포-제주 노선의 승객 수요 예측 정확도를 평가하고, 본 연구를 통해 얻어진 예측 결과는 UAM의 효율적인 계획과 운영에 필요한 기초 정보를 제공할 것이다.

Ⅱ. 수요 예측 이론적 배경

2-1 승객 수요 예측

승객 수요 예측은 항공사 운영의 핵심 요소 중 하나로, 운항 스케줄 최적화, 자원 배분, 비용 절감, 서비스 품질 향상 등에 직접적인 영향을 미친다. 정확한 수요 예측을 통해 항공사는 노선 계획을 최적화하고, 항공기 및 승무원 배치를 효율적으로 수행할 수 있다. 또한, 정확한 수요 예측은 예약 시스템의 과부하를 방지하고, 승객에게 안정적인 서비스를 제공하는 데 기여한다. 특히, 도심 항공 모빌리티 시대를 대비하여, 단거리 비행 및 빈번한 운항이 예상되는 상황에서는 승객 수요 예측의 중요성이 더욱 부각된다. UAM 운영의 특성상, 수요 예측의 정확성은 효율적인 운항 계획과 안전한 비행 운영을 보장하는 데 필수적이다.

2-2 시계열 데이터

시계열 데이터란 일정한 시간 동안 수집된 일련의 순차적으로 정해진 데이터 셋의 집합이다 [4]. 시계열 데이터의 기본적인 특징은 특정 시점의 값이 다른 시점의 값과 특정한 방식으로의 연관을 가지고 있다는 것이다. 즉, 각 값이 서로 완전히 독립적이지 않다는 것이다. 시계열 데이터는 이러한 특징을 가지고 있는 데이터이며, 시계열 데이터 분석은 이런 서로 연관되어 있는 값들의 관계를 파악하는 것이라고 할 수 있다.

시계열을 구성하는 4가지 요인은 아래 표 1과 같다.

표 1. 시계열 데이터 구성 요소
Table 1. Time series data component.

Component	Description
Trend Factor	Long-term fluctuation factors that affect time series data due to changes in population, inflation or deflation, etc.
Cycle Factor	Mid-term fluctuation factors, which typically cycle with a cycle of 2 to 10 years, are grouped and analyzed as trend factors when the data observation period is not long.
Seasonal Factor	Fluctuation factors that occur on a one-year cycle, corresponding to relatively short-term fluctuations compared to trends or cycles
Irregular Factor	This refers to error variation that is difficult to measure and predict, that is, remaining variation that cannot be explained by the three factors above.

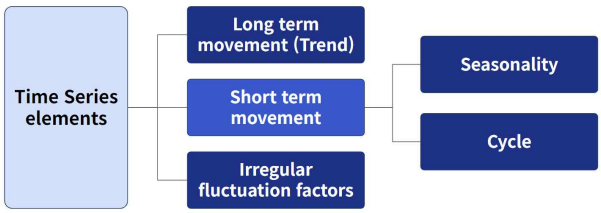


그림 1. 시계열 데이터 구성요소 간의 관계
Fig. 1. Relationships between time series data components.

시계열 분석이라고 하는 것은 이런 시계열을 해석하고 이해하는 데 쓰이는 여러 가지 방법을 연구하는 분야이다. 이 분야는 크게 규칙적 시계열 분석과 불규칙적 시계열 분석으로 나뉜다. 규칙적 시계열이란 트렌드와 분산이 불변하는 시계열 데이터를 말하고, 불규칙적 시계열이란 트렌드 혹은 분산이 변화하는 시계열 데이터를 말한다.

2-3 SARIMA 모델

SARIMA모델은 시계열 데이터의 계절적 패턴을 고려한 통계적 예측 모델이다. ARIMA(autoregressive integrated moving average) 모델에 계절적 차분과 계절적 자기 회귀 및 이동 평균 항을 추가하여 계절성을 반영한 것이 특징이다 [4]. SARIMA 모델은 데이터의 트렌드와 계절성을 분리하여 분석하며, 계절적 주기를 가진 시계열 데이터 예측에 강점을 가진다. 모델 구축 과정은 먼저 데이터의 안정성을 확인한 후, 차분(differencing)을 통해 비정상성을 제거한다. 이후, 자기 회귀 항과 이동 평균 항을 설정하여 모델을 피팅(fitting)하고, 적합도 검정을 통해 최적의 파라미터를 도출한다. SARIMA 모델은 항공 수요와 같이 계절적 패턴이 뚜렷한 데이터에 효과적으로 적용될 수 있다.

2-4 Prophet 모델

Prophet 모델은 Facebook에서 개발한 시계열 예측 모델로, 강력한 추세 및 계절성 분석 기능을 갖추고 있다. Prophet은 데이터의 불규칙한 트렌드와 계절성을 효과적으로 캡처하며, 공휴일 등의 특별 이벤트를 포함한 예측이 가능하다. Prophet 모델은 자동화된 설정과 직관적인 파라미터 조정을 통해 사용자가 쉽게 예측 모델을 구축할 수 있도록 설계되었다 [5]. 특히, 데이터의 누락이나 비정상적인 변동에 강인한 성능을 보이며, 주기적인 패턴을 효과적으로 분석한다. 모델의 기본 구조는 트렌드(trend)와 계절성(seasonality), 그리고 휴일 효과(holiday effect)로 구성되며, 각 구성 요소는 독립적으로 조정이 가능하다. Prophet은 복잡한 시계열 데이터를 신속하게 처리하고, 예측 결과를 시각적으로 확인할 수 있는 장점을 제공한다.

2-5 CatBoost 모델

CatBoost 모델은 Yandex에서 개발한 기계 학습 알고리즘으로, 특히 범주형 변수를 효율적으로 처리하는 데 강점을 가진다. CatBoost는 Gradient Boosting 알고리즘의 일종으로, 예측 성능이 우수하고 과적합(overfitting)을 방지하는 다양한 기술을 적용하고 있다 [6]. 특히, CatBoost는 데이터의 순서를 고려한 순차적 학습 방식을 채택하여, 범주형 변수의 처리 과정에서 발생하는 통계적 편향을 줄인다. 또한, 고도의 병렬 처리가 가능하여 대규모 데이터셋에서도 빠른 학습 속도를 보인다. CatBoost 모델은 항공 수요 예측과 같은 복잡한 데이터 구조를 가진 문제에서도 높은 성능을 발휘하며, 다양한 하이퍼파라미터 튜닝을 통해 최적의 모델을 구축할 수 있다. 특히, 변수 중요도(feature importance)를 명확히 제시하여 예측 결과의 해석을 용이하게 한다.

2-6 RandomForest 모델

RandomForest 모델은 다수의 결정 트리(decision tree)를 결합한 앙상블 학습 방법으로, 데이터의 복잡한 패턴을 잘 학습하여 높은 예측력을 제공한다 [7]. RandomForest는 각각의 결정 트리가 데이터의 무작위 샘플과 변수의 무작위 서브셋을 사용하여 독립적으로 학습한 후, 최종 예측을 위해 이들의 결과를 평균내거나 다수결 투표 방식으로 결합한다. 이 과정은 모델의 안정성과 일반화 능력을 향상시키며, 과적합을 방지하는 데 효과적이다. RandomForest는 다변수 예측 문제에서 특히 유용하며, 변수 중요도를 평가하는 기능을 통해 예측 모델의 해석을 돕는다. 또한, 모델의 복잡도를 조정하는 하이퍼파라미터가 비교적 단순하여 사용자가 쉽게 적용할 수 있다. 항공 수요 예측에서 RandomForest는 다양한 승객 수요 패턴을 학습하고, 예측의 정확성을 높이는 데 유용한 도구로 활용될 수 있다.

Ⅲ. 연구 방법론

3-1 연구 데이터 수집

본 연구에서는 대한항공의 실제 운항 데이터 중 김포발 제주행 노선의 데이터를 활용하였다. 해당 데이터는 2019년 1월 1일부터 2023년 12월 31일까지의 기간 동안의 운항 정보를 포함하며, 총 약 55,000건의 데이터가 확보되었다. 데이터에는 각 비행편의 출발일, 도착일, 출발 시간, 도착 시간, 항공기 기종, 탑승 인원 등의 세부 정보가 포함되어 있다. 이러한 대규모 데이터셋은 승객 수요 예측 모델 구축을 위한 풍부한 정보를 제공하며, 다양한 시계열 패턴을 분석하는 데 유용하다.

3-2 데이터 전처리

데이터 전처리는 정확한 예측 모델을 구축하기 위한 필수적인 과정으로, 다음과 같은 단계로 진행되었다. 먼저, 원시 데이터에서 필요한 변수들을 선택하고, 결측치 및 이상치를 처리하였다. 특히, 탑승 인원 데이터의 결측치는 인접 날짜의 데이터를 활용하여 보완하거나, 비정상적인 값은 제거하였다. 다음으로, 각 비행편의 정보를 일일 단위로 집계하여, 일별 총 승객 수를 산출하였다. 이를 통해 일일 승객 수 데이터가 생성되었으며, 그림 2. 와 같이 시계열 데이터로서의 형태를 갖추게 되었다. 마지막으로 시계열 데이터의 정상성을 가정하는 SARIMA, Prophet 모델을 사용하기 위해 차분(differencing)과 로그 변환(log transformation)을 수행하였다. 차분은 데이터의 비정상성을 제거하고, 트렌드와 계절성을 분리하여 안정적인 시계열을 형성하는 데 사용되었다. 로그 변환은 데이터의 분포를 정규화하고, 변동성을 줄여 예측 모델의 성능을 향상시키는 역할을 하였다. 데이터의 정상성이란 시간의 흐름에 따라 일정한 분산과 일정한 평균을 가지고 있는 것을 말한다. 만약 데이터가 정상성을 가지고 있지 않다면, 현재의 패턴이 미래에 똑같이 재현되지 않기 때문에, 예측 기법을 적용하더라도 유의미한 결과를 얻을 수 없다. 데이터의 정상성을 확인하기 위해 ACF(auto correlation function), PACF(partial auto correlation function) 도표를 활용하였다. ACF는 현재의 값이 과거의 값과 어떤 관계를 갖고 있는지 모두 보여주는 함수이며, PACF는 과거의 값이 현재의 값에 준 영향만을 고려하여 상관관계를 보여주는 함수이다. 정상성을 가진 데이터라면 correlation이 빠르게 내려가고 0으로 수렴하는 것을 볼 수 있다. 그림 3. 과 같이 Raw Data의 ACF 도표를 보면 correlation이 내려가긴 하지만 느리게 감소하여 신뢰구간에는 전혀 들어가지 못하므로 정상성을 가진 데이터가 아닌 것으로 판단된다. 그림 4. 와 같이 차분, 로그 변환을 수행한 Data는 ACF 도표를 보면 correlation이 빠르게 내려가면서 신뢰구간에 들어가는 것으로 보아 정상성을 가진 데이터라고 판단된다.

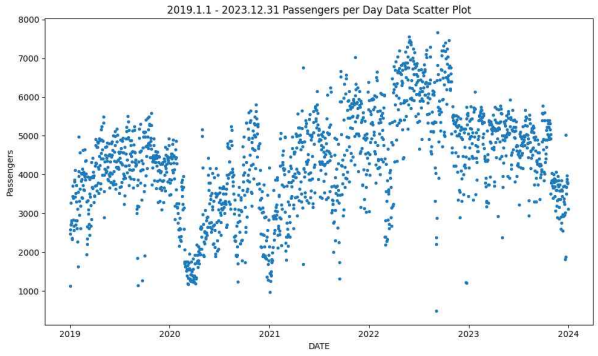


그림 2. 전처리 수행 후 일일 승객 수 데이터
Fig. 2. Daily passenger count data after performing raw data preprocessing.

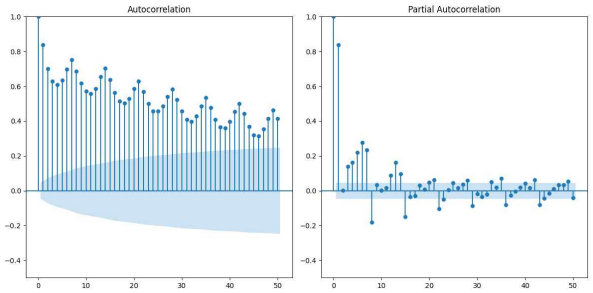


그림 3. Raw Data의 ACF/PACF 그래프
Fig. 3. ACF/PACF graph of raw data.

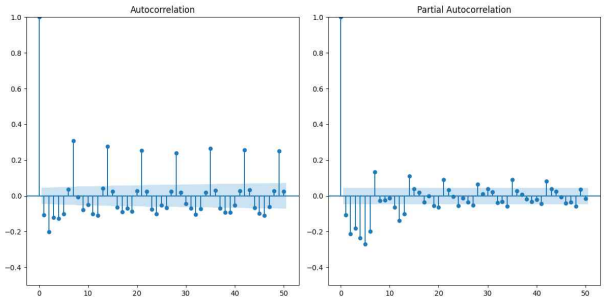


그림 4. Raw Data에 차분, 로그 변환 후 ACF/PACF 그래프
Fig. 4. ACF/PACF graph after differential and log transformation on raw data.

3-3 모델 구축

본 연구에서는 다양한 예측 모델을 구축하여 성능을 비교 분석하였다. 선택된 모델은 SARIMA, Prophet, CatBoost, RandomForest로, 각 모델을 구축한 내용은 다음과 같다.

SARIMA 모델은 계절적 패턴을 고려한 시계열 예측 모델로, 주말이나 계절과 같은 승객 수요 변화에 대한 데이터의 계절성을 효과적으로 반영할 수 있다. SARIMA 모델 구축 과정은 데이터의 자기상관성을 분석하여 적절한 모델을 선택하였다. 그 다음 최적의 자기회귀 항(AR : auto regressive model),

이동 평균 항(MA : moving average model), 그리고 계절적 항을 설정하였고, 설정된 모수를 기준으로 모델을 피팅하여 예측을 수행하도록 구성하였다.

Prophet 모델은 Facebook에서 개발한 시계열 예측 모델로, 추세와 계절성을 자동으로 감지하고 조정할 수 있는 장점을 가진다. Prophet 모델은 비정상적인 변동과 공휴일 등의 이벤트를 포함한 예측이 가능한 것이 특징이며, 본 연구에서는 공휴일을 포함한 예측을 수행하도록 구성하였다.

CatBoost 모델은 기계 학습 알고리즘으로, 범주형 변수를 효율적으로 처리하며, 높은 예측 성능을 자랑한다. 운항데이터는 일일 데이터로 월요일부터 일요일까지 각 요일마다 하나의 범주 데이터로 인식하여 예측을 수행하도록 모델을 구성하였다.

RandomForest 모델은 다수의 결정 트리를 결합한 앙상블 학습 방법으로, 데이터의 복잡한 패턴을 잘 학습하여 높은 예측력을 제공한다. RandomForest 모델은 변수 중요도를 평가하는 기능을 통해 예측 결과의 해석을 용이하게 하며, 다양한 하이퍼파라미터 튜닝을 통해 최적의 모델을 구축할 수 있다. 본 연구에서는 RandomForest 모델의 결정 트리 개수를 기본 값인 100개로 설정하였고, 랜덤한 값을 설정하는 계수를 기본값인 42로 설정하여 재현 가능한 결과를 얻도록 설정하였다.

3-4 성능 측정 지표

모델의 예측 성능을 평가하기 위해 다양한 성능 측정 지표를 사용하였다. 주요 성능 지표는 다음과 같다.

평균 절대 오차(MAE, mean absolute error)는 예측 값과 실제 값의 절대 오차의 평균을 계산하여 모델의 정확도를 평가한다. MAE는 예측 오차의 크기를 직관적으로 이해할 수 있게 해주며, 값이 작을수록 예측 모델의 정확도가 높음을 의미한다. MAE는 다음과 같이 계산된다.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y_i}\right| \tag{1}$$

평균 제곱근 오차(RMSE, root mean squared error)는 예측 값과 실제 값의 차이를 제곱한 후 평균을 내고 이를 제곱근한 값이다. RMSE는 예측 오차의 크기를 강조하여 큰 오차가 있을 때 민감하게 반응하는 특성을 가지며, 큰 오차에 대해 더 큰 패널티를 부여한다. RMSE는 다음과 같이 계산된다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \tag{2}$$

평균 절대 백분율 오차(MAPE, mean absolute percentage error)는 예측 값과 실제 값의 절대 오차를 실제 값으로 나눈 후 백분율로 표현한 지표이다. MAPE는 예측 오차를 비율로 나타내어, 데이터 규모에 관계없이 모델 성능을 비교할 수 있는 장점을 가진다. MAPE는 다음과 같이 계산된다.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y_i}}{y_i}\right| \times 100 \tag{3}$$

결정 계수(R^2, coefficient of determination)는 예측 모델이 실제 데이터를 얼마나 잘 설명하는지를 나타내는 지표이다. R^2 값은 0에서 1 사이의 값을 가지며, 값이 1에 가까울수록 모델의 설명력이 높음을 의미한다. R^2는 다음과 같이 계산된다.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} \tag{4}$$

이러한 성능 지표들은 각각의 특성과 장점을 바탕으로 모델의 예측 정확도를 다각적으로 평가할 수 있게 해준다. MAE는 직관적으로 예측 오차의 크기를 이해할 수 있도록 도와주며, RMSE는 큰 오차에 민감하게 반응하여 모델의 예측 성능을 강조한다. MAPE는 예측 오차를 비율로 나타내어 다양한 규모의 데이터셋에 대한 비교를 용이하게 하며, R^2는 모델의 설명력을 정량적으로 평가할 수 있게 한다. 이 성능 지표들을 활용하여, 본 연구는 각 모델의 예측 정확도를 비교 분석하고, 최적의 승객 수요 예측 모델을 선정하였다.

Ⅳ. 결과 및 토의

4-1 모델 성능 평가

본 연구에서는 대한항공의 김포발 제주행 운항 데이터를 이용하여 다양한 예측 모델의 성능을 평가하였다. 모델 성능 평가 지표로는 앞에서 기술한 것과 같이 평균 절대 오차(MAE), 평균 제곱근 오차(RMSE), 평균 절대 백분율 오차(MAPE), 결정 계수(R^2)를 사용하였다. 이러한 지표를 통해 각 모델의 예측 정확도와 설명력을 평가하였다.

그 결과로, 그림 7.에서 볼 수 있듯이 SARIMA 모델은 MAE가 409.931, RMSE가 562.190, MAPE가 9.723, R^2가 0.504로 나타났다. 이는 SARIMA 모델이 중간 정도의 예측 오차를 가지고 있으며, 데이터의 변동성을 약 50% 정도 설명할 수 있음을 의미한다. Prophet 모델은 MAE가 481.782, RMSE가 623.575, MAPE가 11.349, R^2가 0.390으로, SARIMA 모델보다 다소 낮은 성능을 보였다. Random Forest 모델은 MAE가 136.133, RMSE가 188.385, MAPE가 3.223, R^2가 0.944로 매우 높은 예측 정확도를 보였다. 이는 Random Forest 모델이 데이터의 복잡한 패턴을 잘 포착하고, 매우 낮은 예측 오차와 높은 설명력을 가지고 있음을 보여준다. CatBoost 모델은 MAE가 225.228, RMSE가 310.840, MAPE가 5.349, R^2가 0.848로, Random Forest 모델보다 다소 낮은 성능을 보였지만 여전히 높은 예측 정확도를 나

타냈다.

그림 5. 에서는 2023년도의 실제 값과 4개 모델의 예측 값을 모두 도시하였고, 그림 6.에서 볼 수 있듯이 실제 값과 예측한 값의 차이를 절대값으로 표현한 내용을 보면 Random Forest 모델과 CatBoost 모델의 값이 현저하게 낮은 것을 볼 수 있다. 이러한 결과는 머신러닝 기반 모델이 전통적인 시계열 모델에 비해 훨씬 더 높은 성능을 보일 수 있음을 시사한다. 특히, Random Forest 모델은 모든 평가 지표에서 가장 우수한 성능을 보였으며, 이는 모델이 데이터의 비선형성과 복잡한 패턴을 효과적으로 처리할 수 있음을 의미한다.

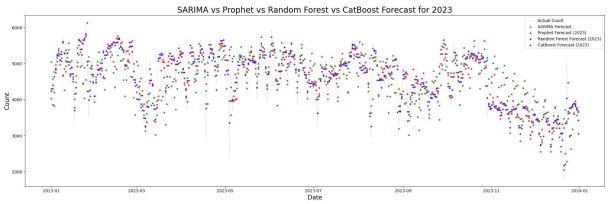


그림 5. 4개 모델의 2023년도 예측 결과

Fig. 5. 2023 passenger number forecast results from four models.

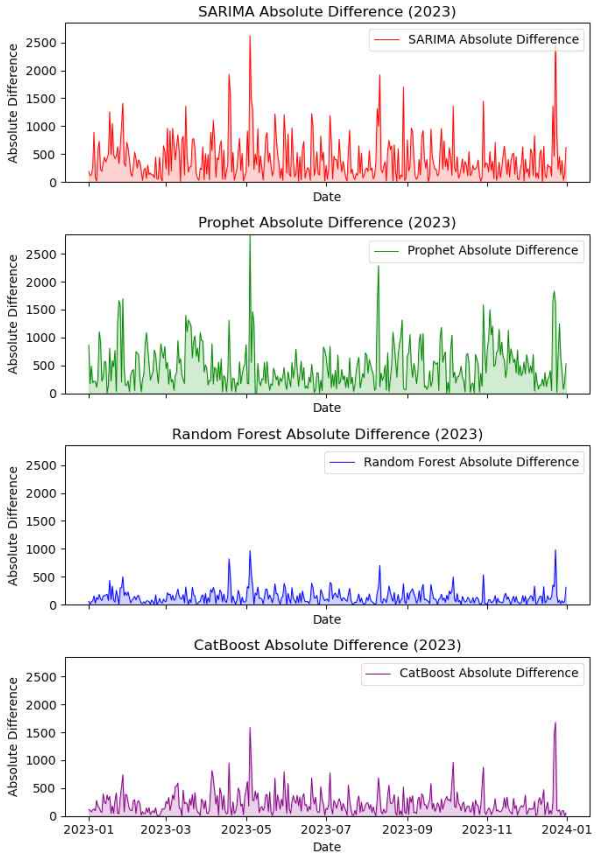


그림 6. 2023년도의 승객수 예측값과 실제값 차이의 절대값

Fig. 6. Absolute value of the difference between the predicted number of passengers in 2023 and the actual value.

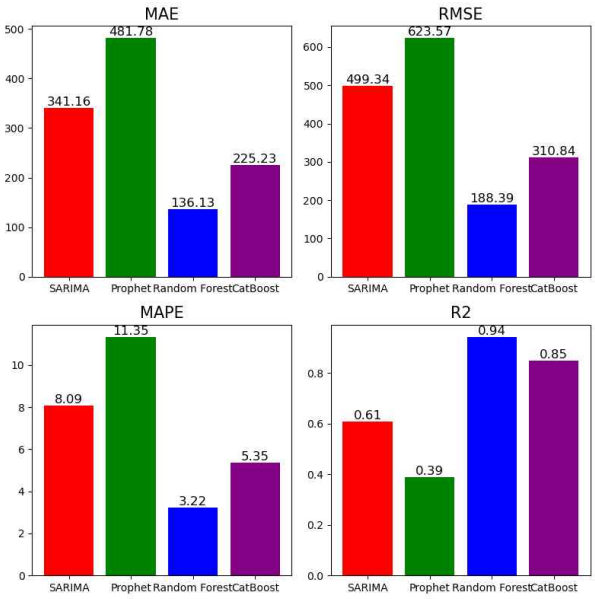


그림 7. 4개 모델의 MAR, RMSE, MAPE, R^2

Fig. 7. Comparison of MAR, RMSE, MAPE, R^2 of four models.

4-2 모델 간 비교 및 분석

모델 간 비교를 표현하는 그림 7.에서 볼 수 있듯이 Random Forest 모델이 모든 지표에서 가장 우수한 성능을 보였음을 알 수 있었다. Random Forest 모델은 MAE, RMSE, MAPE 지표에서 다른 모델에 비해 월등히 낮은 값을 보였으며, R^2 값이 0.944로 데이터 변동성을 거의 완벽하게 설명할 수 있음을 보여주었다. 이는 Random Forest가 여러 결정 트리를 결합하여 예측 성능을 향상시키는 앙상블 학습 방법을 사용하기 때문이다.

CatBoost 모델도 높은 성능을 보였지만, Random Forest 모델보다는 다소 낮은 예측 정확도를 보였다. CatBoost는 Gradient Boosting 알고리즘을 사용하여 데이터의 패턴을 학습하고, 카테고리형 데이터를 효과적으로 처리할 수 있는 강점을 가지고 있다. 그러나 일부 기간에서는 Random Forest보다 큰 오차를 보이는 경향이 있었다.

SARIMA와 Prophet 모델은 상대적으로 낮은 성능을 보였다. SARIMA 모델은 계절성과 트렌드를 잘 반영하지만, 데이터의 복잡한 변동성을 충분히 반영하지 못하는 한계가 있다. Prophet 모델은 Facebook에서 개발된 시계열 예측 모델로, 계절성, 휴일 효과 등을 반영할 수 있지만, 데이터의 비선형성과 복잡한 패턴을 포착하는 데 한계가 있다.

이러한 비교 분석을 통해 머신러닝 기반 모델이 전통적인 시계열 모델에 비해 더 높은 예측 성능을 보인다는 것을 확인할 수 있었다. 이는 머신러닝 모델이 데이터의 복잡한 패턴과 비선형성을 더 잘 포착할 수 있기 때문이다.

4-3 예측 결과 해석

모델의 예측 결과를 통해 분석한 바에 따르면, Random Forest 모델은 실제 데이터와 매우 근접한 예측값을 제공하였다. 이는 모델이 일일 단위의 변동성을 잘 포착하고, 계절적 패턴과 추세를 정확히 반영한 결과이다. 특히, 예측 오차가 낮아 실제 운항 데이터와 매우 일치하는 결과를 보였다.

CatBoost 모델도 유사한 경향을 보였으나, 일부 기간에서는 다소 큰 오차를 보였다. 이는 모델이 특정 패턴이나 외부 요인을 완벽히 반영하지 못한 결과로 해석될 수 있다. 예를 들어, 특정 기간 동안의 예외적인 이벤트나 급격한 수요 변동에 대해 충분히 학습하지 못했을 가능성이 있다.

SARIMA와 Prophet 모델은 일정한 패턴을 따르지만, 데이터의 복잡한 변동성을 충분히 반영하지 못하는 경향을 보였다. 이는 예외적인 이벤트나 비정상적인 변동성에 대한 반응이 부족한 모델의 한계로 볼 수 있다. SARIMA 모델은 계절성과 추세를 반영하지만, 데이터의 비선형성과 복잡한 패턴을 반영하는 데 한계가 있다. Prophet 모델은 휴일 효과 등을 반영할 수 있지만, 데이터의 비선형성을 포착하는 데 한계가 있다.

이러한 결과는 머신러닝 모델이 데이터의 복잡한 패턴과 비선형성을 더 잘 포착할 수 있음을 보여준다. 특히, Random Forest 모델은 높은 유연성과 데이터의 다양한 특징을 효과적으로 반영할 수 있는 능력 덕분에 뛰어난 예측 성능을 보였다.

4-4 한계와 개선 방향

본 연구는 대한항공의 특정 노선 데이터만을 사용하였기 때문에, 다른 노선이나 항공사에 일반화하는 데 한계가 있다. 이는 연구 결과를 다양한 항공사와 노선에 적용할 수 없음을 의미한다.

그리고 외부 변수 부족으로 인해 예측 모델의 성능이 제한될 수 있다. 날씨, 경제 지표, 사회적 이벤트 등의 외부 변수를 포함하지 않았기 때문에, 모델이 실제 상황을 충분히 반영하지 못할 수 있다. 특히, UAM의 경우 기상 조건과 도심 내 교통 상황이 중요한 변수로 작용할 것으로, 외부 변수를 반영하여 수요예측을 수행하는 것이 필요하다.

그리고 머신러닝 모델을 사용할 경우 높은 성능을 보이지만, 모델의 해석 가능성이 낮다는 한계가 있다. 이는 모델의 예측 결과를 이해하고 설명하는 데 어려움을 초래할 수 있다. 이러한 한계는 실무적 적용에서 중요한 문제로 작용할 수 있다.

향후 연구에서는 다음과 같은 개선 방향을 고려할 수 있다.

외부 변수를 추가하여 모델의 예측 성능을 향상시키는 것을 고려해볼 수 있다. 예를 들어, 날씨, 경제 지표, 사회적 이벤트 등 외부 변수를 모델에 포함한다면 보다 정밀한 예측을 할 수 있을 것으로 보인다.

또한 머신러닝 및 딥러닝 모델의 해석 가능성을 높이기 위한 연구가 필요하다. 예로 SHAP(shapley additive explanations) 값을 활용한 변수 중요도 분석 등을 통해 모델의 예측 결과를 명확히 이해하고, 이를 기반으로 한 의사결정의 신뢰성을 높일 수 있다. 해석 가능한 AI(explainable artificial intelligence) 기법을 도입하는 것도 고려해볼 수 있는데, 이를 통해 모델의 투명성을 향상시키고, 사용자가 예측 결과를 보다 쉽게 이해할 수 있도록 할 것이다.

Ⅴ. 결 론

본 연구에서는 대한항공의 김포발 제주행 운항 데이터를 바탕으로 SARIMA, Prophet, Random Forest, CatBoost 모델을 사용하여 승객 수요 예측을 수행하였다. 연구 기간은 2019년 1월 1일부터 2023년 12월 31일까지로, 일일 단위의 승객 수 데이터를 분석하였다. 각 모델의 성능을 MAE, RMSE, MAPE, R^2 지표로 평가하고 비교 분석하였다.

연구 결과, Random Forest 모델이 모든 성능 지표에서 가장 우수한 성능을 보였으며, CatBoost 모델이 그 뒤를 이었다. SARIMA와 Prophet 모델은 상대적으로 낮은 성능을 보였다. 이는 머신러닝 기반 모델이 데이터의 복잡한 패턴과 비선형성을 더 잘 포착할 수 있음을 시사한다. 특히, Random Forest 모델은 높은 유연성과 데이터의 다양한 특징을 효과적으로 반영할 수 있는 능력 덕분에 뛰어난 예측 성능을 보였다.

이러한 결과는 도심 항공 모빌리티(UAM) 준비를 위한 승객 수요 예측에 있어서 머신러닝 기반 모델이 유용할 수 있음을 보여준다. 향후 UAM의 운영 계획 수립에 있어 이들 모델을 활용함으로써 보다 정확한 수요 예측을 할 수 있을 것으로 기대된다. 이번 연구 결과는 항공사 운영과 계획 수립에 중요한 시사점을 제공한다. 높은 예측 정확도를 가진 모델을 활용함으로써 항공사는 효율적인 운항 계획을 수립하고, 승객 수요 변동에 신속하게 대응할 수 있다. 예를 들어, 성수기와 비수기 기간 동안의 승객 수요를 정확히 예측하여 항공편 스케줄을 최적화할 수 있다. 또한, 예측된 수요를 기반으로 항공기 배치와 승무원 스케줄링을 최적화하여 운영 효율성을 높일 수 있다.

특히, 머신러닝 모델의 활용을 통해 예측 성능을 더욱 향상시킬 수 있음을 보여준다. 이는 항공사 뿐만 아니라 도심 항공 모빌리티(UAM)와 같은 새로운 교통 수단의 도입에도 중요한 역할을 할 수 있다. UAM 운영자들은 이러한 모델을 활용하여 승객 수요를 정확히 예측하고, 효과적인 운항 계획을 수립할 수 있을 것이다.

Acknowledgments

본 연구는 국토교통부/국토교통과학기술진흥원의 지원으로 수행되었음 (과제번호 RS-2022-00143625).

References

[1] United Nations, World Urbanization Prospects (The 2018 Revision), Department of Economic and Social Affairs, New York: United Nations., ST/ESA/SER.A/420, pp. 21, 2019.

[2] J. -W. Lim(2024, March), UAM(urban air mobility), leading the 3D future transportation system. Available: https://www.samsungsds.com/kr/insights/uam.240313.html

[3] Y. -R. Kim, J. H. Lim and Y. C. Choi, "A study on the difference of selection attributes and importance of FSC and LCC by age group," *The Korean Society for Aviation and Aeronautics*, Vol. 25, No. 4, pp. 91-100, Dec. 2017. DOI: https://doi.org/10.12985/ksaa.2017.25.4.091

[4] D. -C. Han, D. -W. Lee, and D. -Y. Jung, "A study on the traffic volume correction and prediction using SARIMA algorithm," *The Journal of The Korea Institute of Intelligent Transport Systems*, Vol. 20, No. 6. pp. 1-13, Dec. 2021. DOI: https://doi.org/10.12815/kits.2021.20.6.1

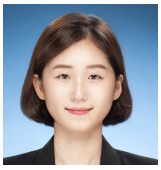
[5] S. J. Taylor, and Benjamin Letham, Forecasting at scale, Sept. 2017. Available: [Internet] https://facebook.github.io/prophet

[6] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: gradient boosting with categorical features support," *The Journal of The Korea Institute of Intelligent Transport Systems*, Vol. 20, No. 6. pp. 1-13, Dec. 2021. DOI: https://doi.org/10.48550/arXiv.1810.11363

[7] H. -J. Han, "Random forest for stationary time series : The case of forecasting inflation in korea," *The Journal of Korean Economic Association*, Vol. 71, No. 3. pp. 37-73, August. 2023. DOI: https://doi.org/ 10.22841/kjes.2023.71.3.002

김 정 훈 (Junghoon Kim)

2013년 8월 : 서울시립대학교 컴퓨터과학과 (공학사)
2013년 8월 ~ 2020년 4월 : ㈜ 대한항공 항공기술연구원 통합시험환경
2020년 4월 ~ 현재 : ㈜ 대한항공 항공기술연구원 M&S
※관심분야 : 도심항공교통, 운항통제 시스템, IT 개발, 체계 종합



최 선 미 (Seonmi Choi)

2017년 2월 : 한국항공대학교 항공우주공학과 (공학사)
2019년 9월 ~ 2021년 7월 : (주) 대한항공 정비본부
2021년 7월 ~ 현재 : (주) 대한항공 항공기술연구원 M&S
※ 관심분야 : 도심항공교통, 운항통제 시스템, 시스템 통합 및 인터페이스

조 희 덕 (Heeduk Cho)

1995 ~ 대한항공 입사, 운항관리 업무, 2000-2002 부산공항 운항관리 지원센터
2003~2015 통제센터, 항공기 스케줄, 운항통제 운영시스템 관리, 2015-2019 일본지역 통제운항관리 지원센터
2022 ~ UAM 부분 T/F
* 관심분야 : UAM 운항통제, 민항기 운항통제, 관련시스템 관리 운영