Journal of KIIT. Vol. 23, No. 7, pp. 13-21, Jul. 31, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.7.13

머신러닝과 SHAP 분석 기반 항공기 지연 예측 모델

김다연*[1], 송민경*[2], 이석종*[3], 장유진*[4], 이한준**

Flight Delay Prediction Model based on Machine Learning and SHAP Analysis

Dayeon Kim*[1], Minkyeong Song*[2], Seokjong Lee*[3], Youjin Jang*[4], and Hanjun Lee**

요 약

기상으로 인한 항공편 지연은 항공 스케줄에 연쇄적인 영향을 미치며 경제적 손실을 유발한다. 이에 여러 관련 선행연구가 이루어졌으나 주로 전통적인 통계기법을 활용한 연구가 주류를 이루며, 아직까지 국내 공항을 대상으로 이루어진 연구는 부족하다. 이에 본 연구에서는 인천국제공항의 저비용 항공편과 기상 데이터를 활용하여 머신러닝 기반의 항공기 지연 예측 모델을 제안한다. 이를 위하여 Random Forest, XGBoost, LightGBM 알고리즘을 활용한 모델을 구축하여 성능을 비교 분석하였으며 SHAP 분석을 통해 모델의 일반화 성능과 해석 가능성을 높였다. 실험 결과, Random Forest 모델이 가장 우수한 성능을 보였으며, 기온, 이슬점 온도, 해면 기압 등이 주요 변수로 도출되었다.

Abstract

Weather-induced flight delays have a cascading impact on flight schedules and result in economic losses. While several related studies have been conducted, most have primarily relied on traditional statistical methods, and research specifically targeting domestic airports remains insufficient. Therefore, this study proposes a machine learning-based flight delay prediction model utilizing low-cost carrier flight and weather data from Incheon International Airport. To achieve this, models were developed using Random Forest, XGBoost, and LightGBM algorithms, and their performance was compared and analyzed. Additionally, SHAP analysis was conducted to enhance the model's generalization performance and interpretability. Experimental results demonstrated that the Random Forest model achieved the highest performance, with temperature, dew point, and sea-level pressure identified as key variables.

Keywords

machine learning, flight delay, random forest, SHAP, Incheon international airport

* 명지대학교 경영정보학과 학부과정
- ORCID[1]: https://orcid.org/0009-0008-4630-767X
- ORCID[2]: https://orcid.org/0009-0007-5312-1833
- ORCID[3]: https://orcid.org/0009-0009-6964-8036
- ORCID[4]: https://orcid.org/0009-0004-1362-2316
** 명지대학교 경영정보학과 부교수(교신저자)
- ORCID: https://orcid.org/0000-0002-9005-3661

· Received: Apr. 04, 2025, Revised: Apr. 24, 2025, Accepted: Apr. 27, 2025
· Corresponding authors: Hanjun Lee
Dept of Management Information Systems, Myongji University, Korea
Tel.: +82-2-300-0772, Email: hjlee1609@gmail.com

Ⅰ. 서 론

항공기 지연은 항공사의 운영 비용 증가, 공항 혼잡도 상승, 탄소 배출 증가 등의 문제를 초래할 뿐만 아니라, 공항 이용자들의 불편과 피해를 증가시킨다[1]. 활주로 지체는 항공기 지연으로 인한 항공사, 승객 및 화주에게 부담된 시간비용이며, 연간 약 313억 원으로 나타난다[2]. 특히 인천국제공항과 같은 대형 허브공항의 경우, 단일 항공편의 지연이 연쇄적으로 확산되어 전체 항공 네트워크의 운영에 영향을 미칠 수 있다[3]. 항공기 지연의 원인은 다양하지만, 그중 기상 요인은 예측이 어렵고 통제가 불가능하다는 특성으로 인해 운영 측면에서 부담으로 작용한다. 미국 교통통계국(BTS, Bureau of Transportation Statistics)에 따르면, 2023년 기준 전체 항공기 지연 시간 중 기상 요인이 차지하는 비율은 약 26.8%로 보고되었다[4]. 이러한 점에서 기상 요인에 의한 지연 발생 가능성을 예측하고 신속히 대응할 수 있는 체계를 마련하는 것이 항공 산업의 운영 효율성과 서비스 품질 향상을 위한 핵심 과제로 부상하고 있다.

기존에는 통계적 분석을 기반으로 지연을 예측하는 연구가 주를 이루었으나, 머신러닝 기술의 지속적인 발전으로 인해 고도화된 데이터 기반 예측 모델이 점차 활용되고 있다. 머신러닝은 대량의 데이터를 분석하여 복잡한 패턴을 학습하는 데 뛰어난 성능을 보이지만, 모델의 의사결정 과정이 블랙박스 형태로 작동하여 해석 가능성이 낮다는 한계를 갖는다. SHAP(Shapley Additive Explanations) 분석은 이에 대한 대안으로 활용되고 있다. SHAP은 협력게임 이론(Cooperative game theory)에서 유래한 Shapley Value를 기반으로 하며, 머신러닝 모델의 결과를 분석할 때 개별 특징(Feature)이 기여하는 정도를 정량적으로 평가하는 기법이다[5]. 머신러닝을 기반으로 항공기 지연 예측 모델을 개발하고 SHAP 분석을 적용한다면 모델의 해석이 가능할 것이다.

이에 본 연구에서는 다양한 머신러닝 알고리즘을 활용하여 항공기 지연 예측 모델을 개발하고 SHAP 분석을 통하여 모델에 사용된 변수들의 영향력을 파악함으로써 항공기 지연에 영향을 미치는 주요 요인을 식별하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 선행연구를 검토하고, 3장에서는 연구의 접근 방법과 데이터 활용 방식을 설명한다. 4장에서는 연구 결과를 분석하고 논의하며, 마지막으로 5장에서 연구의 결론과 시사점을 제시한다.

Ⅱ. 선행연구

2.1 통계분석 기반 항공기 지연 예측 연구

선행연구는 크게 전통적인 통계 분석을 활용한 연구들과 머신러닝을 활용한 연구들로 나눌 수 있다. 국내 주요 공항 중 김포공항, 김해공항, 제주공항을 대상으로 수행한 선행연구[6]에서는 기상 요인과 항공기 결항 및 지연과의 관련성을 분석한 바 있다. 해당 연구에서는 회귀 분석과 확률 모델을 기반으로 김포공항은 안개가, 제주공항은 강풍 및 안개가 항공기 지연의 주요 원인임을 결과로 제시하였다[6].

또 다른 선행연구[7]에서는 김포 제주 노선 항공기 운항을 중심으로, 극한 기상현상이 항공기 지연 시간과 결항 확률에 미치는 영향을 분석하였다. 해당 연구에서는 패널 고정효과 모형으로 지연 시간을 추정하고, 프로빗 모형의 한계효과 분석을 통하여 결항에 대한 확률을 추정하여 제시하였다[7].

이상에서와 같은 통계 기반 연구들은 기상 및 과거 지연 데이터를 활용하여 변수 간의 관계를 분석하는 데 강점이 있다. 그러나 정규성이나 독립성과 같은 가정에 기반하기 때문에 다양한 요인의 비선형적 상호작용을 반영하기 어렵다는 점에서, 다변량 요인의 영향을 받는 항공기 지연 예측에는 한계가 있다. 이러한 한계를 보완하기 위해, 최근에는 복잡한 변수 간 상호작용을 반영하고 더 높은 예측 성능을 확보하기 위해 머신러닝 기법이 도입되고 있다. 머신러닝 기법은 이러한 복잡한 관계를 추론하고 유의미한 패턴을 학습하는 데 효과적이므로, 데이터 특성과 분석 목적에 따라 유용하게 활용될 수 있다.

2.2 머신러닝 기반 항공기 지연 예측 연구

미국 뉴욕의 3개 공항 데이터를 활용한 연구[8]

에서는 분류 및 회귀 모델을 개발하고, 딥러닝, 로지스틱 회귀, 결정 트리, Random Forest, XGBoost (eXtreme Gradient Boosting) 등을 적용하여 성능을 분석하였다. 연구 결과, 강수량, 풍속, 가시성이 지연 여부를 결정하는 핵심 변수로 도출되었으며, 온도, 풍속, 강수량은 지연 시간 예측에 중요한 영향을 미치는 것으로 나타났다.

제주국제공항의 항공기 지연 데이터를 분석한 연구[9]에서는 Random Forest 모델을 구축하여 기상 조건과 지연 패턴 간의 관계를 분석하였다. 연구 대상 422,218편 중 약 24.8%(104,593편)가 지연되었으며, 이 중 기상 요인으로 인한 지연은 30,350편(약 29%)으로 나타났다. 세부적으로 강수와 관련된 지연은 16,878편이었지만 주요 변수로 작용하지 않았으며, 15KT 이상의 강풍이 불 때 전체 항공편 55,488편 중 19,098편이 지연되어 중요한 요인으로 분석되었다. 또한, 순간돌풍(GUST)과 급변풍(WS_RWY) 발생 시 각각 38.2%와 35.3%의 높은 지연률을 보였다. 저시정 시 출발 항공편의 55.8%가 지연되었으나 전체 발생 빈도는 낮았고, 뇌전 발생 시 41.4%의 지연률을 보였으나 주요 변수로 작용하기에는 빈도가 제한적이었다. 연구 결과, 풍속(WSPD), 순간돌풍(GUST), 급변풍(WS_RWY)이 상관관계를 이루며 핵심 기상 변수로 작용함이 확인되었고, 단일 변수가 아닌 복합적인 기상 조건을 종합적으로 고려해야 함이 강조되었다.

BTS에서 제공하는 2015~2019년 시카고 오헤어 국제공항(ORD) 데이터[1]를 활용한 연구에서는 로지스틱 회귀와 Random Forest 모델을 적용하고, 언더샘플링 및 SMOTE(Synthetic Minority Over sampling TEchnique) 샘플링 기법을 활용하여 지연 예측 모델을 구축하였다. 분석 결과, 기상 변수를 포함한 Random Forest 모델의 정확도는 0.718로 나타났으며, 기상 정보가 제외될 경우 0.604로 감소하여 기상 요인이 지연 예측에 중요한 영향을 미친다는 점이 확인되었다. 변수별 중요도는 직접 수치화되지 않았으나, 계절별 분석을 통해 특정 시기에 따라 지연 패턴이 달라질 수 있음을 시사하였다.

또한, JFK(John F. Kennedy) 국제공항의 2015년 항공편 데이터와 OpenWeatherMap API를 활용한 연구[10]에서는 로지스틱 회귀, SVM(Support Vector Machine), Random Forest, 인공 신경망(Artificial neural network) 등의 모델을 비교 분석하였다. 연구 결과, 항공기 지연은 스케줄과 높은 연관성을 보이지만, 기상 조건에 따라 지연 확률이 달라지는 경향이 확인되었다. 특히, 뇌우 발생 시 지연 확률이 증가하였으며, 기압, 온도, 풍속, 상대 습도 등이 지연에 영향을 미치는 주요 변수로 도출되었다. 또한, 눈이 내릴 경우 습도와 풍속이 중요한 역할을 하였으며, 맑은 날씨에서도 풍속과 구름 비율(Cloud fraction)이 지연 발생에 영향을 미치는 것으로 나타났다. 이슬비(Light intensity drizzle) 발생 시 온도와 습도가 주요 변수로 작용하는 등, 누적된 기상 변화의 영향을 고려할 필요성이 제기되었다.

이처럼 머신러닝 기법을 적용한 연구들은 기존 통계적 분석이 비선형적 특성을 충분히 반영하지 못하는 한계를 극복하고, 변수 간의 복잡한 관계를 보다 정밀하게 분석할 수 있도록 하였다. 특히, 강수량, 풍속, 순간돌풍 등의 기상 변수가 항공기 지연에 미치는 영향을 효과적으로 식별하는 데 기여하였다. 그러나 기존 연구들은 뉴욕, 시카고 등 특정 해외 공항을 중심으로 이루어졌으며, 국내 공항을 대상으로 한 연구는 상대적으로 부족한 실정이다. 이에 본 연구에서는 국내에서 이용객 수가 가장 많은 인천국제공항을 대상으로, 기상 조건을 반영한 항공기 지연 예측 모델을 구축하고자 한다. 이를 통해 국내 공항의 지연 원인을 보다 정밀하게 분석하고, 맞춤형 예측 및 대응 전략 수립에 기여할 것으로 기대된다.

III. 연구 방법

3.1 데이터 수집

본 연구에서는 항공정보포털시스템[11]에서 제공하는 항공기 출도착 현황 데이터와 기상자료개방포털[12]에서 제공하는 공항 기상 관측 데이터를 활용하였다. 코로나19로 인한 영향이 분석 결과에 영향을 미치지 않도록, 본 연구에서는 2022년 9월부터 2024년 8월까지의 기간을 분석 대상으로 하였다.

항공기정보포털시스템은 국토교통부와 한국항공협회가 운영하는 항공 정보 플랫폼으로서 국내외

항공 소식, 항공기 출발 및 도착 안내 서비스 등을 제공하고 있다. 본 연구에서는 활용한 항공기 출도착 현황 데이터는 출도착, 날짜, 공항, 항공사, 편명을 기준으로 계획 시간, 예상 시간, 도착 시간, 여객기 구분, 지연 현황을 포함한다. 본 연구의 목적은 인천국제공항에서 출발하는 저비용 항공편의 지연 여부를 분류하는 머신러닝 모델을 개발하는 것에 있으므로, 인천국제공항의 출발 여객편에 대한 데이터로 한정하였다.

기상자료개방포털은 기상청이 운영하는 오픈 API (Application Programming Interface) 서비스로서 지상, 해양을 포함한 총 30종류의 날씨 데이터를 제공한다. 본 연구에서 활용한 공항기상관측 데이터는 인천국제공항을 대상으로 수집하였으며 풍향, 풍속, 시정, 운량, 운고, 기온, 이슬점 온도 등의 기상관측 자료를 포함하고 있다. 풍향 데이터는 0도부터 360도까지 각도 단위로 이루어져 있고, 강수량 데이터는 mm 단위로 이루어져 있다. 일기현상 데이터는 0부터 99까지 숫자로 표기되어 있으며 각 숫자가 의미하는 기상현상이 존재한다. 운고는 구름의 높이를 의미하는 것으로 ft 단위로 이루어져 있다. 운량은 구름의 양을 뜻하며, 0부터 7까지 숫자가 커질수록 구름의 양이 많아지는 것을 의미한다. 시정은 물체가 보이는 최대 거리인 가시거리를 뜻하고 이슬점 온도는 이슬이 맺히는 온도를 의미하며 안개와 지상의 결로현상과 밀접한 관련이 있다. 해면기압은 해발고도가 0m인 해수면 상의 기압을 말하며 단위는 hPa이다.

3.2 데이터 전처리 및 변수 추출

수집한 두 데이터셋을 분석에 적합하도록 단일 데이터셋으로 병합하였다. 두 데이터셋 병합 전, 변수 형식의 일관성을 확보하기 위해 데이터 형식을 정제하였다. 항공기 데이터셋의 경우 날짜와 시간이 별도의 컬럼으로 존재하므로, 기상 데이터셋의 일시 데이터를 날짜와 날씨측정시간으로 분리하였다. 해당 과정 이후 두 데이터셋을 날짜와 시간을 기준으로 통합하여 단일 데이터셋을 생성하였다.

다음으로, 단일 데이터셋에서 항공기 관련 주요 변수를 전처리하였다. 인천공항을 나타내는 지점명과 지점코드 등 중복된 항목을 제거하여 데이터 품질을 향상시켰다. 종속 변수인 현황은 지연 시간이 15분 미만인 경우 출발(0), 지연 시간이 30분 이상인 경우 지연(1)으로 이진 분류하였다. 국토교통부 항공정책실이 발표한 항공통계 작성 메뉴얼에 따라 지연 기준을 15분으로 재정의하였으나 모델이 15분을 기준으로 미세한 차이를 정확히 예측하지 못하는 문제를 고려하여, 15분 초과 30분 미만의 데이터는 제거하였다. 또, 현황 변수는 출발 또는 지연의 값만 가져야 하므로 결측치에 해당하는 행은 제거하였다. 항공사는 국내 저비용 항공사 중 점유율이 높은 에어부산, 제주항공, 티웨이항공, 진에어만 추출하였다.

다음으로, 기상 변수를 전처리하였다. 층별 운량은 하늘의 전체 구름의 양을 나타내는 전운량으로 대체하였다. 층별 운고의 경우 전운고 값을 나타내는 변수를 추가하였는데 각 층별 운고의 합으로 처리하였고, 결측치가 있는 행은 제거하였다. 일기현상 데이터 중 안개와 황사에 해당하는 코드만 선별하여 새로운 변수로 생성 후, 기존 일기 현상 데이터는 제거하였다. 풍향 데이터는 0부터 90까지 북동풍, 90부터 180까지 남동풍, 180부터 270까지 남서풍, 270부터 360까지 북서풍으로 범주화하여 처리하였다. 이후에 원핫인코딩(One-Hot encoding)을 통하여 NE, SE, SW, NW의 각 방향에 대한 이진 변수를 생성하였으며, 참일 경우 1, 거짓일 경우 0으로 나타낼 수 있게 변경하였다.

국내선과 국제선의 차이를 반영하기 위해, 도착공항 정보를 기반으로 국내선은 0, 국제선은 1로 처리하였다. 전체 데이터를 기준으로 국내선 21건, 국제선 43,485건으로 변수를 구성하였다.

휴일과 항공기 지연의 관계를 파악하고자 공휴일 및 주말 정보를 바탕으로 휴일 여부를 나타내는 이진 변수를 추가하였다. 인천공항에서 출발하는 항공편의 이용객 수가 많을수록 공항의 혼잡도가 증가할 가능성이 높아져 항공기 지연에 영향을 미칠 것으로 예상하였기 때문에, 해당 날짜를 기준으로 남은 휴일 수를 산출한 변수도 추가하였다.

최종 전처리 된 데이터는 43,506행 21열을 가지며, 지연되지 않고 출발한 항공편 11,557개와 지연된 항공편 31,949개로 구성된다. 본 연구에서 사용된 데이터의 예시는 그림 1과 같으며 전체 변수에 대한 기술통계량은 표 1과 같다.

표 1. 변수의 기술통계량
Table 1. Descriptive statistics of variables

Parameter	Mean	Std	Min	Max
DelayStatus	0.734	0.442	0	1
WindSpeed_KT	7.718	4.414	0	34
Visibility_m	8814.307	2301.883	50	10000
TotalCloudCover	3.904	3.118	0	8
Temp_C	14.152	11.15	-15.2	35.1
DewP_C	8.784	12.226	-25.1	27.3
SLP_hPa	1016.178	8.963	993.3	1039.7
LP_hPa	1015.344	8.929	992.6	1038.8
Precipitaion	0.188	1.631	0	61.9
CBH_m	9141.268	9031.458	0	49000
Fog	0.207	0.405	0	1
YellowDust	0.016	0.125	0	1
Dom/Int	0.999	0.022	0	1
IsHoliday	0.32	0.467	0	1
LeftHolidays	0.524	0.874	0	6
WindDirection_NE	0.192	0.394	0	1
WindDirection_SE	0.179	0.383	0	1
WindDirection_SW	0.229	0.42	0	1
WindDirection_NW	0.399	0.49	0	1

3.3 모델 구축

모델 구축에 앞서 훈련 데이터와 테스트 데이터를 7:3 비율로 분배하고, 모델 적합을 실시하였다. 데이터 불균형을 완화하고 과적합을 방지하기 위하여 언더샘플링 기법을 적용하였다.

다음으로 출발과 지연 데이터를 분류하는 모델을 만들기 위해 Random Forest, XGBoost, LightGBM (Light Gradient Boosting Machine)을 적용한 총 세 가지 종류의 분류 모델을 구축하였다.

이후, 모델의 성능 향상을 위해 하이퍼 파라미터 튜닝을 진행하였다. Random Forest 모델의 경우 최적의 트리 개수를 결정하기 위해 앙상블 기법을 사용하고, 정확도와 F1-Score 값을 기준으로 최적의 성능을 보이는 값을 선정하였다. 해당 과정을 통해 트리 개수(n_estimators)의 값은 200으로 채택하였다. LightGBM 모델은 이진 분류를 수행하는 것을 목적으로 설정하였고, 평가지표로 Area Under the Curve(AUC)를 활용하였다. 리프 노드 개수(num_leaves)는 31로 선정하였고, 트리의 최대 깊이(max_depth)는 -1로 선정하여 데이터에 따라 자동으로 결정될 수 있게 하였다. 리프 노드 최소 가중치의 합(min_child_weight)의 경우 과적합 방지를 위해 1로 설정하였다. 마지막으로 XGBoost 모델은 앙상블 학습을 통하여 트리 개수(n_estimators)는 100, 트리 최대 깊이(max_depth)는 5로 설정하였다. 평가지표(eval_metric)로 logloss를 활용하여 모델 예측 성능을 평가하였다.

이상의 연구 과정을 도식화하면 그림 2에서 보는 바와 같다.

	DelayStatus	WindSpeed_KT	Visibility_m	TotalCloudCover	Temp_C	DewP_C	SLP_hPa	LP_hPa	Precipitation_mm	CBH_m	Fog	YellowDust	Dom/Int
0	0	4.0	10000.0	0.0	20.8	18.9	1013.8	1013.0	0.0	0.0	0	0	1
1	0	9.0	10000.0	1.0	22.6	18.1	1014.6	1013.8	0.0	20000.0	0	0	1
2	0	8.0	10000.0	2.0	23.6	18.6	1015.1	1014.3	0.0	20000.0	0	0	1
3	0	8.0	10000.0	2.0	23.6	18.6	1015.1	1014.3	0.0	20000.0	0	0	1

그림 1. 데이터 예시
Fig. 1. Sample data

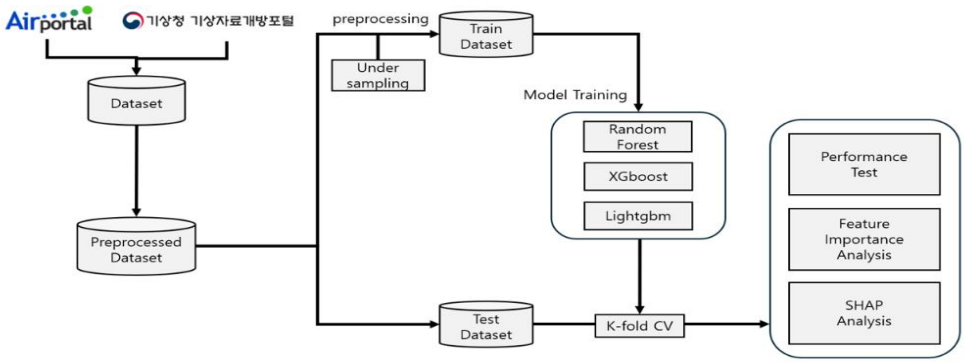


그림 2. 연구 흐름도
Fig. 2. Research flow

Ⅳ. 연구 결과

앞서 구축한 세 가지 머신러닝 모델의 성능을 정밀도(Precision), 재현율(Recall), F1 점수(F1 Score), 정확도(Accuracy)를 지표로 비교 분석하였다. 각 지표의 계산식은 다음과 같다. 여기서 TP(True Positive), TN(True Negative)은 각각 참 양성, 참 음성, FP(False Positive), FN(False Negative)은 각각 거짓 양성, 거짓 음성을 의미한다. 정밀도는 모델이 양성이라고 판단한 사례 중 실제로 양성인 비율을 나타내며, 재현율은 실제 양성 중에서 모델이 정확히 양성으로 예측한 비율이다. F1 점수는 정밀도와 재현율의 조화 평균으로, 두 지표의 균형을 고려한 성능 지표이다. 정확도는 전체 예측 중에서 올바르게 분류된 비율을 의미한다.

$$\mathrm{Precision} = \frac{TP}{TP+FP},\ Recall = \frac{TP}{TP+FN},$$

$$F1\,Score = \frac{2 \times Precision \times Recall}{\mathrm{Precision} + Recall},$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

모델 간 성능 비교결과는 표 2와 같다. 실험 결과, Random Forest 모델이 전반적으로 균형 잡힌 성능을 보였으며, 이에 K-Fold 교차검증을 통해 일반화 성능을 확보한 모델을 최종 모델로 채택하였다.

표 2. 제안 모델의 성능
Table 2. Performance of the proposed model

Model	Precision	Recall	F1 Score	Accuracy
Random Forest	0.78	0.73	0.75	0.73
LGBM	0.78	0.72	0.74	0.72
XGBoost	0.74	0.62	0.65	0.62

최종 모델은 정확도 0.73, 정밀도 0.78, 재현율 0.73, F1 점수 0.75 수준의 성능을 보였다. 특히, 해당 모델은 다른 모델들에 비해 비교적 0(정상 출발) 클래스에 대해 높은 성능을 보였다.

다음으로 모델에 포함된 20개의 변수들을 대상으로 변수 중요도 분석을 수행하였다. 변수 중요도 분석에서는 각 변수가 모델 예측에 기여한 정도를 확인할 수 있으며 분석 결과는 그림 3과 같다.

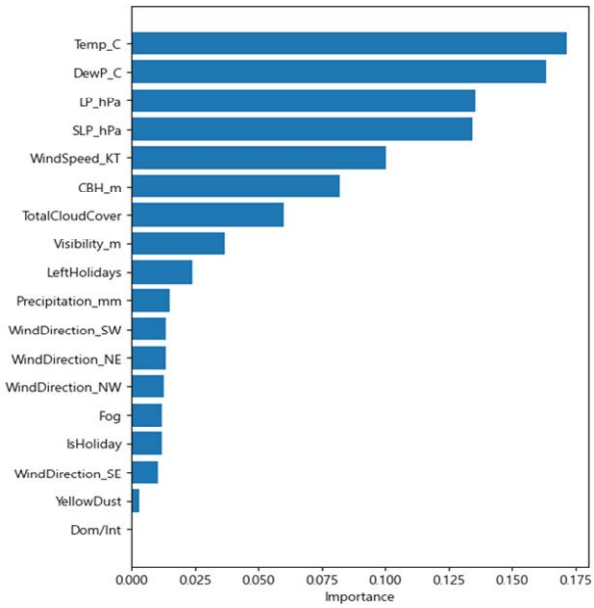


그림 3. 변수 중요도 분석 결과
Fig. 3. Feature importance analysis

결과에 따르면 기온(Temp_C), 이슬점온도(DewP_C), 해면기압(SLP_hPa), 현지기압(LP_hPa), 풍속(WindSpeed_KT), 전운고(CBH_m) 등의 변수가 중요도가 높은 변수로 식별되었다. 반면 남은 휴일 수(LeftHolidays), 휴일 여부(IsHoliday) 등의 변수는 상대적으로 중요도가 낮은 것으로 나타났다. 이러한 결과는 기존 선행 연구들과도 유사한 양상을 나타낸다. 앞서 살펴본 뉴욕 3개 공항 데이터를 활용한 연구에서는 강수량, 풍속, 가시성이 지연 여부를 예측하는 데 핵심적인 변수로 분석되었으며, 지연 시간 예측에서는 기온, 풍속, 강수량이 주요한 영향 요인으로 나타났다[8]. 제주국제공항을 대상으로 한 연구에서도 풍속, 순간돌풍, 급변풍이 항공기 지연의 핵심 변수로 분석되었고, John F. Kennedy 국제공항(JFK) 대상 연구[9]에서는 기압, 풍속, 상대습도 등의 변수가 항공기 지연 발생 가능성에 영향을 미치는 것으로 나타났다. 한편, 시카고 오헤어 국제공항(ORD) 대상 연구[1]에서는 기상 변수를 포함한 Random Forest 모델이 0.718의 정확도를 보였으나, 기상 정보를 포함하지 않은 모델의 정확도는 0.604로 감소하여, 기상 요인이 지연 예측의 필수적인 요소임을 확인한 바 있다. 본 연구에서도 기온, 이슬점온도, 풍속, 기압 등의 기상요인이 주요 변수로 나타나 선행 연구와 유사한 결과를 보였다.

반면, 강수량은 선행 연구와 달리 중요도가 낮은 변수로 분석되어 차이를 보였다. 이는 항공기 지연에 영향을 미치는 기상요인이 분석 대상 공항의 고유한 기상적 특성에 의해 달라질 수 있음을 시사한다.

본 연구에서는 추가적으로 SHAP 분석을 수행하여 모델을 해석하였다. SHAP 분석은 변수들이 모델의 예측에 미치는 영향력을 각 변수의 값을 고려하여 보여준다. 그림 4는 SHAP 분석 결과를 나타내며, X축은 항공기 지연예측에 대한 영향의 크기와 방향을 나타내며 각 변수의 색상은 빨간색에 가까울수록 변수값이 클 때, 파란색에 가까울수록 변수값이 작을 때를 의미한다. 색상 대비가 크고 좌우 폭이 클수록 해당변수의 영향력이 큰 것으로 해석할 수 있다.

분석 결과에 따르면, 기온(Temp_C), 이슬점온도(DewP_C), 전운량(TotalCloudCover), 현지기압(LP_hPa), 해면기압(SLP_hPa), 풍속(WindSpeed_KT) 등의 기상 요인이 상대적으로 높은 중요도를 보였다. 특히, 기온, 풍속, 이슬점온도가 매우 높거나 매우 낮을 경우, 항공기 지연 가능성이 증가하는 경향을 보였다. 이는 극단적인 기온이 비행기 엔진 성능 저하, 기상 악화, 난류 발생 등의 원인이 될 수 있기 때문으로 해석된다[7]. 또한, 현지기압과 해면기압 역시 높거나 낮을 때 모두 영향을 미치는 변수로 나타났다. 일반적으로 저기압 상태에서는 기상 악화(강풍, 강수 가능성 증가)로 인해 지연이 증가하지만, 기압이 지나치게 높은 경우에도 특정한 대기 흐름 변화나 풍속 증가 등의 영향으로 지연 가능성이 증가할 수 있음을 의미한다. 한편, 풍속(KT)과 운고가 높은 경우에도 지연 발생 가능성이 증가하는 경향을 보였으며, 이는 강풍 및 높은 구름층이 비행 안전에 영향을 미쳐 출발 지연을 유발할 가능성이 높음을 의미한다. 또한, 북동풍, 남서풍 등 특정 풍향(WindDirection_NE, WindDirection_SW)이 지연 발생에 상대적으로 더 큰 영향을 미치는 것으로 나타났다. 이는 특정 풍향이 항공기 운항에 불리한 영향을 줄 수 있음을 시사하며, 인천국제공항의 지리적 위치 및 주변 기후 조건과 관련이 있을 가능성이 있는 것으로 해석된다[13][14].

Ⅴ. 결론 및 시사점

본 연구에서는 항공기 지연을 예측하는 모델을 구축하고, 변수중요도와 SHAP 분석을 통하여 모델을 해석하였다. 기존 연구에서는 항공기 지연 예측을 위해 통계적 분석 또는 머신러닝 모델을 활용하였으나, 모델의 의사결정 과정이 블랙박스 형태로 작동하는 한계가 있었다.

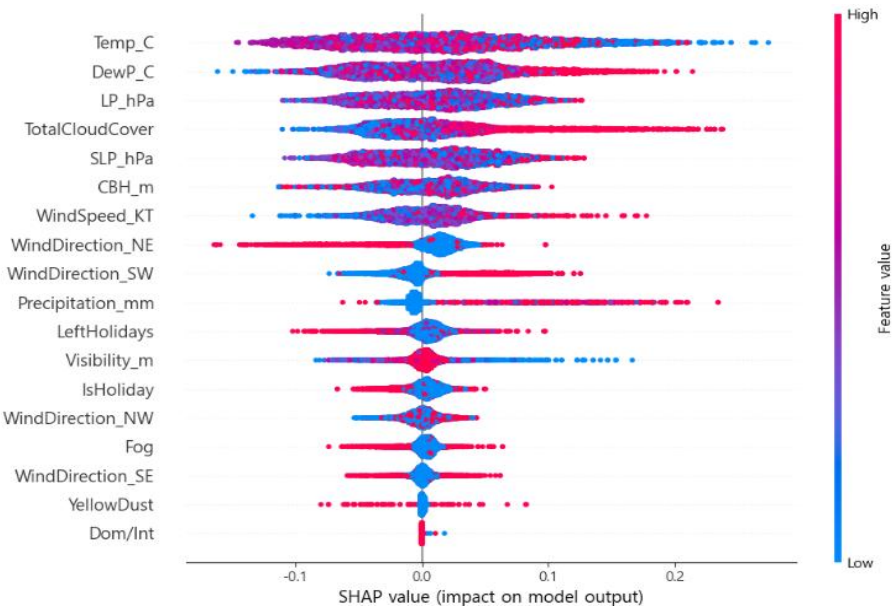


그림 4. SHAP 분석 결과
Fig. 4. SHAP analysis for delay prediction

이에 본 연구에서는 SHAP 분석을 통해 모델의 예측 결과를 설명하고, 각 변수의 상대적인 중요도를 평가할 수 있었다.

분석 결과, 기온, 이슬점온도, 해면기압 등과 같은 기상 요인이 항공기 지연에 주요한 영향을 미치는 것으로 나타났다. 기상 요인이 항공기 지연에 미치는 영향을 다룬 선행연구는 존재하지만, 본 연구는 국내 주요 거점공항인 인천국제공항의 실제 데이터를 기반으로 실증 분석을 수행하였다는 점에서 차별성을 가진다. 특히 기존 연구들이 주로 통계적 기법에 의존한 반면, 본 연구는 머신러닝 기반의 SHAP 분석을 활용하여 변수 간 복잡한 상관관계를 규명하였으며, 이를 통해 기존 연구에서 부족했던 모델 해석 가능성을 제시하였다는 점에서 실무적인 의의가 크다.

본 연구의 결과는 항공기 지연 예측 모델의 신뢰성을 향상시키고, 공항 운영 및 항공사 운항 전략 수립에 실질적인 시사점을 제공한다. SHAP 분석을 활용하여 모델의 의사결정 과정이 해석 가능해짐에 따라 항공기 지연 발생에 대한 원인을 체계적으로 파악할 수 있으며, 이를 기반으로 효과적인 대응 전략을 마련할 수 있다. 또한 기상 요인의 영향을 정량적으로 분석함으로써, 공항 기상 관제 시스템과 연계한 운항 스케줄 최적화에도 활용될 수 있을 것이다.

그러나 본 연구는 2022년 9월부터 2년간의 데이터를 기반으로 분석을 수행하였으므로, 장기적인 기후 변화를 충분히 반영하기 어려웠다. 향후 연구에서는 분석 기간을 확대하고, 계절별 및 연도별 변화를 고려한 연구를 수행함으로써 기후 변화가 항공기 지연에 미치는 영향을 확인할 필요가 있다. 또한, 예측 모델의 신뢰성을 높이기 위해, 다른 설명 가능한 인공지능(XAI, Explainable AI) 기법을 적용하여 변수별 영향이 조건에 따라 어떻게 변화하는지를 심층적으로 분석하는 연구가 필요하다. 이를 통해 신뢰성 높은 예측 모델을 구축하고, 항공 운항의 안정성과 효율성을 극대화하는 데 기여할 수 있을 것이다.

References

[1] J. Hong, J. Lee, M. Kwon, D. Kim, and S. Cho, "Analysis of Aircraft Delay Prediction based on Machine Learning Algorithms", Proc. of the 2024 Korean Institute of Communications and Information Sciences (KICS) Summer Conference, Jeju, Korea, pp. 1217-1220, Jun. 2024.

[2] S. J. Kwak, Y. S. Lee, and J. S. Lee, "Analysis and Implications of Aviation Sector SOC Accounts - Final Report", INational Assembly Research Service, pp. 1-131, Dec. 2014.

[3] M. H. Kim, S. W. Park, and J. H. Bae, "Flight Delay and Cancellation Analysis and Management Strategies", Korea Transport Institute, Nov. 2020.

[4] Bureau of Transportation Statistics, "Understanding Reporting Causes of Flight Delays and Cancellations", U.S. Department of Transportation, https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations. [accessed: Apr. 21, 2025]

[5] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions", Proce. of the 31st Conference on Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, pp. 1-10, Dec. 2017. https://doi.org/10.48550/arXiv.1705.07874.

[6] J. W. Lee, K. K. Ko, T. S. Kwon, and K. K. Lee, "A Study on the Critical Meteorological Factors Influencing the Flight Cancelation and Delay: Focusing on Domestic Airports", Journal of the Korean Society of Aeronautical Operations, Vol. 19, No. 1, pp. 29-37, Mar. 2011. https://doi.org/10.12985/ksaa.2011.19.1.029.

[7] S. H. Kim, "The Impact of Extreme Weather Events on Flight Delays & Cancellations with a Focus on Gimpo-Jeju Air Route, M.S. thesis, Dept. of Environmental Planning", Seoul National University, Seoul, South Korea, Feb. 2021.

[8] G. H. Lee and S. K. Song, "Predicting Flight Delays based on Deep Neural Network", A Collection of Academic Presentation Papers by the Korea Information Society, pp. 1075-1076, Jan. 2021. https://doi.org/10.1186/s40537-020-00380-z.

[9] C. S. Lee, Z. M. Paing, H. M. Yeo, D. S. Kim, and H. J. Baik, "Development of a Prediction Model and Correlation Analysis of Weather-induced Flight Delay at Jeju International Airport UsingMachine Learning Techniques", Journal of the Korean Society for Aviationand Aeronautics, Vol. 29, No. 4, pp. 1-20, Dec. 2021. https://doi.org/10.12985/ksaa.2021.29.4.001.

[10] B. G. Kim and Y. A. Kim, "Analysisthe Impact of Weather Data on Airline Flight Delays", Journal of the Korea Academia-Industrial Cooperation Society, Vol. 23, No. 2, pp. 222-228, Feb. 2022. https://doi.org/10.5762/KAIS.2022.23.2.222.

[11] Air Portal, https://www.airportal.go.kr/airport/aircraftInfo.do [accessed: Apr. 21, 2025]

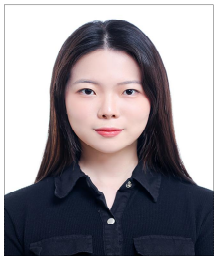
[12] Open MET Data Portal, https://data.kma.go.kr/cmmn/main.do [accessed: Apr. 21, 2025]

[13] Ministry of Land, Infrastructure and Transport, "Aviation Weather", Aviation Safety Policy Division, Ministry of Land, Infrastructure and Transport, pp. 7-310, Jan. 2021.

[14] J. W. Park, "A Study on the Meteorological Effects that Influence on the Flight Operation Obstacles on the International Airports in Korea", M.S. thesis, Department of Atmospheric Science, Chosun University, Aug. 2013.

저자소개

김 다 연 (Dayeon Kim)



2022년 3월 ~ 현재 : 명지대학교 경영정보학과 학부과정
관심분야 : 머신러닝, 데이터 분석

송 민 경 (Minkyung Song)

2022년 3월 ~ 현재 : 명지대학교 경영정보학과 학부과정
관심분야 : 머신러닝, 데이터 분석

이 석 종 (Seokjong Lee)

2020년 3월 ~ 현재 : 명지대학교 경영정보학과 학부과정
관심분야 : 머신러닝, 데이터 분석, ERP

장 유 진 (Youjin Jang)

2021년 3월 ~ 현재 : 명지대학교 경영정보학과 학부과정
관심분야 : 머신러닝, 데이터 분석, ERP

이 한 준 (Hanjun Lee)

2001년 2월 : 서울대학교 컴퓨터공학과(공학사)
2004년 2월 : 서울대학교 컴퓨터공학과(공학석사)
2016년 8월 : 고려대학교 경영학과 MIS 전공(경영학박사)
2007년 7월 ~ 2018년 2월 : 한국국방연구원 선임연구원
2018년 3월 ~ 2020년 2월 : 한남대학교 조교수
2020년 3월 ~ 현재 : 명지대학교 부교수
관심분야 : 머신러닝, 자연어 처리, 정보화 정책