Journal of the Korea Academia-Industrial cooperation Society
Vol. 23, No. 2 pp. 222-228, 2022

https://doi.org/10.5762/KAIS.2022.23.2.222
ISSN 1975-4701 / eISSN 2288-4688

기상 데이터가 항공사 운항 지연에 미치는 영향 분석

김봉기[1*], 김영아[2]
[1]경상국립대학교 컴퓨터공학부, [2](주)En-Core

Analysis the Impact of Weather Data on Airline Flight Delays

Bong-Gi Kim[1*], Yeong-A Kim[2]
[1]School of Computer Engineering, Gyeongsang National University
[2]En-Core Co.,Ltd

요 약 4차 산업 혁명을 대변하는 핵심 키워드로는 인공지능, 빅데이터, 사물인터넷, 클라우드, 블록체인 등이 있다. 데이터의 범람으로 수많은 데이터들 중에서 필요한 지식을 뽑아내어, 현재 주요하게 떠오르는 핵심 이슈를 파악하고 미래를 예측하기 위해 데이터를 분석하고 해석할 수 있는 빅데이터의 활용이 대단히 중요해졌다. 본 본문에서는 항공기 지연 및 결항 예측을 위한 분석 프레임워크를 제안하고, 공항에 영향을 주는 다양한 요인들 중에서 기상 데이터가 항공 운항 지연 및 결항에 미치는 영향을 분석함으로써 항공사 운항 지연 예측 시스템 구축의 토대를 마련하고자 한다. 로지스틱 회귀, 랜덤 포레스트, 서포트 벡터 머신 등 몇 가지 알고리즘을 이용한 탐색적 데이터 분석 결과 날씨가 항공 운항 지연에 영향을 준다는 결론을 얻었다. 본 연구의 결과를 기초로 하여 향후 항공기의 지연 및 결항 확률 예측력을 높이는 다양한 원인들에 대한 심층적인 분석과 이에 따른 항공기 지연 및 예측 시스템 구축에 대한 연구가 계속되어야 할 것이다.

Abstract Artificial intelligence, big data, the internet of things, cloud, and blockchain can be cited as keywords representing the 4th industrial revolution. Due to the overflow of data, it has become important to extract necessary knowledge from a large number of data and use big data that can analyze and interpret data to identify major issues currently emerging and predict the future. This research intends to suggest an analysis framework that can predict flight delays and cancellations and lay the groundwork for implementing a flight delay prediction system. In particular, this system is developed by analyzing the effect of weather data on flight delays and cancellations among various factors that affect airports. Through the exploratory data analysis using several algorithms such as logistic regression, random forest, and support vector machine, this research concluded that weather affects flight delays. Based on the conclusion of this research, in-depth analysis of various reasons that raise the predictive power of probability of flight delays and cancellations is necessary. In addition, the conclusion indicates that the study of the establishment of a flight delay prediction system should continue.

Keywords : Weather Data, Big Data, Data Analysis, Machine Learning, EDA

1. 서론

4차 산업혁명을 대변하는 핵심적인 키워드로는 사물인터넷, 인공지능, 빅데이터, 블록체인 등이 있다. 정보기술의 혁신적 발전과 모바일 기기의 대중화 그리고 인터넷의 광역화 및 초고속화는 사람들이 언제 어디서든 쉽고 편리하게 정보처리가 가능한 환경을 제공하고 있다. 이러한 환경하에서 사람들은 SNS(Social Network

본 논문은 2020, 2021년도 경상국립대학교 대학회계 연구비 지원에 의하여 연구되었음.
*Corresponding Author : Bong-Gi Kim(GNU)
email: bgkim@gnu.ac.kr
Received January 7, 2022 Revised January 21, 2022
Accepted February 4, 2022 Published February 28, 2022

Service)의 이용확산을 가져왔고, SNS를 통해 수집된 정보를 재생산하여 타인에게 전달하는 과정을 수행하고 있다. 이는 급속한 데이터의 양적 팽창을 가져왔고 기존에 우리가 접할 수 있었던 데이터 보다 훨씬 방대하고 다양한 구조를 가진다는 것을 알 수 있다[1].

아날로그 환경에서의 정형화된 데이터에 비해 실시간으로 생성되어 주기가 짧은 데이터와 텍스트 및 이미지 등을 포함하는 방대한 규모의 데이터를 활용하여 사회적 가치를 창출하고자 하는 빅데이터 분석 및 활용에 대한 관심과 수요는 나날이 높아지고 있다. 최근 학계에서도 빅데이터를 분석 기반으로 하는 다양한 연구들이 늘어나고 있으며, 민간 기업과 연관 산업에서도 빅데이터를 분석하고 이를 통한 효과적인 마케팅과 커뮤니케이션 전략을 수립하고 진행하고 있다. 또한 빅데이터 분석을 통한 수익화를 위해 다양한 비즈니스가 등장하고 있다. 이는 정보화 사회에서 정치・경제・사회・문화 등 여러 분야에 걸쳐 데이터 정보 분석이 중요해졌으며, 데이터를 누가, 어디에, 어떻게 활용하느냐에 따라 그 가치가 결정되는 시대를 맞이했기 때문일 것이다. 더욱이 데이터의 범람으로 수많은 데이터 중에 필요한 지식을 뽑아내서 현재 주요하게 떠오르는 이슈를 파악하고 미래를 예측하기 위해 그것을 분석하고 해석할 수 있는 빅데이터 활용이 중요해졌으며 이에 대한 연구가 활발히 진행되고 있다[2].

본 연구에서는 데이터 분석 방법들을 고찰해 보고, 공항의 다양한 지연 및 결항 사유들 중에서 기상 데이터가 항공 운항의 지연 및 결항에 미치는 영향을 분석해 봄으로써 차후 항공사 운항 지연 예측 시스템을 구축하는 기반을 마련하고, 항공기 지연에 따른 승객 및 항공사의 불편을 해소하고자 하는 것이 목적이다.

본 논문의 구성은 다음과 같다. 2장에서는 빅데이터 관련 기술과 기계 학습에 관련된 연구에 대하여 살펴보고, 3장에서는 본 논문에서 제안하는 분석 프레임워크로 항공사 운항 지연 예측을 모델링하여 검증하고, 기상 데이터가 항공 운항의 지연 및 결항에 미치는 영향을 탐색적 분석을 통해 결과를 살펴보고 4장에서는 결론을 내린다.

2. 관련 연구

2.1 빅데이터 관련 기술 고찰

빅데이터란 기존의 데이터 단위를 넘어서는 엄청난 양(Volume), 데이터의 생성과 흐름이 매우 빠르게 진행되는 속도(Velocity), 사진・동영상 등 기존의 구조화된 데이터가 아닌 다양한(Variety) 형태의 정보 등 3가지 속성을 가지는 큰 데이터이다.

목적을 정하고 그 목적에 맞도록 데이터를 어떻게 효과적으로 저장하고 분석할 수 있을까 하는 데이터베이스 기술과 빅데이터는 차이가 있다. 빅데이터는 잘 가공된 데이터가 아닌 버려지고 있는 데이터들에 대한 관심이 부각되면서 기술에 변화를 가져왔다.

빅데이터를 처리하는 과정은 크게 데이터의 생성, 수집, 저장, 처리, 분석, 표현의 과정으로 나누어진다. 데이터의 수집은 내부 파일 시스템이나 데이터베이스 관리 시스템, 센서 등에 접근하여 정형 데이터를 수집하는 내부 데이터 수집과 크롤링, 오픈 API와 같은 툴이나 프로그래밍으로 비정형 데이터를 자동으로 수집하는 외부 데이터 수집으로 구분된다. 데이터 저장은 분산 파일 시스템, NoSQL, 병렬 DBMS, 네트워크 구성 저장 시스템 등 다양한 접근 방식이 있으며 특히 빅데이터는 대용량, 비정형, 실시간 속성을 수용할 수 있고 대량의 데이터를 파일 형태로 저장할 수 있는 기술과 비 정형 데이터를 정형화된 데이터 형태로 저장하는 기술이 필요하다. 빅데이터 처리 기술로는 정형, 비정형 빅데이터 분석에 가장 선호되는 솔루션인 하둡을 이용한다. 맵리듀스 기술은 일반 범용 서버로 구성된 군집화 시스템을 기반으로 〈키, 값〉 입력 데이터 분할 처리 및 처리 결과 통합 기술, Job 스케줄링 기술, 작업 분배 기술, 장애에 대처하는 태스크 재수행 기술 등이 통합된 분산 컴퓨팅 기술이다. 빅데이터 분석에 사용하는 것들은 대부분 통계학과 전산학, 특히 기계 학습과 데이터 마이닝 분야에서 사용한 텍스트 마이닝, 소셜 네트워크 분석, 분류, 군집화 방법을 적용한다. 빅데이터 표현 기술은 데이터 분석 결과를 효과적으로 전달하려고 어렵고 복잡한 정보를 한눈에 쉽게 이해할 수 있도록 간단한 도표나 3D 이미지 등으로 표현하는 정보 표현 기술이 발전하였다[3].

2.2 기계 학습

기계 학습(Machine Learning)은 어떤 목적을 달성하기 위해 기계가 데이터를 이용해 어떤 지식이나 행동을 학습하도록 하는 기술이다. 크게 지도 학습(Supervised Learning)과 비지도 학습(Unsupervised Learning), 강화 학습(Rainforcement Learning)으로 구분된다. 기계가 지식이나 행동을 학습하도록 하는데 사용하는 데이터를 훈련 데이터라고 하는데 이 훈련 데이터가 목표 변수와 설명 변수를 포함하는지에 따라 지도 학습과 비지도 학습으로 나눈다[4].

2.2.1 지도학습

지도 학습은 훈련 데이터를 이용해 데이터에 포함된 특정 변수를 예측하는 모델을 구축하는 방법이다.

$$y = f(x) \quad (1)$$

Where, y denotes target variable, x denotes explanatory variable, f denotes function

식 (1)과 같은 함수가 있을 때 y가 목표 변수, x가 설명 변수, 함수(f)가 모델이다. 지도 학습은 목표 변수의 데이터 형식이 주가와 같은 수치라면 회귀(Regression) 문제로, 남성·여성과 같은 범주라면 분류(Classification) 문제로 구분한다[4]. 계속해서 대표적인 지도 학습 알고리즘을 간단하게 고찰해 보고자 한다.

(1) K-Nearest Neighbor

직관적이고 간단한 알고리즘이다. 새로운 데이터가 주어졌을 때 이를 클래스 A로 분류할지, 클래스 B로 분류할지를 판단하여 더 많은 데이터가 포함되어 있는 범주로 분류하는 방식이다. Fig. 1은 KNN을 이해하는 데 도움이 되는 그림이다. KNN에서는 데이터와 데이터 사이의 거리를 구해야 하는데 거리를 구하는 방식은 유클리드(Euclidean Distance) 거리와 맨해튼(Manhattan Distance) 거리 방식이 있다.

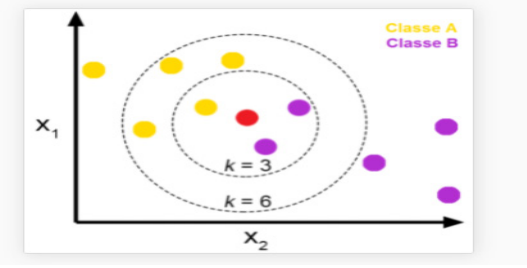


Fig. 1. Principle of KNN
Source : https://bkshin.tistory.com/entry/

2차원에 있는 두 점(a, b) 사이의 거리는 식 (2)의 유클리드 거리 계산 방법과 식 (3)의 맨해튼 거리 계산 방법으로 구할 수 있다.

$$d = \mathrm{sqrt}((a_1 - b_1)^2 + (a_2 - b_2)^2) \quad (2)$$

$$d = |a_1 - b_1| + |a_2 - b_2| \quad (3)$$

(2) Logistic Regression

회귀를 사용하여 데이터가 어떤 범주에 속할 확률을 0에서 1 사이의 값으로 예측하고 그 확률에 따라 가능성이 더높은 범주에 속하는 것으로 분류해주는 지도 학습 알고리즘이다. 일반적인 회귀 모델과의 차이점은 오차 제곱합(SSE)을 최소로 갖는 회귀선을 찾는 것이 아닌 시그모이드(Sigmoid) 함수 최적선을 찾아 이 시그모이드 함수의 반환 값을 확률로 간주하여 분류를 진행한다. 시그모이드 함수는 식 (4)와 같다.

$$\varnothing = 1 / (1 + e^{-z}) \quad (4)$$

Where, e denotes natural constant, z denotes output of linear equations

(3) Decision Tree

의사결정 트리는 어떤 목적에 도달하기 위해 데이터의 각 속성에 따라 조건 분기를 반복하면서 전체 자료를 몇 개의 소집단으로 분류하거나 예측(Prediction)을 수행하는 분석 방법이다. 목표 변수가 범주일 경우 분류나무, 수치일 때는 회귀 나무라고도 한다[4]. 의사결정 트리는 이해하기 쉽고, 예측할 때 사용하는 프로세스가 명백하다. 숫자형 데이터와 범주형 데이터를 동시에 다룰 수 있고, 특정 변수의 값이 누락 되어도 사용 가능한 장점이 있다[5]. Fig. 2는 의사결정 나무를 일반화한 예이다.

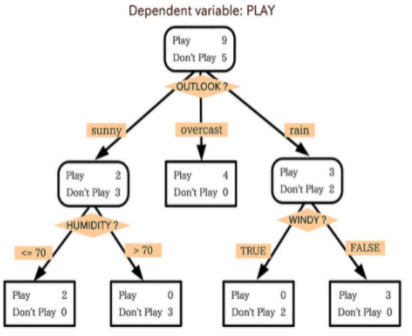


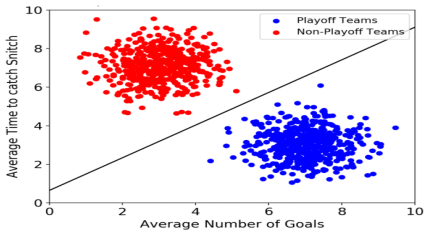
Fig. 2. Example of Decision Tree
Source : https://ratsgo.github.io/machine%20learning/2017/03/26/tree/

(4) 랜덤 포레스트(Random Forest)

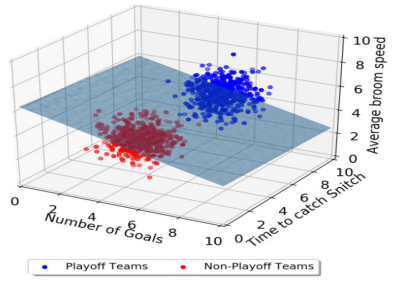
의사결정 나무는 오버피팅(Overfitting)될 가능성이 있다는 약점을 가지고 있다. 랜덤 포레스트는 이 약점을 해결하고자 여러 개의 의사결정 나무를 만들어 다수결로 결과물을 종합하는 방식이다[5]. 즉, 여러 개의 의사결정 트리를 형성하고 새로운 데이터 포인트를 각 트리에 동시에 통과시키며, 각 트리가 분류한 결과에서 투표를 실시하여 가장 많이 득표한 결과를 최종 분류 결과로 선택한다. 분류나무라면 나무의 결과로 투표를 하면 되고, 회귀나무라면 예측값의 평균을 내면 된다[6].

(5) 서포트 벡터 머신(Support Vector Machine)

서포트 벡터 머신은 결정 경계(Decision Boundary), 즉 분류를 위한 기준선을 정의하는 모델이다. 분류되지 않은 새로운 점이 나타나면 경계의 어느 쪽에 속하는지를 확인해서 분류 과제를 수행한다. 결정 경계는 데이터가 2개 속성만 있다면 결정 경계는 선 형태가 될 것이고, 속성이 3개로 늘어나면 3차원으로 그려야 한다. 이때의 결정 경계는 선이 아닌 평면이 된다. Fig. 3은 서포트 벡터 머신의 결정 경계를 나타내는 예이다.



(a) A Decision Boundary in Two Dimensions



(b) An SVM using Data with Three Features

Fig. 3. Example of Decision Boundary in SVM
Source : https://hleecaster.com/ml-svm-concept/

2.2.2 비지도 학습

비지도 학습은 목표 변수가 없고 데이터 그 자체에 주목해 데이터에 잠재된 패턴이나 시사점을 발견하는 방법이다. 대표적인 비지도 학습은 데이터를 유사한 그룹으로 나누는 군집 분석과 데이터의 정보를 잃지 않으면서 변수의 개수를 축소하는 주성분 분석이 있다[4]. 아래에서 대표적인 비지도 학습 알고리즘을 간단하게 고찰해 보고자 한다.

(1) 주성분 분석(Principal Component Analysis)

여러 개의 양적 변수(Quantitative Variable)들 사이의 분산-공분산 관계를 이용하여 변수들의 선형 결합 으로 표시되는 주성분을 찾고, 2-3개의 주성분으로 전체 변동(Variance)의 대부분을 설명하고자 하는 다변량 분석법이다. 주성분 분석의 순서는 다음과 같다.

① 표준화, 정규화(Standardization)
② Z 공분산 매트릭스 계산
③ Z 공분산 매트릭스에서 고유벡터, 고유값 계산
④ 원 데이터를 고유벡터에 정사영(Projection)하여 매트릭스 변환

(2) K-평균 분석(K-Means Analysis)

K-평균 분석 알고리즘은 효율적이고 강력한 성능으로 가장 많이 사용되는 군집화 알고리즘이다. K는 주어진 데이터로부터 그룹화할 그룹, 즉 클러스터의 수를 말하고, 평균은 각 클러스터의 중심과 데이터들의 평균 거리를 의미한다. 이때 클러스터의 중심을 Centroids라고 한다. K-평균 분석 알고리즘은 다음과 같은 과정으로 수행한다[6].

① 군집 K개의 중심 초깃값을 결정한다.
② 거리를 기준으로 모든 데이터 표본을 군집에 할당한다.
③ 각 군집의 데이터 표본의 평균값으로 군집의 중심을 새로 결정한다.
④ 종료 조건을 만족할 때까지 ②와 ③을 반복한다.

(3) Hierachical Clustering Analysis

계층적 트리 모형을 이용해 개별 개체들을 순차적, 계층적으로 유사한 개체 내지 그룹과 통합하여 군집화를 수행하는 알고리즘이다. K-평균 군집과 달리 군집 수를 사전에 정의하지 않고, 학습 이후 군집 수를 선택한다. 계층적 군집 분석을 수행하려면 모든 개체들 간의 거리나 유사도(Similarity)가 이미 계산되어 있어야 하는데 거리 계산은 유클리드 거리 계산식을 이용한다. 군집 방법에는 군집들 사이의 거리를 이용하여 가까운 거리에 있는 군집들을 묶거나(합병법), 먼 거리에 있는 군집들을 차례로 분리해 나가는 방법(분리법)으로 크게 두 가지의 접근 방법이 있으며, 세부적인 방법으로는 최단거리법(Nearest Neighbor Method), 최장거리법(Futherest Neighbor Method), 집단 평균법(Group Average Method), 중앙값법(Median Method), 워드법(Words's Method) 등이 있다. Fig. 4는 계층적 군집 분석에서 개체들이 결합되는 순서를 덴드로그램(Dendrogram)으로 시각화하고 적절하게 트리를 잘라 군집을 나눌 수 있음을 보여준다.

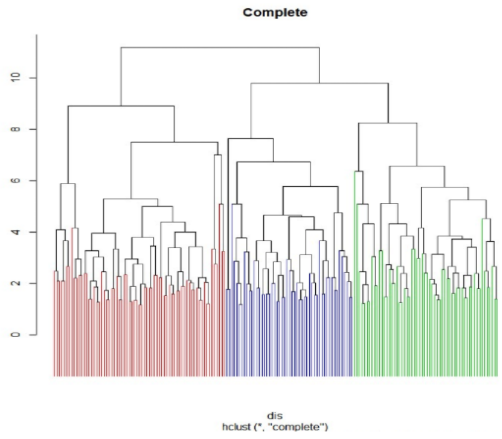


Fig. 4. Result of Clustering with Dendrogram and Cutree
Source : https://ybeaning.tistory.com/25

3. 항공사 운항 지연 예측 분석

최근 들어 항공 교통에 대한 수요는 지속적으로 증가하고 있으며, 이와 더불어 공항의 지연 및 결항이 늘어나고 이로 인해 공항 이용자들의 불편과 피해도 늘어나고 있다. 따라서 공항의 지연 및 결항의 발생 확률을 예측할 수 있다면 공항 운영에 대한 적절한 조치를 취할 수 있어 항공기 지연에 따른 승객 및 항공사의 불편을 해소할 수 있을 것이다.

본 논문에서는 항공기 지연 및 결항 예측을 위한 분석 프레임워크를 제안하고, 예측 모델을 구성하기 위해 비지도 학습 알고리즘을 모델링하고 검증하여 공항의 다양한 지연 및 결항 사유들 중에서 기상으로 인한 항공기 지연 및 결항의 발생 확률을 예측하고자 날씨와 관련된 변수들이 항공기 지연에 미치는 영향을 다양한 데이터 분석 방법으로 탐색하고자 한다.

3.1 분석 프레임워크

데이터는 Kaggle 사이트에서 2015년 John F.Kennedy International Airport에서 출발하는 항공편에 대한 데이터를 기반으로 구축하였다. 데이터 수집은 Openweathermap.org에서 API를 통한 날씨 데이터를 Json 형식으로 수집하였다. 분석은 항공기 지연 및 결항 예측에 영향을 미치는 속성을 단계별로 구성하여 기상 요소들을 예측 모형의 독립 변수로 사용하였다. 예측 모형으로는 Logistic Regression, SVM, Random Forest, Neural Network을 사용하였다. Fig. 5는 항공기 지연 및 결항 예측을 위한 분석 프레임워크를 나타낸 것이다.

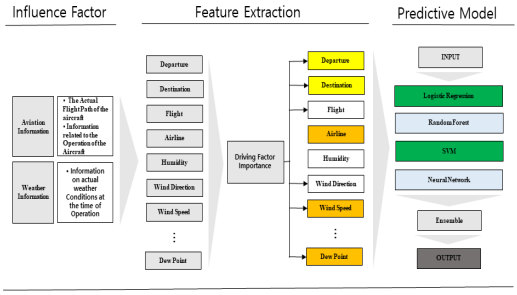
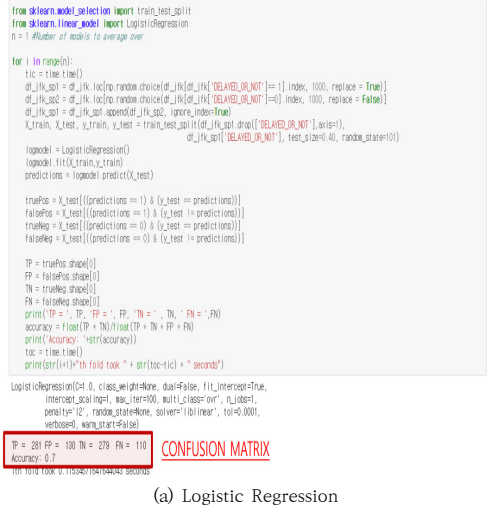


Fig. 5. An Analytical Framework for Predicting Flight Late and Cancellations

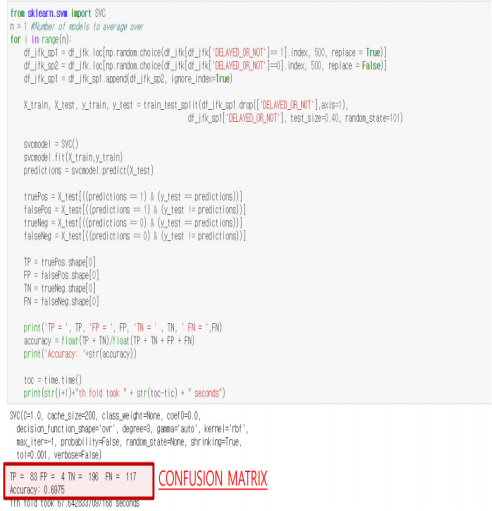
3.2 모델링 검증

2015년 항공편 데이터와 날씨 데이터를 하나의 데이터 프레임(Data Frame)으로 병합하여 분석하였다. 그리고 Numerical Analysis와 Graphical Analysis를 동시에 진행하여 변수들이 지연에 미치는 영향을 다각도에서 탐색할 수 있도록 했다. 각 데이터 셋에 들어 있는 자료구조를 확인하고, 분석에 활용하기 위한 데이터 변형 및 결측치 대체 방법을 강구하여 데이터 분석에 활용할 수 있도록 자료형을 변형하였다. Fig. 6은 예측 모델을 구성하기 위해 비지도 학습 알고리즘 중에서 Logistic Regression, Random Forest, SVM 알고리즘을 이용하여 분석하고 모델링한 결과를 나타낸 것이다. 분석 결과 Logistic Regression의 정확도는 70%, Random Forest의 정확도는 78%, SVM의 정확도는 70%로 나타났다.



(a) Logistic Regression

(b) Random Forest



(c) SVM

Fig. 6. Analysis Result and Modeling Validation

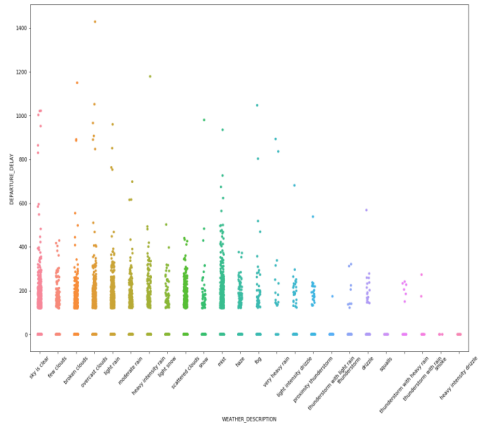
3.3 EDA(Exploratory Data Analysis) 결과 요약

데이터의 평균, 중위수, 최빈값, 표준편차 등 기초 통계량과 중심점, 분포성, 산포성 등 데이터 자체 특성 분석을 통해 데이터의 통계적 특성을 이해하고 일변량, 다변량 통계 분석 및 시각화를 통해 데이터를 이해하고 모델링을 위한 기초 자료로 활용하였다. 이를 통해 날씨 특성별로 항공기 출발 지연 간의 관계를 탐색적 데이터 분석을 하여 Table 1과 같은 결과를 얻었다.

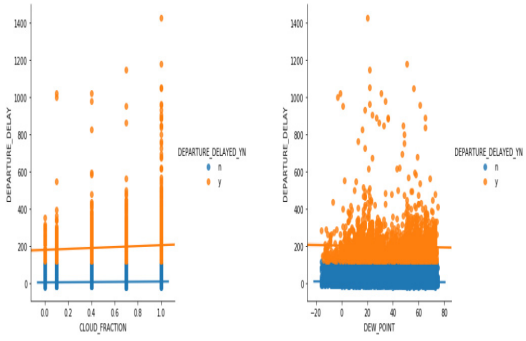
Fig. 7은 Weather-Description과 항공기 지연 간의 상관관계를 분석한 것을 시각화하여 보여준 것이다.

Table 1. Result of EDA

Weather Description	Result of EDA
thunderstorm variable	humidity, pressure, and mslp were judged to be significant
snow variable	humidity and wind_speed were judged to be significant
sky_is_clear variable	wind_speed and cloud_fraction show easily distinguishable
proximity thunderstorm variable	pressure, temperature, wind_speed and relative_humidity were judged to be significant
light intensity drizzle variable	humidity and temperature seem to affect the situation where delay is not possible



(a) Analysis of the Relationship between Weather-Description and Flight Departure Delay



(b) Analysis of the Relationship between Specific Weather Variables and Flight Departure Delay

Fig. 7. Visualization of Relationship Analysis between Weather and Flight Departure Delays : for New York JFK Airport

공항 지연 및 결항 요인을 분석한 결과 항공기의 비정상 운항은 항공기 스케줄과 연관성이 높게 나타나지만, 해당 시간대의 기상 데이터에 따라 차이가 나는 것을 알 수 있다. 즉 날씨가 항공기 지연 및 결항에 영향을 미친다는 것을 확인할 수 있다.

4. 결론

본 논문은 데이터 활용의 중요성을 이해하고자 데이터 분석 방법들을 고찰하고, 날씨가 항공기 운항 지연과 결항에 미치는 영향을 분석하여, 차후 항공기 지연에 따른 승객 및 항공사의 불편을 해소하기 위한 항공사 운항 지연 예측 시스템을 구축하는 토대를 마련하고자 연구를 수행하였다.

본 연구에서는 공항 지연 및 결항 요인을 분석한 결과 항공기의 비정상 운항은 항공기 스케줄과 연관성이 높게 나타나지만, 해당 시간대의 기상 데이터에 따라 차이가 나는 것을 알 수 있었다.

다만 본 연구에서는 날씨 데이터가 항공기 지연에 미치는 영향을 분석하는 것으로 끝이 났지만, 공항 지연과 결항 예측에 있어서 고려해야 할 부분은 매우 많고 까다롭다는 사실을 알 수 있다. 따라서 전체 항공기의 지연, 결항 확률 예측력을 높이기 위해서는 다양한 원인들을 복합적으로 조사하고 변수를 통계하여 분석을 해야 할 것이다. 빅데이터를 활용한 항공기 지연 및 결항 예측은 단순히 항공 운송 이용자의 편의를 증진시킬 뿐만 아니라 항공 운송 사업자인 항공사와 공항의 효율성과 비용 절감에도 크게 기여할 수 있을 것이다. 향후 항공기의 지연, 결항 확률 예측력을 높이는 다양한 원인들에 대한 심층적인 분석과 이에 따른 항공기 지연 및 예측 시스템 구축에 대한 연구가 계속되어야 할 것이다.

References

[1] T. Jeong-Mee Lee, "Understanding Big Data and Utilizing its Analysis into Library and Information Services", Journal of the Korean BIBLIA Society for Library and Information Science, Vol.24, No.4, pp.53-73, 2013.

[2] Do-Sik Choi, "Problems of Big Data Analysis Education and Their Solutions", Journal of the Korea Convergence Society, Vol.8, No.124, pp.265-274, 2017.

[3] Sang-Min Lee, "Big Data and Industrial Application", KISTI, pp.25-28, 2015.

[4] Jaewon Choi, "Data Science Training Book", Insightbook, pp.212-283, 2020.

[5] Joel Grus, "Data Science from Scratch 2/E", Insight Press, pp. 236, 2020.

[6] Geun-woo Choi, "Learning Data Science for the First Time", Hanbit Media Inc., pp. 171, 2020.

김 봉 기(Bong-Gi Kim) [종신회원]

- 1989년 2월 : 숭실대학교 대학원 전자계산학과 (공학석사)
- 1999년 2월 : 숭실대학교 대학원 전자계산학과 (공학박사)
- 1994년 3월 ~ 1999년 2월 : 한림전문대학 조교수
- 1999년 3월 ~ 2021년 2월 : 경남과학기술대학교 컴퓨터공학부 교수
- 1999년 3월 ~ 현재 : 경상국립대학교 컴퓨터공학부 교수

〈관심분야〉
데이터베이스, 빅데이터

김 영 아(Yeong-A Kim) [정회원]

- 2019년 8월 : 경남과학기술대학교 컴퓨터공학과(공학석사)
- 2017년 1월 ~ 현재 : ㈜ En-Core HRD 연구원

〈관심분야〉
빅데이터, 정보보안, 네트워크, 인공지능, ICT, IoT