

## 항공편 결항 예측 모델 연구

조규철\*, 김예지<sup>o</sup>, 전동준<sup>o</sup>

\*인하공업전문대학 컴퓨터정보과,

<sup>o</sup>인하공업전문대학 컴퓨터정보과

e-mail: kccho@inhac.ac.kr\*, yeji4268@naver.com<sup>o</sup>, djeon0912@naver.com<sup>o</sup>

## The Research of Prediction for Flight Cancellation

Cho Kyu Cheol\*, Kim Ye Ji<sup>o</sup>, Jeon Dong Jun<sup>o</sup>

\*Dept. of Computer Science, Inha Technical College,

<sup>o</sup>Dept. of Computer Science, Inha Technical College

### ● 요약 ●

본 연구에서는 항공편 결항 시, 이용객이 겪게 되는 시간적 / 비용적 피해를 최소화하기 위해 머신러닝 · 딥러닝 기법을 이용하여 항공편 결항 예측 모델을 제안한다. 이 모델은 5가지 이진 분류기법을 사용하여 과거 2017년~2021년 제주공항 기상 데이터와 항공편 스케줄 데이터를 병합하여 결항, 출발을 분류한다. 본 연구는 기상으로 인한 항공편 결항의 피해 최소화를 목적으로 한다.

**키워드:** 머신러닝, 인공 신경망, 이진 분류

### I. 서론

코로나 19의 영향으로 지난 2년 동안, 국내/외 여행의 수요가 대폭 감소하였지만, 최근 사회적 거리두기가 전면 해제된 이후로 위축됐던 여행 심리가 빠르게 되살아나면서 여행 수요가 점차 회복되고 있다.

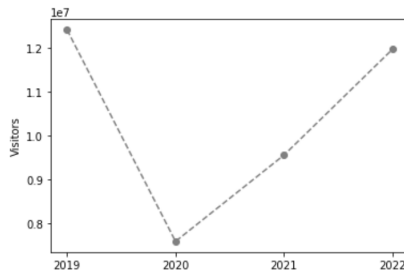


Fig. 1. 제주공항 이용객 추이

그러나 항공편의 지연과 결항은 이러한 여행의 변수 중 하나가 될 수 있다. 현재의 항공기 결항 예측은 해당 공항에 발효되는 기상 특보에 따라 결항 판단을 내리는 절차를 가지고 있다. 그러나 기상 특보의 특성상 급하게 내려지는 경우가 있어, 이용객에게 가는 피해는 불가피하다. 따라서 항공기 지연 및 결항에 대한 피해를 최소화하고자 본 연구를 진행하였다.

본 연구는 항공 기상청과 항공 포털에서 데이터를 크롤링하여 사용하였다. 이진 분류를 위한 모델로는 로지스틱 회귀, 의사 결정 나무, SVM, 나이브 베이즈, DNN 모델을 활용하여 연구를 진행하였다. 이 연구를 통해 기상예보처럼 사용자에게 예측한 결항 여부를 알려주어 사용자의 행동 결정에 도움을 줄 것으로 기대한다.

### II. 데이터 수집 및 전처리

항공기 결항에 가장 영향을 미치는 주요 요인은 기상 요인이 대표적으로 나타났고 기상 데이터를 이용해 관련 연구에 많이 적용되고 있다. 따라서 본 연구에는 기상 데이터와 항공편 스케줄 데이터를 사용한다. 기상 데이터는 제주 공항의 2017년부터 2021년의 데이터를 수집하고 항공편 스케줄에 대한 결항 데이터는 항공 포털에서 수집하였다. 그리고 수집한 데이터를 병합하는 전처리 과정을 통해 하나의 데이터 셋을 생성한다. 전처리 과정으로 만들어진 해당 데이터는 강수량, 기온, 풍향, 풍속 등으로 구성되었고 클래스 값은 지연 이륙은 0, 결항은 1로 구성하였다.

### III. 항공편 결항 예측모델 실험

#### 1. 로지스틱 회귀

로지스틱 회귀는 회귀를 사용한 지도 학습 알고리즘으로, 데이터가 범주에 속할 확률을 예측하고 확률이 더 높은 범주에 속하는 것으로 분류 예측하는 알고리즘이다. 로지스틱 회귀 모델 생성 전, 모델의 정확도와 과적합 방지를 위해 데이터 스케일링 작업을 진행한다.

#### 2. 의사 결정 나무

의사 결정 나무는 클래스 값을 결정하는 규칙을 트리 구조로 나타내어 전체 데이터를 특정 범주로 분류하거나 예측을 수행하는 모델이다. 의사 결정 나무 모델 생성 시, 모델의 복잡도를 조절하여 성능을 개선할 수 있는 `max_depth` 파라미터의 최적값을 찾아가며 진행한다.

#### 3. 서포트 벡터 머신(SVM)

서포트 벡터 머신(SVM)은 데이터 집합과 가장 멀리 떨어진 선 또는 면을 찾아내어 분류를 위한 기준선인 결정 경계로 사용하고 이를 통해 데이터를 분류한다. SVM의 `Gamma` 파라미터는 하나의 데이터 샘플이 영향력을 행사하는 거리로, 결정 경계의 유연성을 조절하여 모델의 과적합을 방지하고 최적의 결정 경계를 탐색하기 위해 사용된다.

#### 4. 나이브 베이즈 분류

나이브 베이즈는 특성들 사이의 독립을 가정하는 베이즈 정리를 적용한 확률 분류기의 일종이다. 나이브 베이즈의 장점으로는 지도 학습 환경에서 효율적으로 훈련이 될 수 있고, 다양한 데이터 유형에 사용 가능하고 적은 데이터로도 효율적인 성능을 낼 수 있다는 것에 있다. 분류에 사용하고자 하는 데이터는 연속형 데이터이기 때문에 가우시안 베이즈 분류를 사용한다. 가우시안 베이즈의 경우 모델 생성 시 사용 데이터의 비율과 무관하게 사전 확률을 지정하여 사용된다.

#### 5. DNN(Deep Neural Network)

인공 신경망 기법 중 하나인 DNN은 기존 ANN의 파라미터의 최적값을 찾기 어렵다는 단점을 보완하여 모델 내 은닉층 수를 통해 최적의 파라미터 값을 찾을 수 있다. DNN의 은닉층의 개수, 은닉층을 구성하는 노드 개수, 드롭아웃 비율, 배치 사이즈 조절을 통해 최적의 성능을 내는 값을 찾는다.

### IV. 성능 평가 및 모델 비교

항공편 결항 예측 모델을 위한 데이터 셋에 대해서 5가지의 예측 분류 모델들을 이용하여 각 모델 별로 정확도, 정밀도, 재현율, F1-score를 측정하여 성능을 비교하였다.

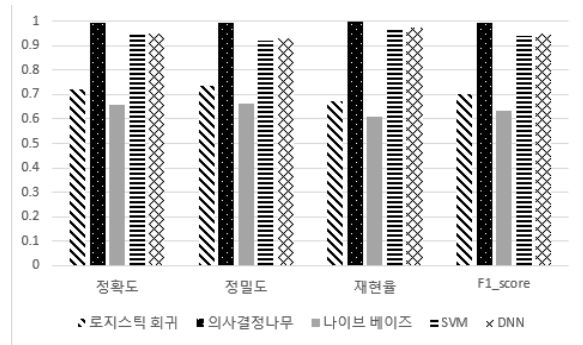


Fig. 2. 항공편 결항예측모델 성능 비교

그림 2와 같이 지표별로 순위가 동일하게 나타났고, 의사결정나무, DNN, SVM, 로지스틱 회귀, 나이브 베이즈 순으로 높은 성능을 보여준다.

### V. 결론

항공편의 결항 여부 예측을 위해 5가지의 분류 모델을 설계하고 각 모델의 성능 평가 결과로 분류가 가장 잘 된 모델을 제공하여 결항을 예측함으로써 항공편 결항으로 인한 이용자의 피해 최소화와 행동 결정에 도움을 줄 것이다.

### REFERENCES

- [1] Korea Airports Corporation, "https://airport.co.kr"
- [2] Aviation Meteorological Office "https://amo.kma.go.kr"
- [3] AirPotal , "https://www.airportal.go.kr"