항공기 지연 예측 프로젝트

디지털스마트 부산 아카데미 SW 전문인재 양성사업

3조 이위성 이영빈 이선오

DIGITAL SMART BUSAN ACADEMY

SHINSEGAE I&C

1. 서론

- **문제 상황**

(1) 항공사 측 문제점 (연방항공국 조사 데이터의 수치를 기반)

2019년 미국 항공사 지연율 20%, 고객 지연 보상의 비용이 329억 달러, 한화로 약 44조 7000억의 비용이 듦, 항공사 측은 이러한 비용에 대한 많은 부담감을 느끼고 있음

(2) 이용객 측 문제점 (항공서비스 소비자피해 실태조사 내용)

1시간 이전 + 지연 보상에 대한 애매한 기준으로 많은 부분에서 보상이 이뤄지지 않는 상황이 발생, 이에 이용객들이 많은 불편함을 느낌

- **목표**

항공사와 이용객의 항공기 지연 관련 문제를 해결을 위한 항공기 지연 예측 머신러닝 및 딥러닝 모델 개발

2. 본론

(1) 사용 데이터 셋

데이콘 항공기 지연 예측 데이터 set

- Target 변수 : 항공기의 Delay 여부 (Not_delay, Delay)
- Feature 변수 : 출발/도착공항_주, 출발/도착공항고유번호, 이동거리, 출발/도착_예측시간, 항공사 등 (항공기의 지연 여부와 각 Column의 특징을 알 수 있는 변수 18개로 구성)

→ 해당 데이터를 전처리하여 결측치 복구 및 제거, 사용하지 않을 Column을 제거 한 후 15개의 Column을 사용

(2) 데이터 EDA (탐구 데이터 분석 과정)

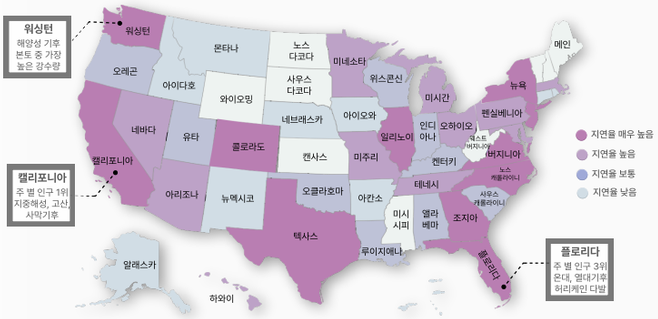
	월	일	출발 시간	도착 시간	출발 항공사 ID	도착 항공사 ID	운행 거리
월	1.00	-0.00	-0.00	-0.01	-0.02	0.02	0.02
일	0.01	1.00	-0.00	-0.00	0.00	-0.00	0.00
출발 시간	-0.00	-0.00	1.00	0.68	-0.03	0.02	-0.01
도착 시간	-0.01	-0.00	0.68	1.00	-0.01	0.01	0.01
출발 항공사 ID	-0.01	-0.00	-0.03	-0.01	1.00	0.02	0.06
도착 항공사 ID	-0.01	-0.00	0.02	0.01	0.02	1.00	0.06
운행 거리	-0.02	-0.00	-0.01	0.01	0.06	0.06	1.00

- **상관관계분석**

출발 시간과 도착 시간만 높은 수치를 보여줌
대체적으로 상관관계가 없음

장점
- 독립성 유지
- 다중공산성 문제 감소

단점
- 복잡한 상호작용 무시
- 고차원→ 일반화 성능 떨어짐



- **지역별 지연율 시각화**

대체적으로 인구가 많고, 날씨가 좋지 않은 곳에 지연이 빈번하게 일어남

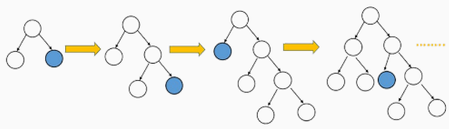
(3) 항공기 지연 예측 모델 개발 구현

01	Scaler / Encoder	Standard Scaler / Label Encoder
02	Feature	['월', '일', '출발 예측 시간', '도착 예측 시간', '경유 여부', '출발 공항 코드', '출발 공항 ID', '출발 공항(주)', '도착 공항 코드', '도착 공항 ID', '도착 공항(주)', '이동 거리', '지연 여부']
03	Split	Train : Test : Validation (%) = 70 : 20 : 10
04	Target	지연 여부 [Not Delayed : 0 / Delayed : 1]
05	Model	DL : 다층 퍼셉트론(MLP) ML : Logistic Regression, K-Nearest Neighbors, XGBoost, LightGBM

- **불균형한 Target 갯수**
  → Under/Over Sampling을 통해 Label의 갯수 균일화
- **'출발/도착공항_주' 같은 경우는 Label Encoder 처리**

사용 모델 : Logistic Regression, KNN, XGB, LGBM, MLP
모델의 평가기준 : 정확도 / 재현율을 기준으로 높은 순서

- **최적 모델 : LGBM (Light Gradient Boosting Model)**

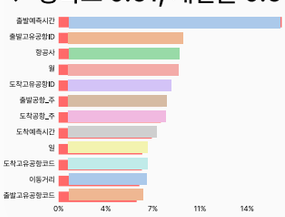


모델 소개 : 분류와 회귀 영역에서 뛰어난 예측 성능을 가진 XGB 모델의 학습 속도를 개선한 모델, 의사 결정 알고리즘 사용

정확도 : **0.61** 재현율 : **0.61**

n_estimators=1000
force_col_wise=True
num_leaves=64
boost_from_average=False

Parameter Tuning을 위한 GridSearch, Hyper opt 적용 결과
→ 정확도 0.61, 재현율 0.60로 기존 LGBM 모델보다 낮은 수치



LGBM 모델의 최종적 Feature 중요도
- 출발예측시간 : 17%
- 출발고유공항ID : 9.1%
- 항공사 : 8.8%

출발지와 해당 위치 항공사가 많은 영향을 주는 모습을 보여주고 있음
→ 해당 위치의 날씨와 지역적 특성에 영향

결론적으로, 기존의 LGBM 모델을 이용한 Target 변수인 항공기 지연여부를 Feature를 통해 분류해주는 모델 생성을 할 수 있었음

3. 결론

(1) 의의
- 탐색적 데이터 분석(EDA)를 통한 항공 관련 인사이트 발견
- 다양한 모델 구현을 통한 일정 성능을 가진 LGBM 모델 개발

(2) 활용 방안
- 항공사 자체 데이터를 사용해 항공편 지연을 예측 가능
- 지연 확률이 높은 항공편에 부가 비용 예측 / 새로운 가격 책정
- 운항 스케줄의 조정 등의 대응 정책 구축 가능

(3) 한계점
- Feature의 다양성이 적음 → 날씨/계절등의 데이터의 부재
- 예측 모델의 성능이 높지 않아 활용 가능성이 낮음

(4) 개선 방안
- 항공편 데이터 추가 확보 및 기존 데이터 전처리 방법 고안 후 다양성 확보
- 다양한 모델 실험을 통한 성능의 개선을 시도