[양식 6] 1-2PBL 결과보고서

[■ Project) Based Learning 결과보고서

1 팀 구성

제 (오! 분석해조) 팀 구성

팀명	오! 분석해조			
프로젝트 주제	(항공 지연 원인) 빅데이터 분석			
구분	학과	학번	성명	역할
팀장	항공컴퓨터학과	202400755	이해인	보고서 작성. 데이터 정제, 데이터 가공, 최종 R 스튜디오 코딩
팀원	항공컴퓨터학과	201800427	신동우	데이터셋 발표 데이터 취득 자료 조사, 데이터 분석, 시각화
팀원	홍보정보학과	202100647	조정헌	ppt 제작, 데이터 취득 자료 조사
팀원	항공컴퓨터학과	202400723	김마리	K-NN 발표, K-NN 자료 조사 데이터 취득 자료 조사

3 활동결과

가. (오! 분석해조) 팀 활동 경과

일시	주요활동 내용	참석인원
2025.04.09.	◦ 역할 분담 ◦ 세부 주제 정하기	4
2025.04.16.	◦ 세부 주제 정하기 ◦ 데이터 조사 바탕으로 방향 정하기	4
2025.05.14.	◦ 방향 정한 후 데이터셋 탐색 및 분석 ◦ 발표 형식 정하고 연습해보기	4

❒ 활동일지

<table>
<tr><td rowspan="3">미팅
개요</td><td>일시</td><td colspan="2">2025.04.09.</td><td>문제해결과정</td><td>역할 분담</td></tr>
<tr><td>참석자</td><td colspan="4">신동우, 조정헌, 이해인, 김마리</td></tr>
<tr><td>본 미팅의
주요활동</td><td colspan="4">조이름 선정 및 주제 회의</td></tr>
<tr><td rowspan="3">진행
사항</td><td>구분</td><td colspan="3">활동 내용</td><td>조치 사항</td></tr>
<tr><td>이번
미팅에서
한 일</td><td colspan="3">조 이름을 5조로 정한 뒤, KOSIS(통계청), 데이터저널리즘코리아, 지방자치단체, Kaggle 등의 홈페이지에서 데이터를 탐색하며 주제를 논의함.</td><td>각 조원들의 의견을 듣고 역할을 나눈 뒤, 각 조원들이 관심 있어 하는 분야에서 주제를 정할 것.</td></tr>
<tr><td>논의 사항</td><td colspan="4">세부 주제가 정해지지 않음.</td></tr>
<tr><td rowspan="3">추후
계획</td><td>구분</td><td>활동 내용</td><td>역할 분담</td><td>추진 일정</td></tr>
<tr><td>다음
미팅에서
해야 할 일</td><td>이번 미팅에서 이야기된 항공기 지연, 난입 예측, 날씨 예측 등의 주제를 바탕으로 주제를 선정하기로 함.</td><td>주제 선정을 어떻게 할지 논의하고, 각자 관련된 데이터셋을 탐색해 오기로 함.</td><td>2025년 4월 16일</td></tr>
<tr><td>기타</td><td colspan="3">주제 선정은 아직 명확하지 않지만, 데이터를 찾을 때 항공과 관련된 내용이 많이 탐색되어 해당 주제로 선정될 가능성이 높다고 예상됨.</td></tr>
<tr><td>피드백</td><td>교수님의
피드백
사항</td><td colspan="3"></td></tr>
</table>

<table>
<tr><td rowspan="3">미팅
개요</td><td>일시</td><td colspan="2">2025.04.16.</td><td>문제해결과정</td><td>데이터 선택, 취합, 정제</td></tr>
<tr><td>참석자</td><td colspan="4">신동우, 조정헌, 이해인, 김마리</td></tr>
<tr><td>본 미팅의
주요활동</td><td colspan="4">지난 회의에서 선정한 항공기 지연 예측 주제를 바탕으로, 본 미팅에서는 분석에 사용할 데이터를 선택하고, 취합 및 정제하는 방향을 중심으로 논의를 진행함.</td></tr>
<tr><td rowspan="3">진행
사항</td><td>구분</td><td colspan="3">활동 내용</td><td>조치 사항</td></tr>
<tr><td>이번
미팅에서
한 일</td><td colspan="3">항공기 지연 예측에 활용할 수 있는 다양한 공개 데이터셋(Kaggle 항공 지연 데이터, 기상청 날씨 정보, 공항별 항공편 이력 등)을 조사하고, 각 데이터의 특징과 분석 가능 항목을 공유함. 항공 지연에 영향을 줄 수 있는 변수(출발 시간, 기상 조건, 항공사, 요일 등)를 중심으로 필요한 데이터를 어떤 방식으로 조합할지 논의함. 이후 분석을 위한 데이터셋을 만들기 위해 불필요한 열 제거, 결측값 처리 방안, 변수 단위 통일 등 정제 작업의 방향을 논의함.</td><td>여러 데이터셋 중에 기간도 최근이고, 항공사별 지연율도 알 수 있는 데이터셋을 중심으로 데이터를 선택, 취합하였음</td></tr>
<tr><td>논의 사항</td><td colspan="4">결측값이 많은 경우 해당 데이터를 제외할지, 대체값을 사용할지에 대한 논의 진행</td></tr>
<tr><td rowspan="3">추후
계획</td><td>구분</td><td>활동 내용</td><td>역할 분담</td><td>추진 일정</td></tr>
<tr><td>다음
미팅에서
해야 할 일</td><td>정제한 데이터셋을 바탕으로 탐색적 데이터 분석(EDA)을 실시하고, 변수 간 분포 및 상관관계를 시각화하여 데이터의 특성을 파악할 예정임. 또한, 예측 정확도를 높이기 위해 사용할 수 있는 모델링 기법(랜덤 포레스트, 로지스틱 회귀 등)에 대해 간단한 구조를 분석할 예정임.</td><td>신동우: 데이터 취득 자료 조사, 데이터 분석
조정헌: 데이터 취득 자료 조사
이해인: 주차 보고서 작성, 조사된 데이터를 바탕으로 데이터 정제
김마리: 데이터 취득 자료 조사</td><td>2025년 5월 14일</td></tr>
<tr><td>기타</td><td colspan="3">EDA와 예측 모델 설계 수행 준비하기</td></tr>
<tr><td>피드백</td><td>교수님의
피드백
사항</td><td colspan="3"></td></tr>
</table>

<table>
<tr><td rowspan="3">미팅
개요</td><td>일시</td><td>2025.05.14.</td><td>문제해결과정</td><td>EDA와 예측 모델 설계</td></tr>
<tr><td>참석자</td><td colspan="3">신동우, 조정헌, 이해인, 김마리</td></tr>
<tr><td>본 미팅의
주요활동</td><td colspan="3">데이터 전처리를 바탕으로 EDA, 시각화, 모델링을 진행하고 R로 최종 코딩을 마무리 함.</td></tr>
<tr><td rowspan="3">진행
사항</td><td>구분</td><td colspan="2">활동 내용</td><td>조치 사항</td></tr>
<tr><td>이번
미팅에서
한 일</td><td colspan="2">R을 활용해 항공 지연 데이터를 기반으로 EDA를 수행함. 결측치와 이상치를 처리하고, 항공사, 출발 공항, 시간대에 따른 지연 패턴을 시각화를 통해 분석함. ggplot2 패키지를 사용해 주요 결과를 시각적으로 표현함. 이후 지연 요인을 분석하기 위한 예측 모델을 설계하고, 랜덤 포레스트와 로지스틱 회귀 모델을 비교함. 모델 성능은 정확도, 정밀도, 재현율 등을 통해 평가함. 발표를 위해 분석 및 모델링 내용을 정리함</td><td>초기 데이터에서 결측치와 이상치가 다수 발견되어 해당 값을 제거하거나 평균/최빈값으로 대체함. 변수 간 상관관계가 낮은 항목은 분석에서 제외하여 모델 성능 향상을 도모함.</td></tr>
<tr><td>논의 사항</td><td colspan="3">결측치 처리 방법과 모델 선택에 대해 논의함</td></tr>
<tr><td rowspan="3">추후
계획</td><td>구분</td><td>활동 내용</td><td>역할 분담</td><td>추진 일정</td></tr>
<tr><td>다음
미팅에서
해야 할 일</td><td>발표 자료 만들고 준비하기</td><td>각자 담당한 파트별로 발표 자료 만들기</td><td>2025년 5월 21일</td></tr>
<tr><td>기타</td><td colspan="3">R 스튜디오 코딩 오류 확인</td></tr>
<tr><td>피드백</td><td>교수님의
피드백
사항</td><td colspan="3"></td></tr>
</table>

4 최종 결과보고서

■ 제 (오! 분석해조) 팀 프로젝트 결과 (항공 지연 원인, 로지스틱회귀)

1. 서론 및 데이터셋 개요

항공편 지연은 승객, 항공사, 공항 운영자뿐만 아니라 국가 경제 전반에도 중대한 영향을 미친다. 미국 연방항공청(FAA)의 2010년 보고서에 따르면 항공 지연으로 인한 연간 경제적 비용은 약 329억 달러에 달하며, 그중 절반 이상이 승객의 시간 손실과 추가 비용 부담으로 이어진다. 이러한 문제의 심각성을 바탕으로 본 보고서는 항공편 지연의 주요 요인을 규명하고, 그 패턴을 시각적으로 분석하며, 향후 지연을 예측할 수 있는 기반 모델을 탐색하고자 한다.
데이터는 Kaggle에서 제공하는 Airline Delay Cause 데이터셋을 활용하였으며, 이 데이터는 2013년부터 2023년까지의 미국 내 항공편 정보를 포함하고 있다. 주요 변수로는 항공사 코드(carrier), 공항 코드(airport), 항공편 수(arr_flights), 지연된 항공편 수(arr_del15), 다양한 지연 사유(항공사, 날씨, 시스템, 보안, 이전 항공편 등), 총 지연 시간, 취소 및 우회 현황 등이 있다. 이 데이터셋은 날짜 정보를 포함하고 있지 않아 시계열 분석에는 제약이 있으나, 연도(year) 및 월(month) 단위의 추세 분석은 가능하다.

2. 데이터 전처리

분석에 앞서 데이터 정제 과정을 수행하였다. 먼저 수치형 변수의 결측값은 0으로 대체하고, 범주형 변수는 "Unknown"으로 대체하였다. 중복된 행은 제거하였으며, 음수값을 포함하는 이상치는 확인 후 모두 필터링하여 제거하였다. 파생 변수로는 연도와 월 정보를 결합하여 "year_month" 문자열을 생성하였고, 각 항공편에 대해 전체 지연율(delay_rate)과 취소율(cancel_rate)을 계산하여 변수로 추가하였다.
예시 코드:

```
# 예시: year-month 조합 (문자열로 생성)
df$year_month ← paste(df$year, df$month, sep = "-")

# 지연율 및 취소율 계산
df$delay_rate ← df$arr_del15 / df$arr_flights
df$cancel_rate ← df$arr_cancelled / df$arr_flights
```

3. 탐색적 데이터 분석(EDA)

탐색적 데이터 분석을 통해 데이터의 구조적 특성과 변수 간 관계를 이해하고자 하였다. 항공사별 지연율을 분석한 결과, 일부 저비용 항공사 및 지역 항공사의 지연율이 상대적으로 높게 나타났다. 예를 들어, 지연율이 가장 높은 상위 10개 항공사는 평균적으로 전체 항공편의 30% 이상이 지연된 것으로 확인되었다.
공항별로는 혼잡도가 높은 대형 허브 공항이나 기후 조건이 까다로운 지역의 공항에서 지연율이 높게 나타났다. 또한 월별 분석 결과, 겨울철(12월1월) 및 여름철(6월7월)에 지연이 집중되는 경향이 뚜렷하였다. 이는 계절적 기상 요인과 여행 수요의 급증이 주요 원인으로 작용함을 보여준다.
예시 코드:

```r
# ——— 월별 지연 패턴 분석 ———
# year, month 열이 있다고 가정
monthly_delay ← df %>%
  group_by(year, month) %>%
  summarise(
    arr_flights = sum(arr_flights, na.rm = TRUE),
    arr_del15 = sum(arr_del15, na.rm = TRUE),
    .groups = "drop"  # 그룹화를 해제
  ) %>%
  mutate(delay_rate = arr_del15 / arr_flights)

cat("\n월별 지연율:\n")
print(monthly_delay)
```

이와 같이 요약된 지연율 데이터를 기반으로 연도-월 단위의 지연 패턴을 시각화함으로써, 연도별 계절 패턴을 시각적으로 파악할 수 있다.

4. 지연 원인 분석

총 지연 시간의 구성 비율을 분석한 결과, 가장 큰 원인은 이전 항공편 지연(Late Aircraft Delay)으로 약 37.4%를 차지하였다. 이어서 항공사 내부 요인(Carrier Delay, 25.2%), 국가 항공 시스템 지연(NAS Delay, 23.5%), 날씨 문제(Weather Delay, 13.7%) 순이었으며, 보안 지연(Security Delay)은 전체 지연의 0.2%로 가장 낮았다.
항공사별로 주요 지연 원인의 비중을 계산해 본 결과, 일부 항공사는 항공사 자체 요인이나 시스템 지연의 비중이 높았으며, 대형 항공사는 상대적으로 이전 항공편의 지연 비중이 컸다. 이러한 분석은 각 항공사의 운영 특성과 시스템 의존도에 따라 지연 원인이 다르게 나타난다는 점을 시사한다.
예시 코드:

```r
# 항공사별 주요 지연 원인 분석
airline_delay_causes ← df %>%
  group_by(carrier_name) %>%
  summarise(
    carrier_delay = sum(carrier_delay, na.rm = TRUE),
    weather_delay = sum(weather_delay, na.rm = TRUE),
    nas_delay = sum(nas_delay, na.rm = TRUE),
    security_delay = sum(security_delay, na.rm = TRUE),
    late_aircraft_delay = sum(late_aircraft_delay, na.rm = TRUE),
    arr_delay = sum(arr_delay, na.rm = TRUE)
  ) %>%
  mutate(
    carrier_delay_pct = round((carrier_delay / arr_delay * 100), 2),
    weather_delay_pct = round((weather_delay / arr_delay * 100), 2),
    nas_delay_pct = round((nas_delay / arr_delay * 100), 2),
    security_delay_pct = round((security_delay / arr_delay * 100), 2),
    late_aircraft_delay_pct = round((late_aircraft_delay / arr_delay * 100), 2)
  )
```

이 코드는 전체 지연 시간에서 항공사, 기상, 시스템, 보안, 이전 항공기 지연이 각각 차지하는 비율을 계산하고, 항공사별로 주요 지연 원인의 상대적 비중을 분석하는 데 사용된다.

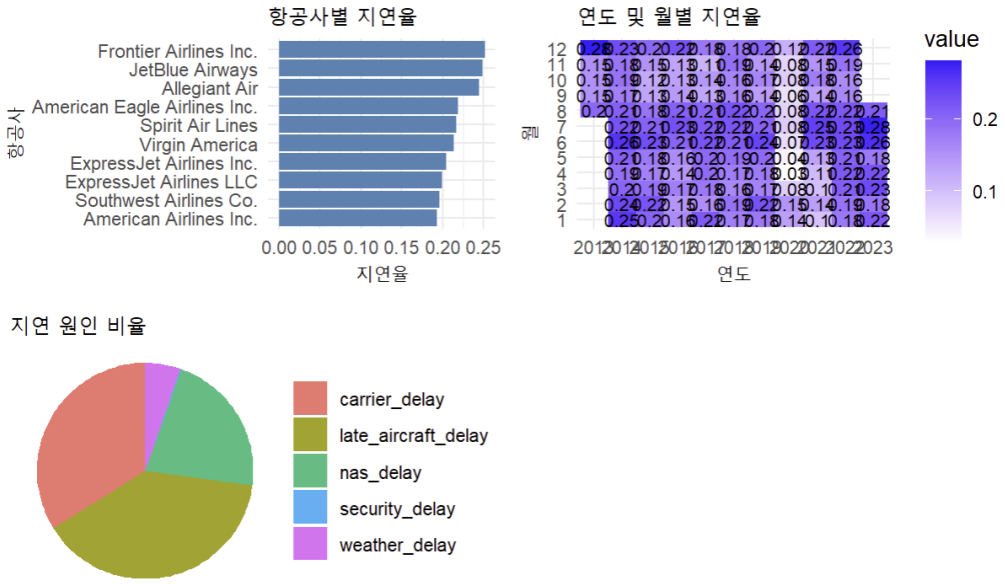
5. 시각화 기법

데이터의 주요 특징과 지연 패턴을 효과적으로 전달하기 위해 다양한 시각화를 수행하였다. 다음은 주요 시각화 기법이다.

- 항공사별 지연율: 지연율 기준 상위 10개 항공사를 막대그래프로 시각화하여 비교하였다.
- 연도 및 월별 지연율 히트맵: 연도-월 형태로 정리된 지연율 데이터를 히트맵으로 표현하여 계절별 추세를 직

관적으로 파악하였다.

- 지연 원인 비율 파이 차트: 전체 지연 시간의 원인을 시각적으로 보여주어 각 요인의 상대적 중요성을 쉽게 파악할 수 있도록 하였다.
- 지연 시간 분포 히스토그램: 평균 지연 시간의 분포를 히스토그램 및 커널 밀도 그래프로 나타내어 이상값 여부와 전반적인 분포 특성을 확인하였다.



이러한 시각화 결과는 보고서의 핵심 메시지를 직관적으로 전달하는 데 효과적이며, 다양한 이해관계자(항공사, 공항 운영자, 정책 입안자)에게 유용한 정보를 제공한다.

6. 예측 모델 구축 및 성능 평가

예측 모델은 각 항공편의 지연 발생 가능성을 이진 분류 문제로 설정하여 구축하였다. 예측 타겟은 delay_rate가 0.2(20%)를 초과하는 경우를 지연으로 간주하였다. 사용된 주요 변수는 다음과 같이 연도, 월, 항공사, 공항, 취소율 등이다. 모델링에는 다음과 같은 분류 모델이 사용되었다.

- 로지스틱 회귀 (Logistic Regression)
- 랜덤 포레스트 (Random Forest)
- 그래디언트 부스팅 머신 (GBM)

R 언어에서는 caret 패키지를 활용하여 데이터 전처리, 학습, 예측, 교차 검증을 통합적으로 수행하였다. 모델 성능 평가는 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 점수, 교차 검증 점수 등을 기준으로 이루어졌다.
분석 결과, 랜덤 포레스트가 가장 높은 예측 정확도와 안정적인 성능을 보였으며, 주요 변수의 중요도를 기준으로 분석한 결과, 항공사와 공항 변수의 영향력이 가장 큰 것으로 나타났다. 이는 실제 운영상의 지연 패턴과도 일치한다.

7. 결론 및 인사이트

(1) 계절성 및 요일 패턴: 겨울철과 여름 휴가철에 지연이 집중되며, 금요일과 일요일의 지연율이 높다.
(2) 항공사 및 공항 특성: 저비용 항공사 및 기후 영향이 큰 지역 공항에서 지연이 빈번하다.
(3) 지연 원인의 차이: 항공사별로 지연 원인이 다르며, Late Aircraft 지연이 전체에서 가장 큰 비중을 차지한다.
(4) 모델 기반 예측 가능성: 머신러닝 모델을 통해 지연 가능성을 사전 예측할 수 있으며, 이는 운영 최적화 및 승객 안내에 활용될 수 있다.

이러한 분석 결과는 항공사에는 운항 스케줄 관리 개선, 공항에는 자원 배분 전략 수립, 승객에게는 지연 회피를 위한 정보 제공 등 다양한 실용적 활용 가능성을 제시한다.

8. 참고 문헌

- Kaggle의 "Flight Delay Data Analysis" 코드
- Fabien Daniel, "Predicting flight delays" 튜토리얼
- Adrian Vera, "Flight Delay EDA“
- 미국 연방항공청 항공 지연 비용 보고서
- 미국 교통통계국 항공 지연 원인 데이터

본 보고서는 항공 지연 데이터를 통해 유의미한 패턴과 예측 가능성을 확인하였으며, 이를 바탕으로 항공업계와 정책 결정자, 그리고 일반 승객에게 실용적인 인사이트를 제공하는 것을 목표로 한다.