감자마을친구들

2019 BIG CONTEST 퓨처스리그

지연 없는 완벽한 비행을 위해

당신의 시간은 금이다.

〈항공 운항 데이터를 활용한 항공 지연 예측〉

INDEX

1. 분석 주제 이해

분석 주제 / 주제 관련 이슈

1.1 분석 주제

1. 분석 주제 이해

- 분석 주제

분석 주제

- 항공 운항 데이터를 활용한 **항공 지연 예측.**
- 항공 시즌 스케줄, 운항데이터 등 항공운항데이터(한국공항공사)와 항공기상 데이터 등을 활용해 항공지연예측 모형을 개발하여서 **9월 16일부터 9월 30일 까지의** 항공편별 지연 여부 예측.

항공 운항 데이터 등 각종 데이터 활용

↓

항공기 운항 지연의 원인과 문제점 파악

↓

지연에 영향을 미치는 요인을 바탕으로
해결방안 제시

1.2 주제 관련 이슈

1. 분석 주제 이해

- 주제 관련 이슈

국내선 항공기 지연, 연결과정서 주로 발생…오후 3시 가장 빈번

저비용항공사 국내선 10대 중 2대 지연…군산공항 지연율 29.8%

[여행의 기술] 폭설로 출발 지연된 항공편, 보상 어디까지?

정비 문제로 잇따른 결항·지연…국토부, 아시아나항공 특별점검

[속보] 제주공항 윈드시어 발동, 비행기 취소

강한 비바람에 김해공항 40편 결항·16편 지연

항공기 지연은
연결편, 정비, 기상, 환자 등
다양한 원인에 의해 발생

2. 데이터 준비 및 탐색

데이터 준비 / 데이터 탐색 / 비식별화 데이터 유추

2.1 데이터 준비

2. 데이터 준비 및 탐색

- 데이터 준비

flightradar24 LIVE AIR TRAFFIC

KAC 한국공항공사 KOREA AIRPORTS CORPORATION

기상자료개방포털

AIM 항공정보관리체계

GREAT CIRCLE MAPPER

항공기상청 Aviation Meteorological Office

	추가 준비 데이터 리스트	비고
1	공항명	공항
2	부지 면적	공항
3	활주로 길이	공항
4	지역명	기상
5	시간 (Y/M/D/H)	기상
6	기온	기상
7	이슬점온도	기상
8	기압	기상
9	공항간 직선 거리	항로
10	항공기별 비행 항로	항로
11	ENROUTE CHART	항로
⋮		

제공 받은 데이터 외, 분석을 위한
추가적인 데이터 준비

2.2 데이터 탐색

2. 데이터 준비 및 탐색

- 데이터 탐색

AFSNT.CSV

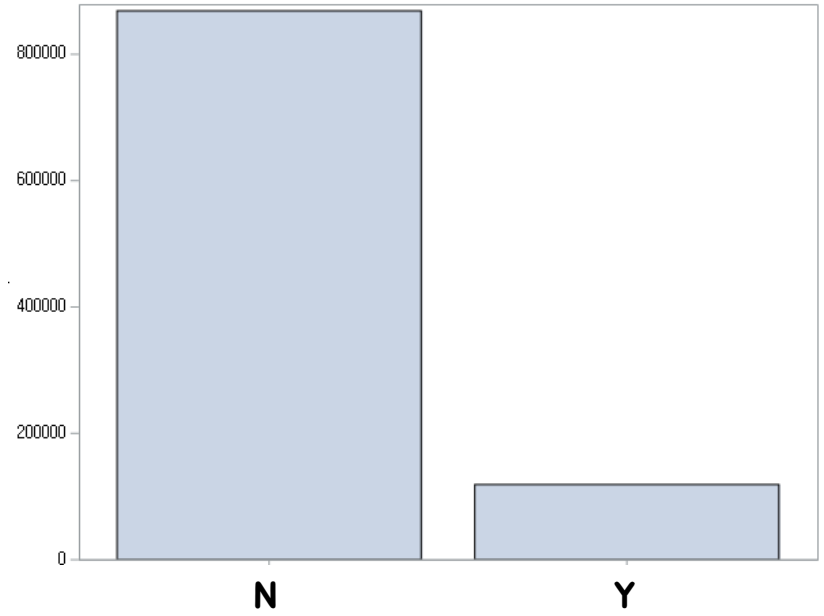
2017.1.1 ~ 2019.6.30 까지의
운항실적 데이터 (Train Set)

변수	형식	설명
SDT_YY	CHAR	연
SDT_MM	CHAR	월
SDT_DD	CHAR	일
SDT_DY	CHAR	요일
ARP	CHAR	공항
ODP	CHAR	상대공항
FLO	CHAR	항공사
FLT	CHAR	편명
REG	CHAR	등록기호
AOD	CHAR	출도착
IRR	CHAR	부정기편
STT	CHAR	계획시각
ATT	CHAR	실제시각
DLY	CHAR	타겟변수 / 지연여부
DRR	CHAR	지연사유
~~CNL~~	~~CHAR~~	~~결항여부~~
~~CNR~~	~~CHAR~~	~~결항사유~~

결항된 항공편은 Test Set에서 제거되므로
결항 관련 변수는 분석에서 제외하기로 결정함.

<타겟변수>

☞ DLY (지연여부)



DLY	빈도	백분율
N	868772	87.96
Y	118937	12.04

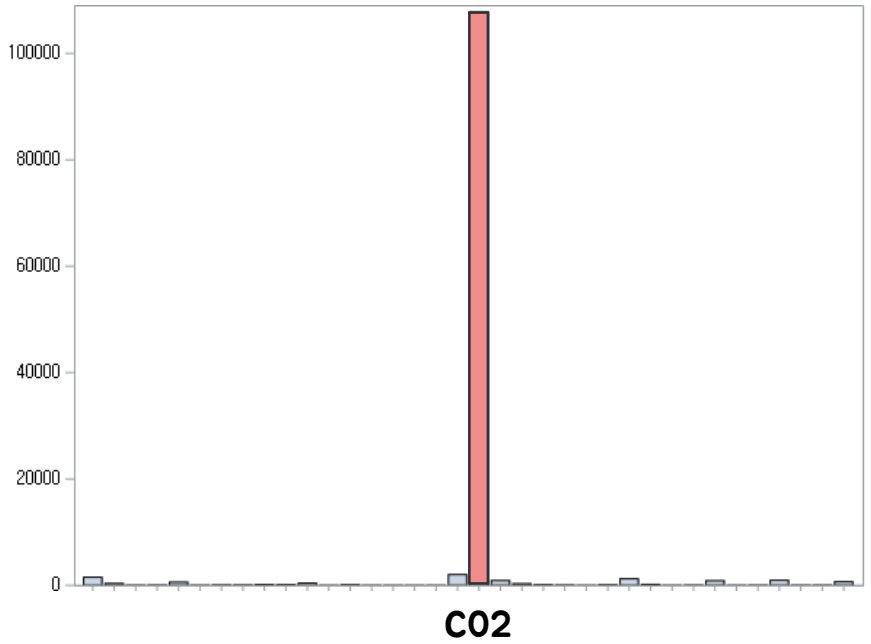
- 타겟변수의 계급별 분포 즉, 지연과 비지연의 비율이 약 1 : 9인 Unbalanced Data임을 확인함.
- 이후 전처리 과정에서 Sampling을 통하여 각 계급에 비슷한 수준의 학습이 가능하게 처리해야함.

2. 데이터 준비 및 탐색

- 데이터 탐색

<지연사유별 지연의 빈도>

☞ DRR (지연사유)



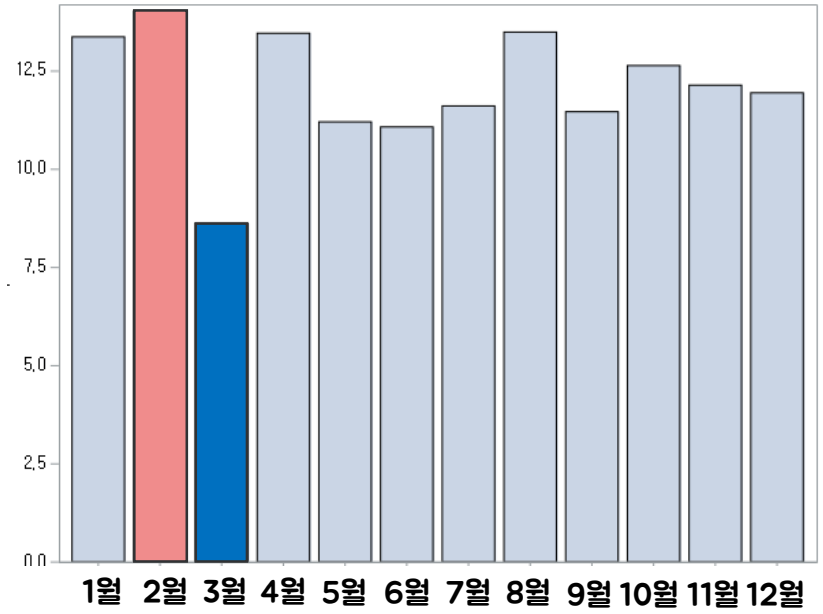
상위 5개 DRR	지연 빈도	백분율
A01	1524	1.28
C01	2031	1.71
C02	107738	90.58
C10	1227	1.03
D01	950	0.80

지연 CODE	분류	백분율
A	기상 사정	2.56
B	공항 사정	0.46
C	항공기 사정	95.57
D	항로 사정	0.82
Z	기타	0.59

- 전체 지연에서 C02(연결에 의한 지연)이 차지하는 비중이 90.58%로 압도적으로 높음.
- 지연 사유의 상위 5개 항목은 순위대로 C02(연결), C01(정비), A01(안개), C10(제방빙), D01(항로혼잡)임.
- 대부분의 지연은 항공기 사정에 의해 발생하며, 그 다음으로는 기상 사정에 의해 발생함.

<월별 지연율 탐색>

☞ SDT_MM (월)

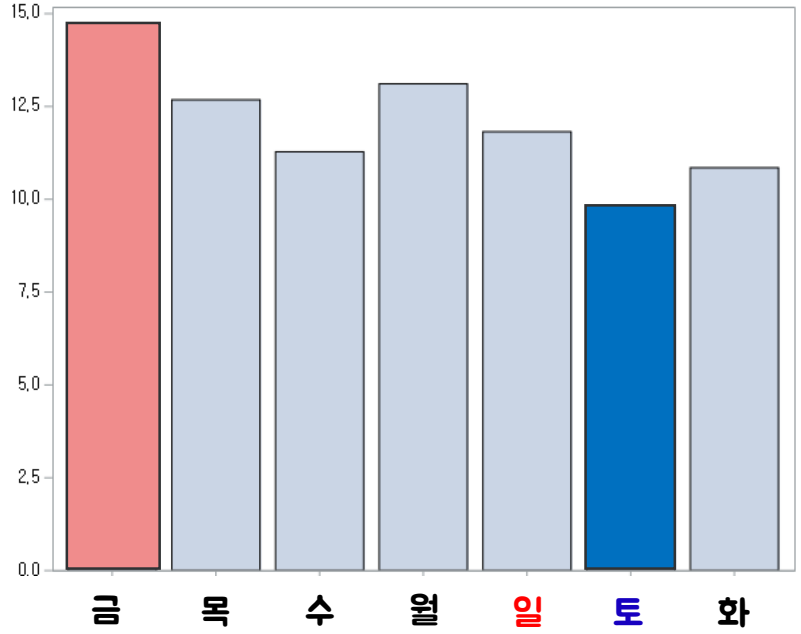


SDT_MM	총 운항 횟수	지연 빈도	지연율
1	96848	12952	13.37
2	88259	12384	14.03
3	98396	8486	8.62
4	98868	13313	13.47
5	103473	11598	11.21
6	100182	11100	11.08
7	68481	7954	11.61
8	68671	9267	13.49
9	67244	7711	11.47
10	68471	8656	12.64
11	63706	7736	12.14
12	65110	7780	11.95

- 3월의 지연율은 10% 미만으로 유일한 한 자리 수 지연율을 나타냄.
- Test Set의 대상인 9월의 지연율은 평균 이하 수준을 나타냄.
- 19년 Data의 경우, 6월까지만 존재하므로 7월~12월은 총 운항 횟수와 지연 빈도가 상대적으로 낮게 나타남.

〈요일별 지연율 탐색〉

☞ SDT_DY (요일)

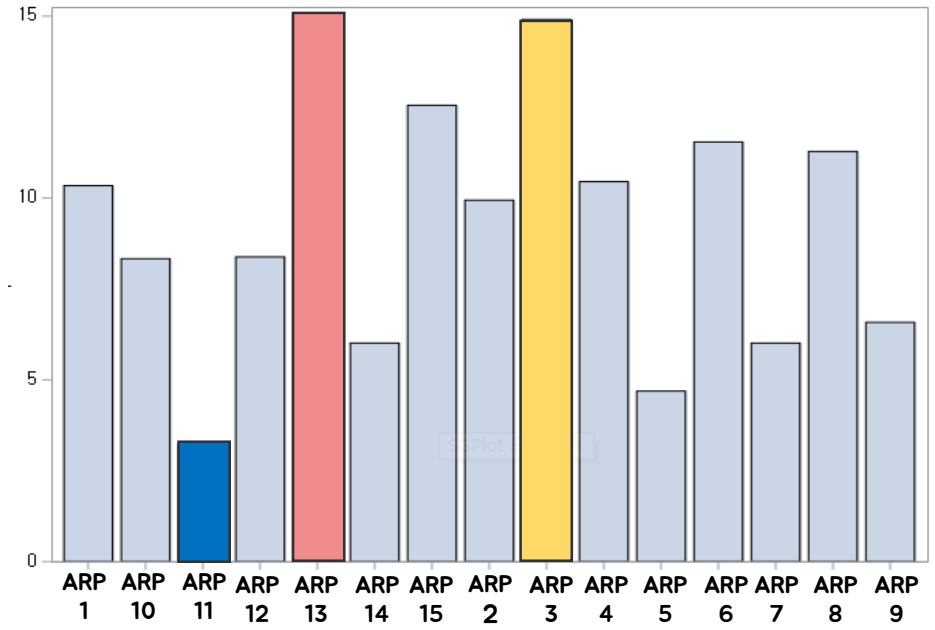


SDT_DY	총 운항 횟수	지연 빈도	지연율
금	142842	20993	14.70
목	139793	17726	12.68
수	139320	15712	11.28
월	141466	18540	13.11
일	143873	17001	11.82
토	141237	13867	9.82
화	139178	15098	10.85

- 상대적으로 평일에 비해 주말을 활용하여 여행을 가는 승객의 수가 많을 것이라는 판단 하에 **승객이 몰리는 주말의 지연 빈도가 평일보다 더 높을 것이라 추측**했음.
- 오히려 주말인 **토요일의 지연율이 가장 낮게** 나타나고, 일주일 중 **금요일의 지연율이 가장 높은** 것으로 나타남.

<공항별 지연율 탐색>

☞ ARP (공항)



ARP	총 운항 횟수	지연 빈도	지연율
ARP1	310665	32129	10.34
ARP10	12	1	8.33
ARP11	3658	119	3.25
ARP12	4880	409	8.38
ARP13	4248	640	15.07
ARP14	1897	114	6.01
ARP15	13696	1719	12.55
ARP2	121513	12081	9.94

ARP	총 운항 횟수	지연 빈도	지연율
ARP3	393607	58693	14.91
ARP4	33623	3513	10.45
ARP5	16321	765	4.69
ARP6	34472	3979	11.54
ARP7	3163	190	6.01
ARP8	33195	3746	11.28
ARP9	12759	839	6.58

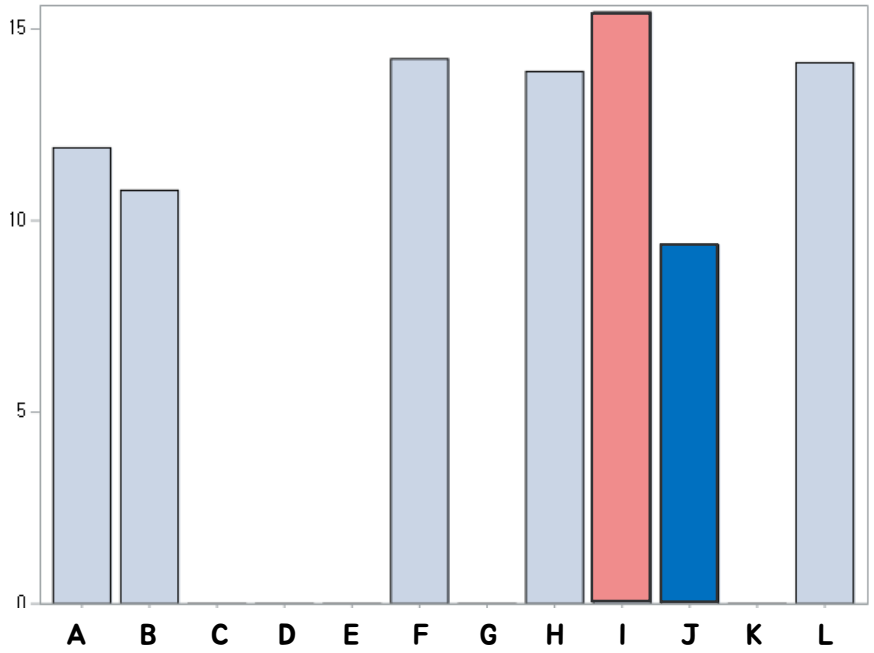
- 국내선 전체 운항의 70% 이상이 ARP1과 ARP3에서 이루어짐.
- ARP3의 경우, 운항 횟수가 가장 많은 공항이면서 동시에 지연율이 15%에 가까운 높은 수치를 기록함.

2. 데이터 준비 및 탐색

- 데이터 탐색

<항공사별 지연율 탐색>

☞ FLO (항공사)

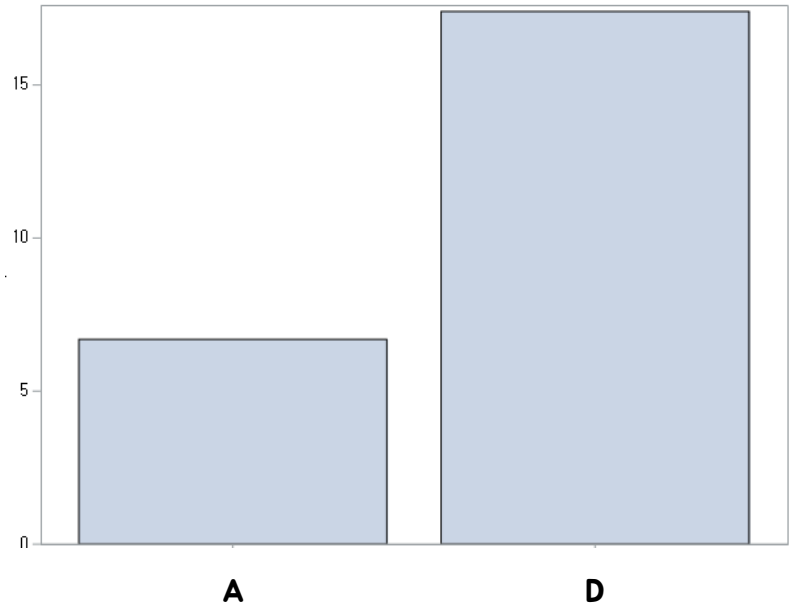


FLO	총 운항 횟수	지연 빈도	지연율
A	177187	21078	11.90
B	135235	14588	10.79
C	3	0	0
D	2	0	0
E	1	0	0
F	88110	12533	14.22
G	1	0	0
H	131935	18323	13.89
I	95074	14679	15.44
J	276447	25916	9.37
K	2	0	0
L	83712	11820	14.12

- 지연이 발생한 항공사 중 I항공사의 지연율이 가장 높고, J항공사의 지연율이 가장 낮음.
- 총 운항 횟수가 3번 이하인 항공사들은 지연이 한 번도 발생하지 않았음.

〈출도착별 지연율 탐색〉

☞ AOD (출도착)



AOD	총 운항 횟수	지연 빈도	지연율
A	493392	33010	6.69
D	493717	85927	17.4

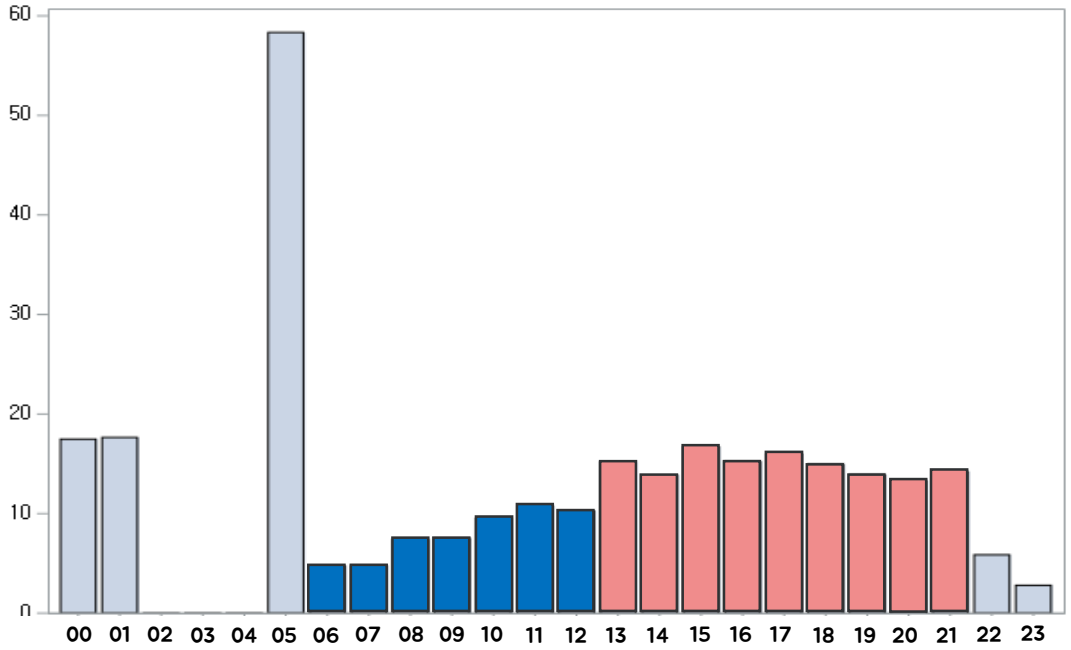
- 공항 기준 출발(D)한 항공편의 지연율이 도착(A)한 항공편의 두배 이상임.
- 공항 기준 도착편의 지연 여부는 해당 항공편의 출발시 지연 여부에 큰 영향을 받을 것으로 추측함.

2. 데이터 준비 및 탐색

- 데이터 탐색

〈계획시간별 지연율 탐색〉

☞ STT (계획시간) / 시간 단위 환산



STT	총 운항 횟수	지연 빈도	지연율
0	103	18	17.48
1	17	3	17.65
2	1	0	0
3	0	0	0
4	0	0	0
5	24	14	58.33
6	20154	958	4.75
7	49236	2349	4.77
8	65559	4835	7.38
9	71721	5387	7.51
10	64972	6174	9.50
11	65522	7055	10.77

STT	총 운항 횟수	지연 빈도	지연율
12	59669	6134	10.28
13	60219	9149	15.19
14	65257	9031	13.84
15	64536	10784	16.71
16	67543	10287	15.23
17	69960	11181	15.98
18	70202	10495	14.95
19	71119	9896	13.91
20	61210	8150	13.31
21	41755	5935	14.21
22	18680	1095	5.86
23	250	7	2.8

- 05시대에 계획되었던 비행편은 60%에 가까운 높은 지연율을 보임.
- 오전보다 **오후의 지연율이 상대적으로 높음.**
- **오후에 계획된 항공편**일수록 오전에 계획된 항공편보다 **연결지연의 발생 위험도가 높을 것**으로 추측함.

2. 데이터 준비 및 탐색

- 데이터 탐색

SFSNT.CSV

2019년 하계 스케줄 중
7월~9월이 포함된 시즌데이터

변수	형식	설명
SSC	CHAR	시즌코드
FLT	CHAR	편명
ORG	CHAR	공항
DES	CHAR	상대공항
STD	CHAR	출발시각
STA	CHAR	도착시각
FLO	CHAR	항공사
MON	CHAR	월
TUE	CHAR	화
WED	CHAR	수
THU	CHAR	목
FRI	CHAR	금
SAT	CHAR	토
SUN	CHAR	일
FSD	CHAR	시작일자
FED	CHAR	종료일자
IRR	CHAR	부정기편

- Test Set에 포함된 기간동안 운항하는 실제 항공편들의 세부 스케줄 확인 가능.
- 각 항공편별 정기 / 부정기 여부 확인 가능.

〈Test Set의 부정기편 유무〉

☞ IRR (부정기편)

SSC	FLT	ORG	DES	STD	STA	FLO	MON	TUE	WED	THU	FRI	SAT	SUN	FSD	FED	IRR
S19	B1806F	ARP4	ARP1	21:15	22:15	B		Y						20190603	20191026	Y
S19	B1853	ARP1	ARP4	7:30	8:25	B			Y					20190603	20191026	Y
S19	H1875	ARP6	ARP3	19:00	20:10	H	Y	Y	Y	Y	Y	Y	Y	20190503	20191026	Y
S19	H1876	ARP3	ARP6	17:15	18:25	H	Y	Y	Y	Y	Y	Y	Y	20190503	20191026	Y

- Test Set에 포함된 기간 중에 운항되는 **부정기 항공편**은 위의 **4개편 분**임을 확인함.
- 해당 항공편들을 Test Set에 조회해본 결과, **Test Set에 포함되어 있지 않음**을 확인함.
- Test Set에 부정기편이 존재하지 않으므로 **Train Set에 있는 부정기편 Data**들을 **제거**하기로 결정함.

AFSNT_DLY.CSV

2019.9.16 ~ 2019.9.30 까지의 데이터
(Test Set)

변수	형식	설명
SDT_YY	CHAR	연
SDT_MM	CHAR	월
SDT_DD	CHAR	일
SDT_DY	CHAR	요일
ARP	CHAR	공항
ODP	CHAR	상대공항
FLO	CHAR	항공사
FLT	CHAR	편명
AOD	CHAR	출도착
STT	CHAR	계획시각
DLY	CHAR	타겟변수 / 지연여부
DLY_RATE	NUM	타겟변수 / 지연확률

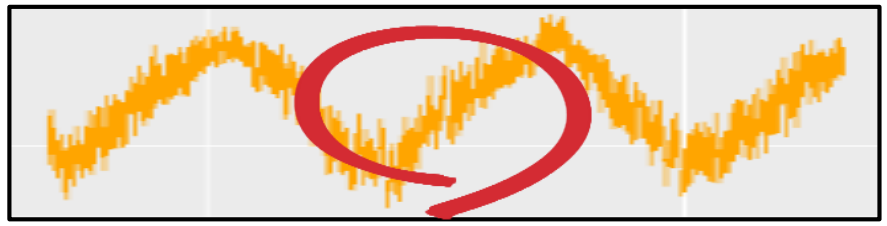
- 타겟의 분류를 위한 **독립변수가 부족**하다는 판단 하에, **파생변수**를 만들기로 결정함.

날씨 관련 데이터

항공기상청과 기상자료개방포털에서 수집한
과거 날씨 데이터

〈날씨 예측의 Risk〉

☞ 시간별 정확한 수치 예측이 사실상 불가능함



〈TEMP〉

〈WS〉

- Test Set 기간에 해당되는 기온, 풍속 등의 날씨 변수를 예측하기 위한 모델을 여러 번 구축함.
- 구축한 모델의 예측값을 실제값과 비교했을 때, **실제 값과 작지 않은 오차**가 확인됨.
- 실제 값과의 오차를 줄이기 위해서 주기성을 나타내는 날씨 변수만을 사용하고, 주기성이 없는 날씨 변수는 과감히 **독립변수에서 제거**하기로 결정함.

2.3 비식별화 데이터 유추

〈비식별화된 공항 정보 유추〉

☞ ARP (공항)

KAC 한국공항공사 KOREA AIRPORTS CORPORATION

공항명	운항(편수)		
	도착	출발	계
김포	29,070	29,109	58,179
합계	29,070	29,109	58,179

ARP	빈도	실제 공항명
ARP1	58179	김포공항
ARP10	0	양양공항
ARP11	684	포항공항
ARP12	970	사천공항
ARP13	884	군산공항
ARP14	425	원주공항
ARP15	2728	인천공항
ARP2	23043	김해공항
ARP3	76032	제주공항
ARP4	6458	대구공항
ARP5	3334	울산공항
ARP6	6929	청주공항
ARP7	720	무안공항
ARP8	6637	광주공항
ARP9	2485	여수공항

- Train Set의 **ARP별 빈도수**를 한국공항공사 홈페이지에 게시되어 있는 공항별 **실제 운항 횟수와 비교**하여 비식별화 되어있는 공항 정보를 유추함.

<비식별화된 항공사 정보 유추>

☞ FLO (항공사)

KAC 한국공항공사 KOREA AIRPORTS CORPORATION

항공사	항공사명	운항(편수)		
		도착	출발	계
JJA	제주항공	12,515	12,515	25,030

FLO	빈도	실제 항공사명
A	33564	아시아나항공
B	26844	에어부산
C	0	-
D	0	-
E	0	-
F	16288	이스타항공
G	0	-
H	25030	제주항공
I	18066	진에어
J	53194	대한항공
K	0	-
L	16522	티웨이항공

- Train Set의 FLO별 빈도수를 한국공항공사 홈페이지에 게시되어 있는 항공사별 실제 운항 횟수와 비교하여 비식별화 되어있는 항공사 정보를 유추함.

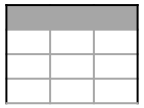
3. 데이터 전처리

파생변수 생성 / 이상치 처리 / 데이터 분할

3.1 파생변수 생성

3. 데이터 전처리

- 파생변수 생성



AFSNT

- 기존에 Train Set으로 제공된 AFSNT.CSV 데이터셋을 활용하여 계절, 분 단위 환산 값 등의 시간 파생변수를 생성하고, 각각의 데이터를 일 단위로 추출하여 일자별 운항 정보를 담고 있는 파생변수를 생성함.

1. 기존의 변수로부터 생성한 파생변수

☞ 시간 변수

변수	형식	설명
Season	CHAR	계절
Yy_dly_rate	NUM	연도별 지연율
Mm_dly_rate	NUM	월별 지연율
Dy_dly_rate	NUM	요일별 지연율
Time	NUM	STT를 분 단위로 환산한 값
Hour	NUM	STT의 시간 단위 (24시간)

- 계절과 같은 특정 시간이 지연에 영향을 미칠 것이라는 판단하에 SDT_MM (월), SDT_DY(요일) 변수 등을 활용하여 **시간 정보를 담고 있는 파생변수**를 생성함.
- 기존에 13:35와 같은 형식으로 시간을 나타내고 있던 데이터를 시간 단위, 분 단위 값으로 환산하여 **수치적인 크기 비교**가 가능하도록 변형시킨 파생변수를 생성함.

☞ 하루 운항 변수

변수	형식	설명
Haru_arp_odp_cnt	NUM	ARP & ODP 기준 하루 운항 횟수
Haru_arp_cnt	NUM	ARP 기준 하루 운항 횟수
Haru_cnt	NUM	국내선 하루 총 운항 횟수
Haru_aod_arp_odp	NUM	AOD별 ARP & ODP 기준 하루 운항 횟수
Haru_aod_arp	NUM	AOD별 ARP 기준 하루 운항 횟수
Haru_aod	NUM	AOD별 하루 운항 횟수
Seq	NUM	해당일의 ARP & ODP 기준 운항 순번

- 지연 사유의 가장 큰 비중을 차지하는 **CO2(연결지연)**은 앞선 항공편들의 영향을 받는 변수이므로, **하루의 운항 횟수가 많을수록, 상대적으로 나중 순서에 운항하는 항공편일수록** 지연에 영향을 더 받을 것이라고 판단함.
- 전체 데이터를 각각의 **일 단위로 추출**하여서 위와 같은 정보를 담고 있는 파생변수들을 생성함.

3. 데이터 전처리

- 파생변수 생성

KAC 한국공항공사 KOREA AIRPORTS CORPORATION

- 공항의 부지 면적이 넓을수록 동시에 수용할 수 있는 항공기의 수가 많을 것이며, 활주로의 길이가 길수록 활주로 사정에 의한 지연이 발생할 경우가 많을 것이라는 추측에 의해 한국공항공사로부터 데이터를 추출해 파생변수를 생성함.

1. 새로 수집한 데이터로부터 생성한 파생변수

☞ 공항 변수

구분 \ 공항명	김포공항	김해공항	제주공항	대구공항	…
부지면적 (m^2)	8,440,923	3,697,435	3,561,679	171,308	…
활주로 길이 (m)	3,400	2,972	3,180	2,749	…

〈출처 : 한국공항공사 전국공항 시설현황〉

변수	형식	설명
AA	NUM	ARP 공항 면적
LOA	NUM	ARP 공항 활주로 길이

Google

- 지연의 빈도수는 상대적으로 운항 횟수가 많은 대형 항공사들이 더 많은 것으로 보이나, 각 항공사별 지연율을 살펴보면 높은 지연율을 차지하는 항공사는 대부분이 저가 항공사임. 따라서 항공사 규모를 나타내는 파생변수를 생성함.

1. 새로 수집한 데이터로부터 생성한 파생변수

☞ 항공사 변수

변수	형식	설명
LCC	CHAR	항공사 규모 (0: 대형 항공사, 1: 저가 항공사)
Flo_dly_rate	NUM	항공사별 지연율

3. 데이터 전처리

- 파생변수 생성

GREAT CIRCLE MAPPER

- 공항-공항 사이의 거리가 멀수록 즉, 장거리 노선일수록 항공편의 운항 지연율이 높게 나타난다는 논문 결과를 참고하여 공항간 거리를 나타내는 파생변수를 생성함.

1. 새로 수집한 데이터로부터 생성한 파생변수

☞ 거리 변수

목적지별 운항지연율(delay), 연도별 운항지연율(delay), 운항거리별 운항지연율(delay)의 평균을 비교한 결과, 모두 장거리 노선의 운항지연율(delay)이 높게 나타났다. 그러므로 운항지연율(delay)은 단거리 노선과 장거리 노선으로 구분되는 운항거리에 따라 차이가 없을 것이라는 분석 내용 3-1은 기각하였다.

〈출처 : 항공기 운항정시성에 대한 실증 연구 논문〉

Distances

From	To	Initial Heading	Magnetic Heading	Distance
GMP	CJU	184° (S)	192° (S)	280 mi

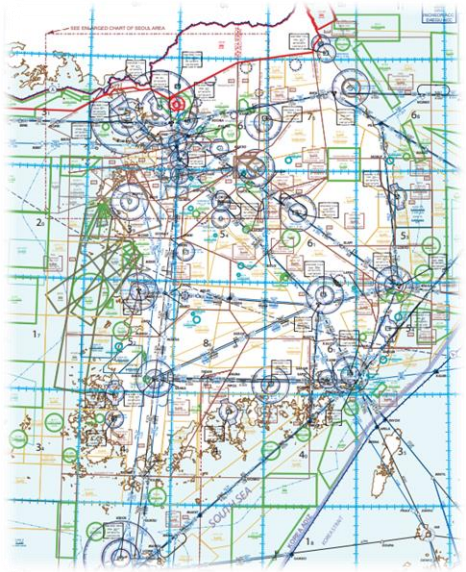
변수	형식	설명
Dist_km_	NUM	공항 간 거리

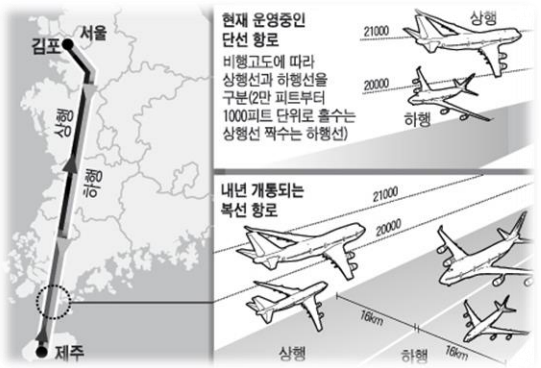
AIM 항공정보관리체계

- 항로에는 상행선과 하행선이 비행고도에 따라 구분되는 단선항로와, 서로 다른 항로를 사용하는 복선항로가 있음. 단선항로일수록 동시간대에 항로가 혼잡할 가능성이 높다는 판단 하에 항로의 종류를 나타내는 파생변수를 생성함.

1. 새로 수집한 데이터로부터 생성한 파생변수

☞ 항로 변수





변수	형식	설명
Flight_road	CHAR	항로 종류 (0: 단선항로, 1: 복선항로)

3. 데이터 전처리

- 파생변수 생성

날씨데이터

- 데이터 탐색 과정에서 항공기 지연 사유의 2위가 기상에 의한 지연임을 확인함. 그 외에도 각 종 논문이나 뉴스 기사에서도 항공기 지연이 기상에 영향을 받음을 쉽게 확인할 수 있음. 따라서 기상 정보를 나타내는 파생변수를 생성함.

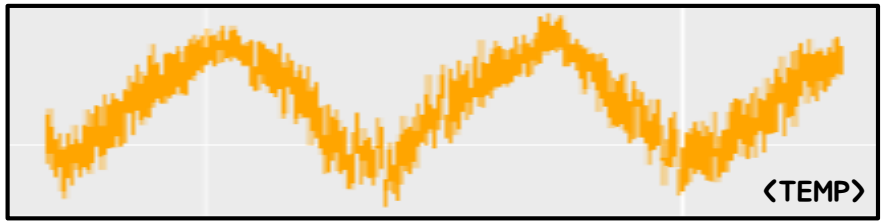
1. 새로 수집한 데이터로부터 생성한 파생변수

☞ 날씨 변수

발생하였다. 특히, 가장 큰 비중을 차지하는 A/C접속이 기상현상과 관련이 있음을 알 수 있어 기상현상이 중요한 원인이 됨을 알 수 있다.

〈출처 : 항공기 운항정시성에 대한 실증 연구 논문〉

- 기상은 단순히 기상에 의한 지연에만 그 영향을 미칠 뿐 아니라, 전체 지연 사유 중 가장 큰 비중을 차지하는 CO2(연결지연)과도 관련 되어 있음을 논문을 통해 확인함.



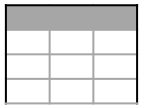
- 주기성을 가지는 기상 변수라고 해도 예측 모델을 구축했을 때 실제 값과의 오차는 필연적으로 발생함.
- 연도가 바뀌어도 같은 주기성을 가지므로, 동일한 월/일/시간의 과거 기상 데이터의 평균 값을 Test Set에 활용함.

변수	형식	설명
TEMP	NUM	기온
DP	NUM	이슬점 온도
PRESS	NUM	기압

3.2 이상치 처리

3. 데이터 전처리

- 이상치 처리



AFSNT

- 분석에 있어 불필요하다고 판단되는 데이터와 Raw Data 생성 과정에서 오기입 된 것으로 판단되는 데이터, 논리적으로 오류가 발견된 데이터를 수정 및 제거함.

1. 분석에 불필요한 데이터 제거

☞ CNL(결항여부), CNR(결항사유)

> 결항된 항공편은 대회 문제에서 제외됩니다.
>
> - 2019 빅콘테스트 FAQ 발췌 -

- 결항은 실제로 운행하지 않은 비행편이기 때문에 지연에 대한 Train을 할 때 불필요할 뿐더러, Test Set에서도 결항된 항공편은 제외 되기 때문에, **결항된 데이터를 제거함.**

☞ IRR(부정기편)

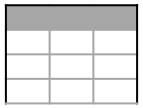
> 조기 출발편은 실제시간이 계획시간보다 10분 내외로 조기 출발 할 수 있지만, 시간 차이가 많이 발생하는 경우는 부정기 페리편에 해당합니다.
>
> - 2019 빅콘테스트 FAQ 발췌 -

- IRR(부정기편)은 N으로 기입되어 있으나, 실제 출발시간이 계획출발시간보다 16분 이상 빠른 데이터가 발견됨.
- 조기 출발한 시간이 15분 이하까지는 정상적인 정기 조기출발편으로 판단하고, **16분 이상일 경우 부정기 페리편으로 판단하여 분석에서 제외함.**

- SFSNT와 AFSNT_DLY 데이터셋 탐색 과정에서 알 수 있었듯이 Test Set에는 부정기편이 존재하지 않으므로 부정기편에 대한 불필요한 학습을 방지하고자 **부정기편 데이터를 제거함.**

3. 데이터 전처리

- 이상치 처리



AFSNT

- 분석에 있어 불필요하다고 판단되는 데이터와 Raw Data 생성 과정에서 오기입 된 것으로 판단되는 데이터, 논리적으로 오류가 발견된 데이터를 수정 및 제거함.

2. Raw Data 생성 과정에서 오기입 된 것으로 판단되는 데이터 수정

☞ REG(등록기호) 불일치

SDT_YY	SDT_MM	SDT_DD	ARP	ODP	FLT	REG	AOD
2018	4	20	ARP1	ARP9	A1735	SEw3MjQ3	D
2018	4	20	ARP9	ARP1	A1735	SEw3Nzcy	A

- 짝지어진 한 쌍의 출발편 및 도착편 데이터임에도 불구하고 REG(등록기호)가 서로 **불일치**하는 데이터를 발견함.
- **출도착 편의 REG 값이 같아지도록** 데이터를 수정하여 재기입함.

3. 논리적 오류가 발견된 데이터 제거

☞ ATT(실제시간)이 자정을 넘었을 때의 논리적 오류

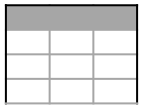
SDT_YY	SDT_MM	SDT_DD	ARP	ODP	STT	ATT	DLY
2017	10	2	ARP3	ARP6	20:25	00:05	N
2018	4	11	ARP6	ARP3	21:40	00:14	N

- 계획시간(STT)보다 실제시간(ATT)이 30분 넘게 늦었음에도 불구하고, **ATT가 자정을 넘으면** DLY가 **N으로 기입되는 오류**를 발견함.
- 해당 데이터들의 DLY를 **Y로 수정**하여 재기입함.

☞ 도착시간이 출발시간보다 빠른 논리적 오류

- 도착 계획시간이 출발 계획시간보다 빠른 데이터를 발견함.
- 도착 실제시간이 출발 실제시간보다 빠른 데이터를 발견함.
- **논리적으로 불가능한 경우이므로 해당 데이터들을 분석에서 제거함.**

3.3 데이터 분할



AFSNT

- 분류 모형을 만들기 위해 전처리된 Train Set을 자체적으로 다시 한 번 Train Set과 Test Set으로 분할함.
- 이 때 생성된 Test Set은 대회의 실제 분석 Test Set인 AFSNT_DLY.CSV의 특징을 반영하고 있어야함.

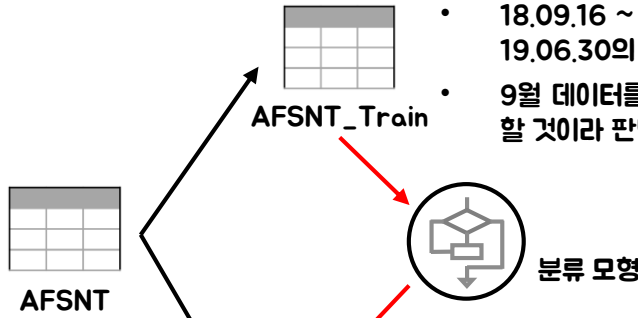
AFSNT_DLY.CSV

2019.9.16 ~ 2019.9.30 까지의 데이터
(Test Set)

1. AFSNT_DLY.CSV의 특징

- 2019년 9월 16일 ~ 9월 30일까지의 데이터만을 가지고 있음.

2. 자체적으로 분할한 Test Set을 해당 기간의 데이터로 구성



- 18.09.16 ~ 18.09.30의 데이터를 제외한 17.01.01 ~ 19.06.30의 데이터로 구성.
- 9월 데이터를 학습시키는 것이 9월 데이터 분류에 유리할 것이라 판단하여 17년 9월 데이터를 학습에 포함.

- 18.09.16 ~ 18.09.30의 데이터로 구성.
- 적합된 분류 모형이 Test에서 좋은 성능을 보인다면, 실제 Test Set의 데이터인 19.09.16 ~ 19.09.30에 대해서도 좋은 성능의 분류를 할 것이라 판단.

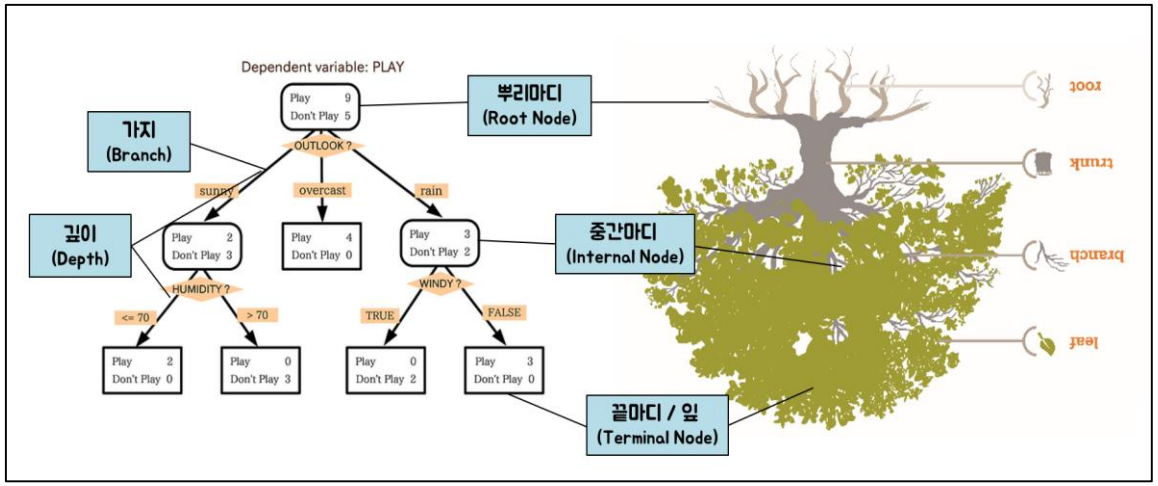
4. 모델링 및 결론

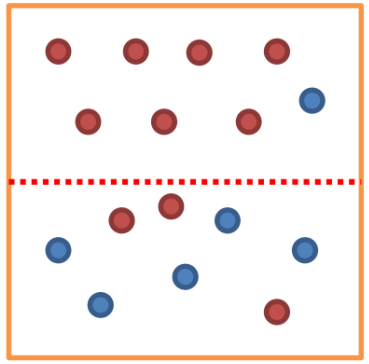
사용 모델 / 모델 비교 및 평가 / 최종 모델 선정

4.1 사용 모델

DECISION TREE

- 데이터를 분석하여 그들 사이의 패턴과 결과를 트리 구조로 모형화하여 나타내는 분류 및 예측 모형.
- 한 번에 하나씩의 설명변수를 사용하여 예측 가능한 규칙들의 집합을 생성하는 알고리즘.

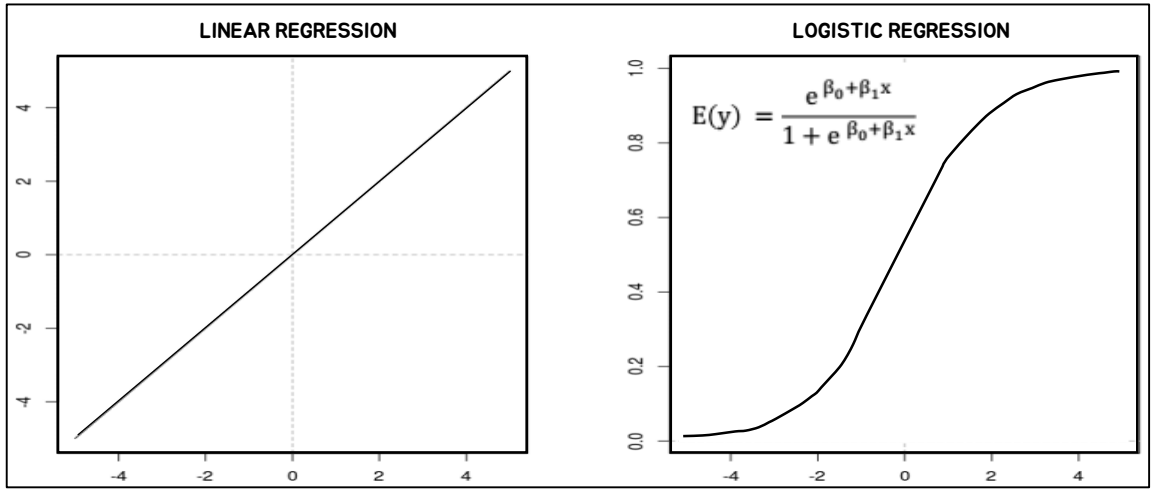




- 분리의 결과로 생기는 직사각형 내부의 **불순도를 감소시키는 방향**으로 분리를 진행함.
- 불순도 측정의 지표로는 지니지수, 엔트로피, 카이제곱 통계량 등이 사용됨.

LOGISTIC REGRESSION

- 선형회귀분석 모델과 동일하게 종속변수와 독립변수 간의 관계를 표현하지만 특히 범주형 종속변수를 다루는 분류 모형.
- 오즈비를 통해 각각의 요인이 결과에 유의미한 영향을 미치는지 여부를 파악할 수 있음.



$$Odds = \frac{P(A)}{1 - P(A)}$$

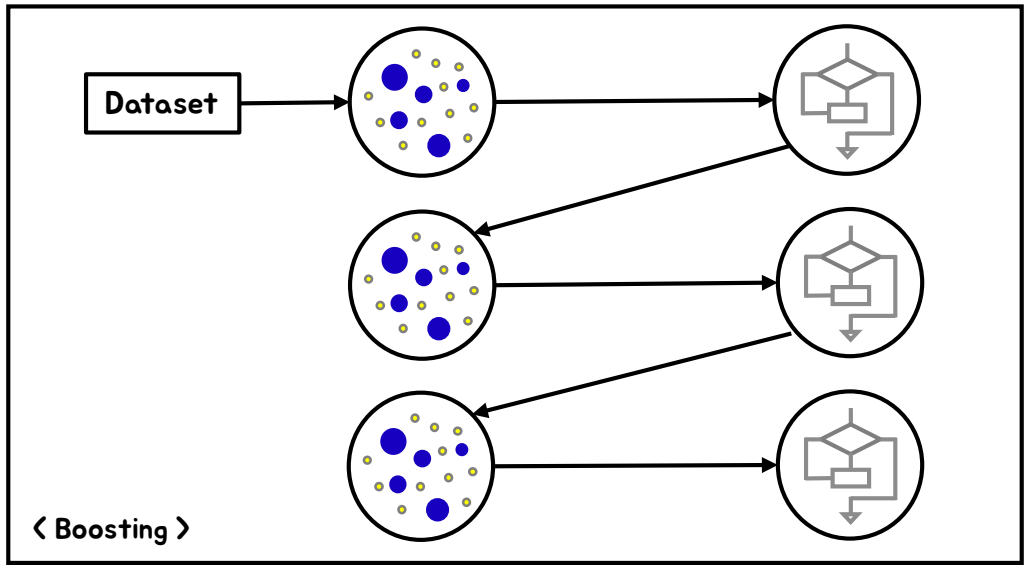
- 오즈(Odds)란, **임의의 사건 A가 발생하지 않을 확률 대비 일어날 확률**임. 즉, 오즈는 항상 음이 아닌 값을 갖고, P(A)가 1에 가까울수록 오즈의 값은 커짐.

4. 모델링 및 결론

- 사용 모델

Gradient Boosting

- 손실함수(Loss Function)을 파라미터로 미분해서 기울기를 구하고, 값이 작아지는 방향으로 파라미터를 조절.
- Gradient가 드러낸 현재 모델의 약점을 Boosting 기법을 통해 보완해 나가는 모형.



- Train Set에서 분류기를 생성하고, 분류 결과를 통해 오분류된 데이터에 대해 가중치를 부여하여 다음 학습에 순차적으로 이용함.
- Gradient Boosting 모형은 경사하강법을 통해 잔차를 계속해서 줄어들게 만들며 예측에 대해 엄청난 성능을 보이며, 머신러닝 알고리즘 중 가장 예측 성능이 높다고 알려진 모형임.

1. 일반적인 GBM 모형의 한계

- GBM은 계산량이 상당히 많이 필요한 알고리즘임.
- 따라서 모델의 실행 시간이 상당히 오래 걸린다는 단점이 있음.
- 이를 효율적으로 구현하기 위해 LightGBM, XGBoost 등의 모형 탄생.

2. Gradient Boosting의 확장된 모형

LightGBM

- Light해서 적은 메모리를 사용하며 고속으로 대용량 데이터를 처리함.
- 정확성에도 초점을 맞추어 가장 성공적인 알고리즘으로 꼽힘.
- 1만줄(rows) 이하의 작은 데이터에서는 Overfitting 될 가능성 높음.

XGBoost

- XGBoost는 속도와 모델 성능에 초점을 맞춘 GBM 모형임.
- Tree를 구성할 때 병렬처리기법을 사용해서 알고리즘 시간을 단축함.

4.2 모델 비교 및 평가

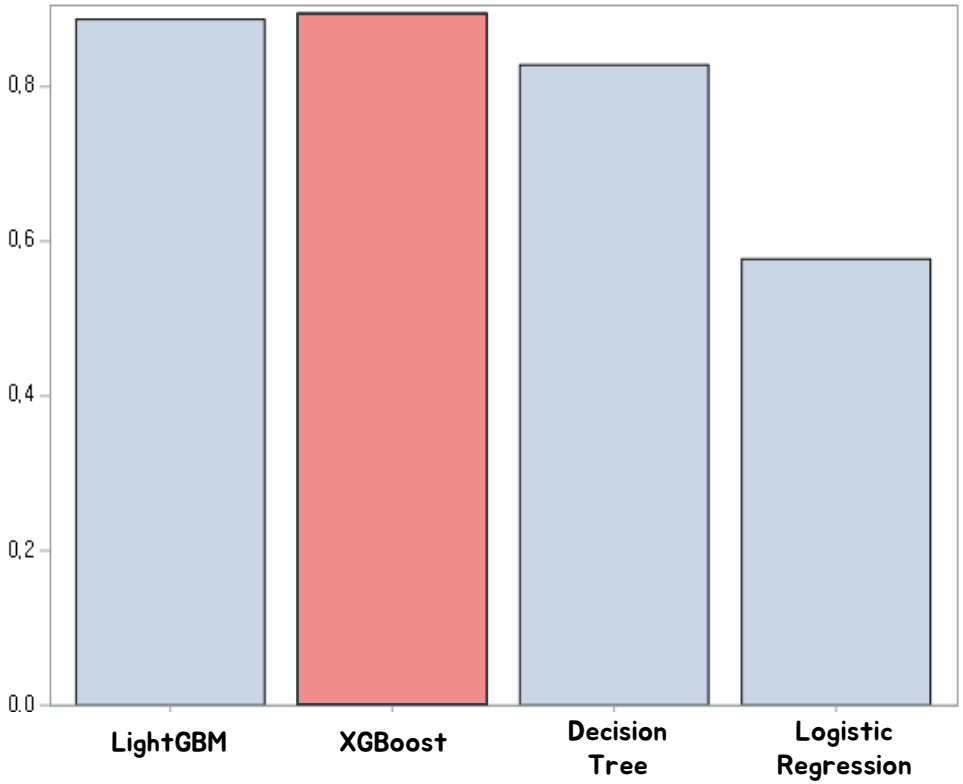
4. 모델링 및 결론

- 모델 비교 및 평가

항공기 지연여부 예측

- LightGBM, XGBoost, Decision Tree, Logistic Regression 의 4개의 서로 다른 분류 모형을 파라미터를 조절하며 적합시켜서 각각의 성능을 5가지 지표에 의해 비교 및 분석.

1. AFSNT_Train



☞ 모델별 성능 비교표

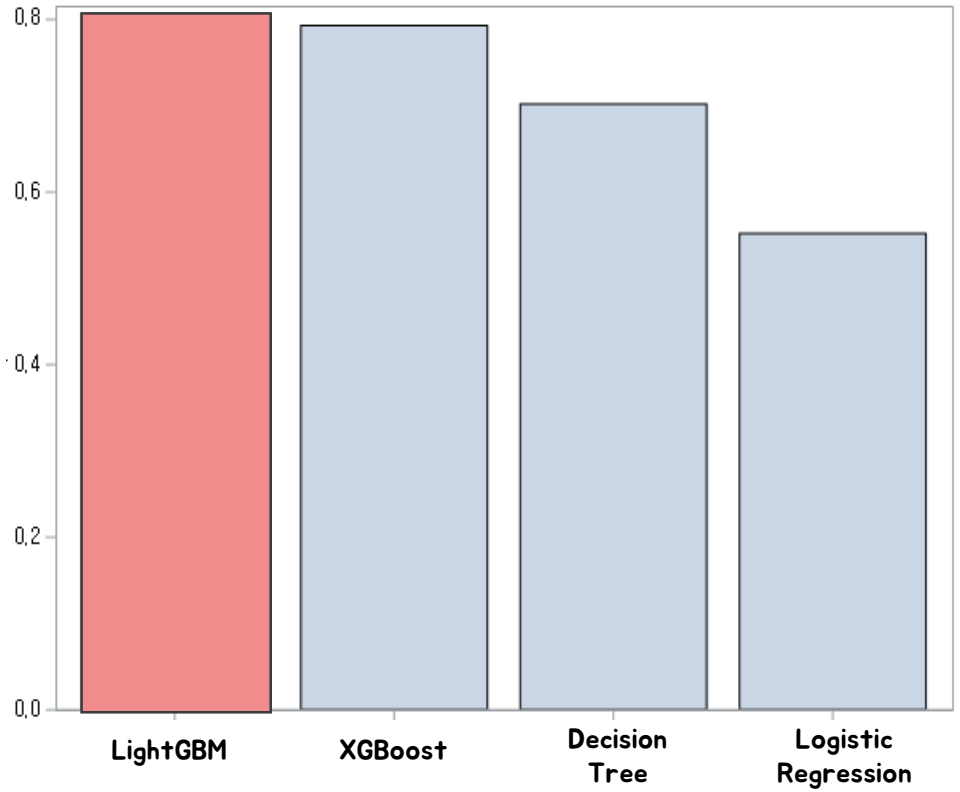
모델	Accuracy	AUROC	Recall	Precision	F1 score
LightGBM	0.808	0.887	0.795	0.369	0.504
XGBoost	0.790	0.895	0.836	0.350	0.493
Decision Tree	0.735	0.824	0.730	0.738	0.734
Logistic Regression	0.577	0.577	0.598	0.573	0.595

- Train Set에 대해서는 XGBoost가 가장 높은 AUROC 값을 보임.
- Accuracy적인 측면에서는 LightGBM이 더 높은 값을 보임.
- Decision Tree를 비롯한 Tree 기반 모델은 모두 0.8 이상의 높은 AUROC 값을 보이는 반면, Logistic Regression 모델은 0.5대의 낮은 값을 보임.

항공기 지연여부 예측

- LightGBM, XGBoost, Decision Tree, Logistic Regression 의 4개의 서로 다른 분류 모형을 파라미터를 조절하며 적합시켜서 각각의 성능을 5가지 지표에 의해 비교 및 분석.

2. AFSNT_Test



☞ 모델별 성능 비교표

모델	Accuracy	AUROC	Recall	Precision	F1 score
LightGBM	0.758	0.806	0.701	0.228	0.344
XGBoost	0.726	0.793	0.705	0.205	0.318
Decision Tree	0.691	0.704	0.611	0.168	0.264
Logistic Regression	0.628	0.552	0.459	0.114	0.183

- 그러나 Test Set에 대해서는 AUROC와 Accuracy적인 두 가지 측면 모두에서 LightGBM이 가장 높은 값을 보임.
- Gradient Boosting 기반 모델에서, Cut-off 조절을 통해 Accuracy 값을 더 높일 수도 있으나, 이번 문제는 단순히 Accuracy의 값을 높이는 것 보다는 지연의 경우를 예측하는 것이 더 중요한 사항이므로 Recall 값이 최소 0.7 이상이 나오는 선에서 파라미터 조절을 진행함.

4.3 최종 모델 선정

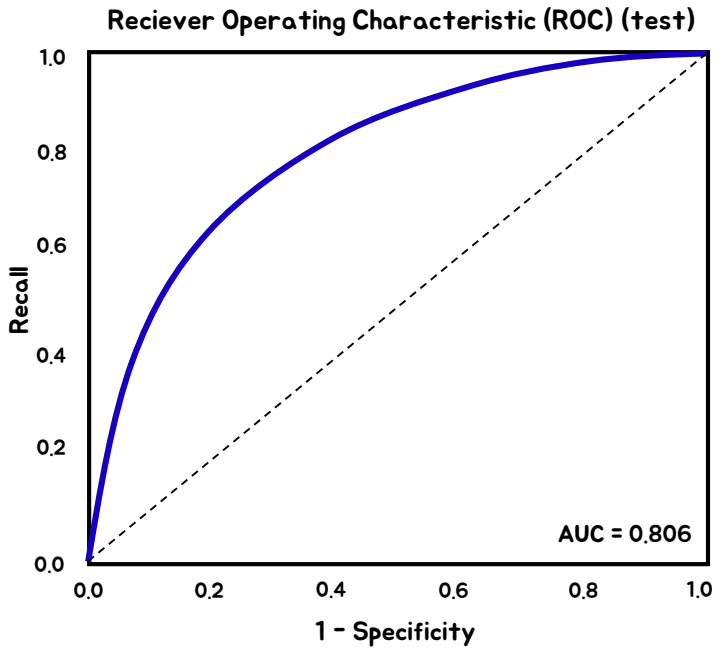
4. 모델링 및 결론

- 최종 모델 선정

LightGBM

- LightGBM 모델이 9월 16일 ~ 30일의 데이터로 이루어진 Test Set에 대해서 가장 높은 Accuracy 값과 AUROC 값을 나타냄.
- 해당 모델은 위와 동시에 0.7 이상의 Recall 값을 보임으로서 10편 중 7편의 지연을 맞추는 성능을 보임.

모델	Accuracy	AUROC	Recall	Precision	F1 score
LightGBM	0.758	0.806	0.701	0.228	0.344



Parameters

- Max_depth : 16
- Num_leaves : 350
- Learning_rate : 0.05
- Max_bin : 95
- Bagging_fraction : 0.85
- Feature_fraction : 0.90

⋮

- Min_child_weight : 0.001
- Min_child_samples : 20

2. LightGBM 변수중요도 (상위 10개 항목)

변수명	변수 설명	변수 중요도
Haru_cnt	국내선 하루 총 운항 횟수	1
Time	STT를 분 단위로 환산한 값	0.768
PRESS	기압	0.736
DP	이슬점 온도	0.69
TEMP	기온	0.689
Haru_aod_arp_odp	AOD별 ARP & ODP 기준 하루 총 운항 횟수	0.558
Seq	해당일의 ARP & ODP 기준 운항 순번	0.492
Haru_arp_cnt	ARP 기준 하루 총 운항 횟수	0.476
Haru_aod_arp	AOD별 ARP 기준 하루 총 운항 횟수	0.392
Dist_km_	공항 간 거리	0.179

- 하루 운항 정보를 담고 있는 변수들이 5개, 기상 변수 3개, 시간 정보를 담고 있는 변수 1개, 항로 변수 1개가 선택되었음.
- 해당 일자에 총 몇 편의 항공편이 운항되는지가 지연에 중요한 영향을 미친다는 점을 유추할 수 있음. (연결 지연과 밀접한 관계가 있을 것으로 추측)

5. 활용방안 및 기대효과

활용방안 / 기대효과

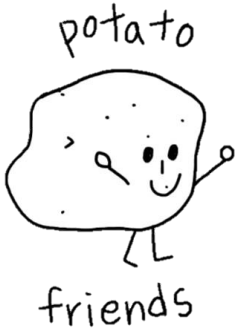
5.1 활용방안

5.2 기대효과

APPENDIX

사용한 분석툴

- 각 분석 단계의 목적에 맞게 가장 효율적인 분석툴을 선정해 사용함.
- 전처리 단계 : SAS
- 모델링 단계 : Python, R



감자마을친구들


태연


성곤


건욱


규선


성찬