항공지연 예측

에어플레인모드 팀

황종수 손희헌 이재원 이소정

목차

1. 탐색적 자료분석

1 탐색적 자료분석

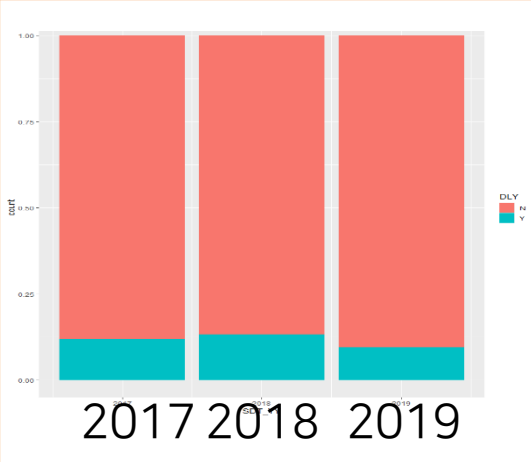
2 전처리 및 변수생성과정

3 모형 비교

4 최종 모형

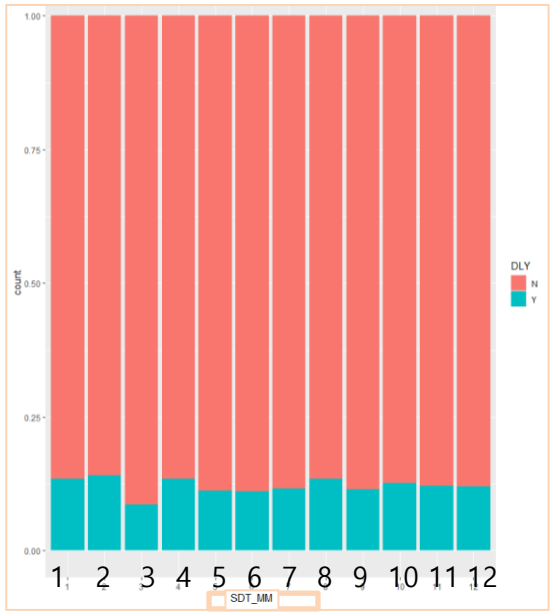
1. 탐색적 자료분석

연도별 지연율



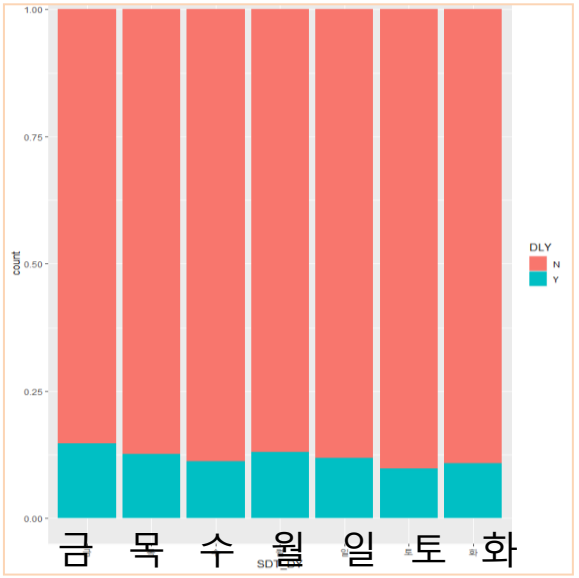
2018년이 지연율이 가장 높았다. 하지만 "2019년 데이터는 7월을 포함한 이후의 데이터는 없음"을 인지 해야한다.

월별 지연율



3개년 평균으로 봤을 때 3월의 지연율이 눈에 띄게 낮았다. 우리가 예측해야하는 9월달은 3월을 제외하고 가장 낮았다.

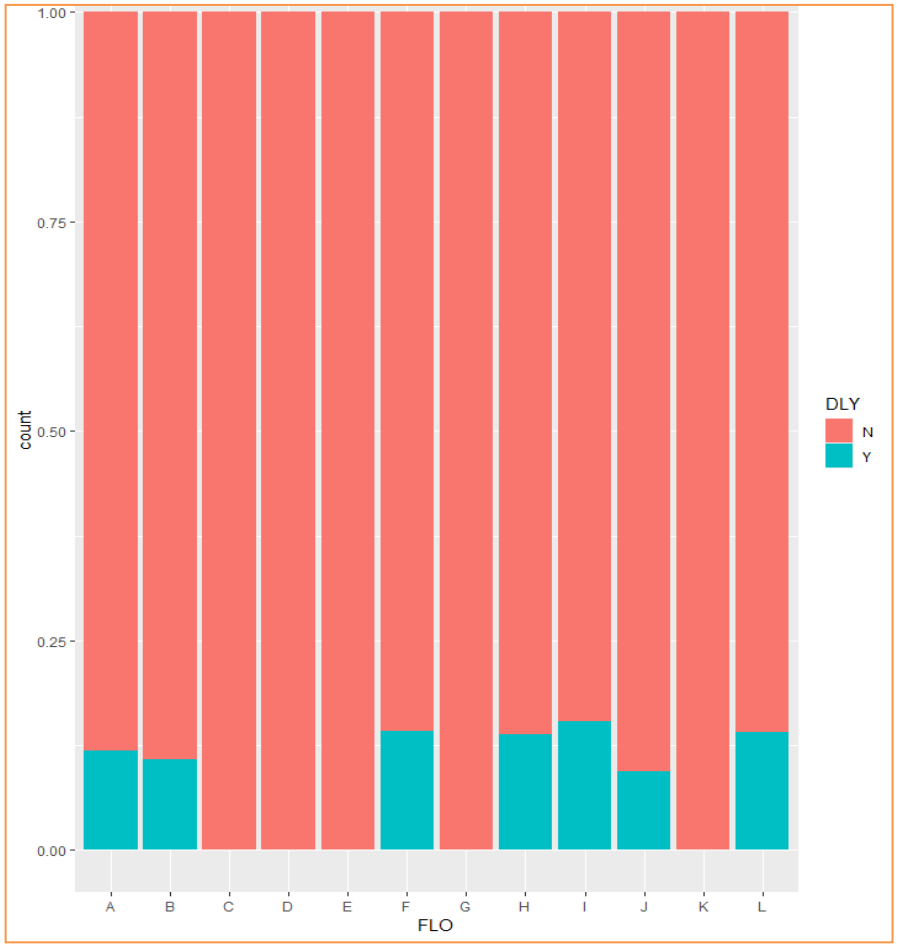
요일별 지연율



금요일의 지연율이 가장 높은 반면 수요일과 토요일에 출발하는 지연율이 유의하게 낮았다.

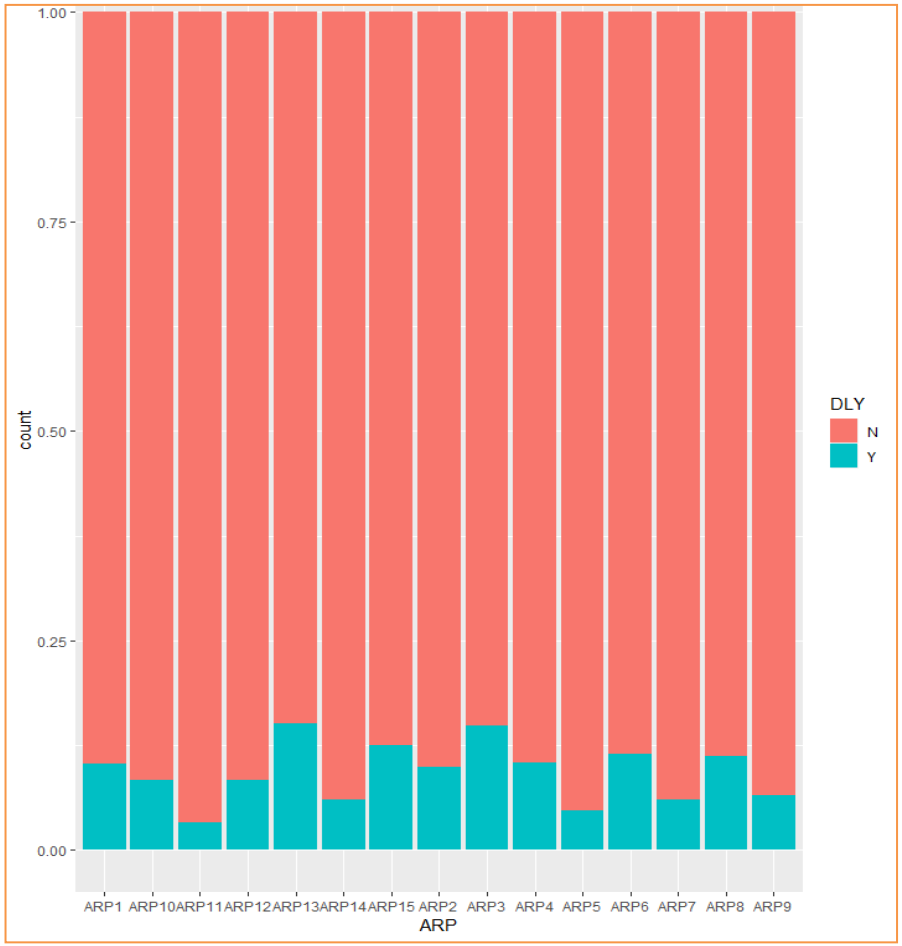
1. 탐색적 자료분석

항공사별 지연율



12개 항공사 중에서 7개 항공사가 지연율이 존재함을 보여 주고 있으며, 그 정도가 모두 다름을 알 수있다.

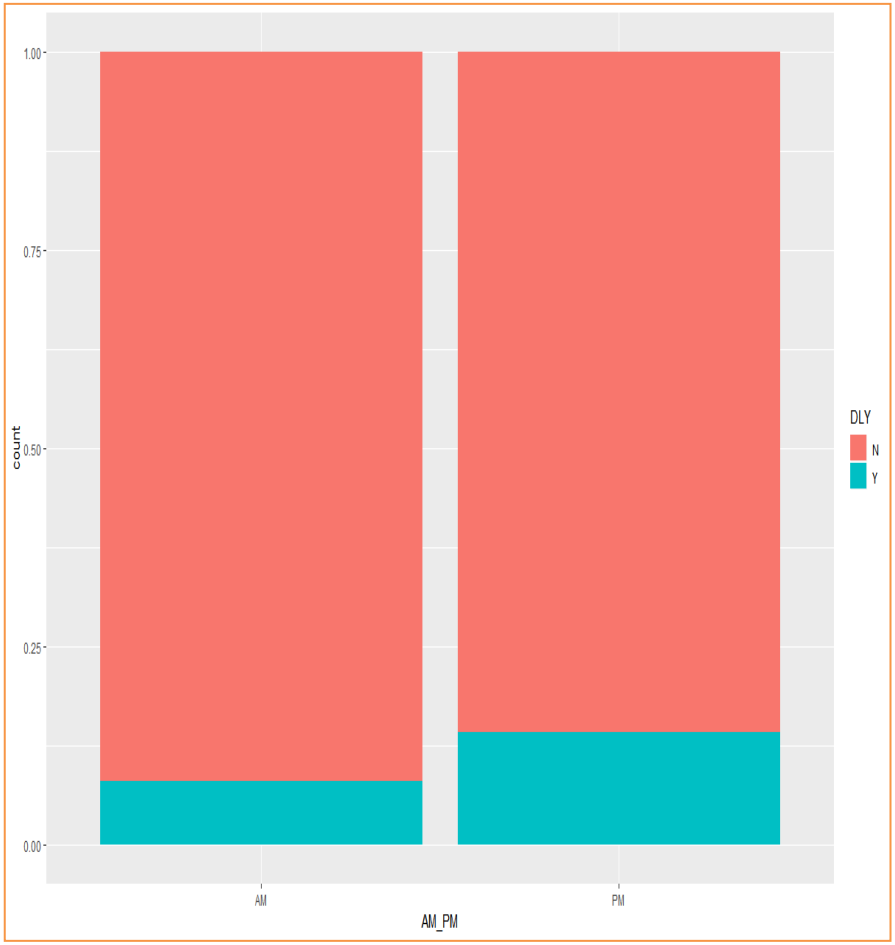
공항별 지연율



공항별 지연율은 각각의 공항마다 다른 지연율을 나타내고있다. 이를 파생변수를 만드는데 이용한다면, 예측에 유의함을 기대 할 수있다.

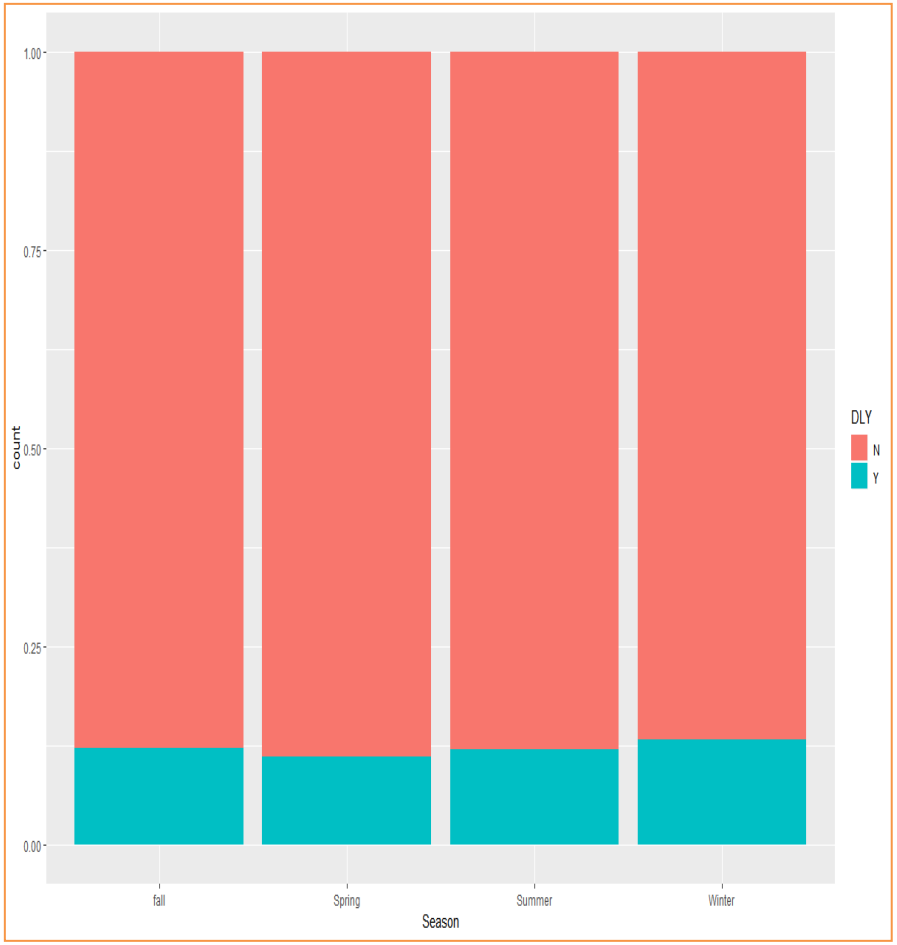
1. 탐색적 자료분석

오전/오후 지연율



낮 12시를 기점으로 오전 오후로 나뉘어 지연 비율을 보았을 때,오후의 지연율이 비교적 큼을 나타낸다.

계절별 지연율



봄의 지연율이 낮은 편이고 겨울의 지연율이 상대적으로 높은 편임을 확인할 수 있다.

2. 전처리 및 변수생성과정

1 탐색적 자료분석

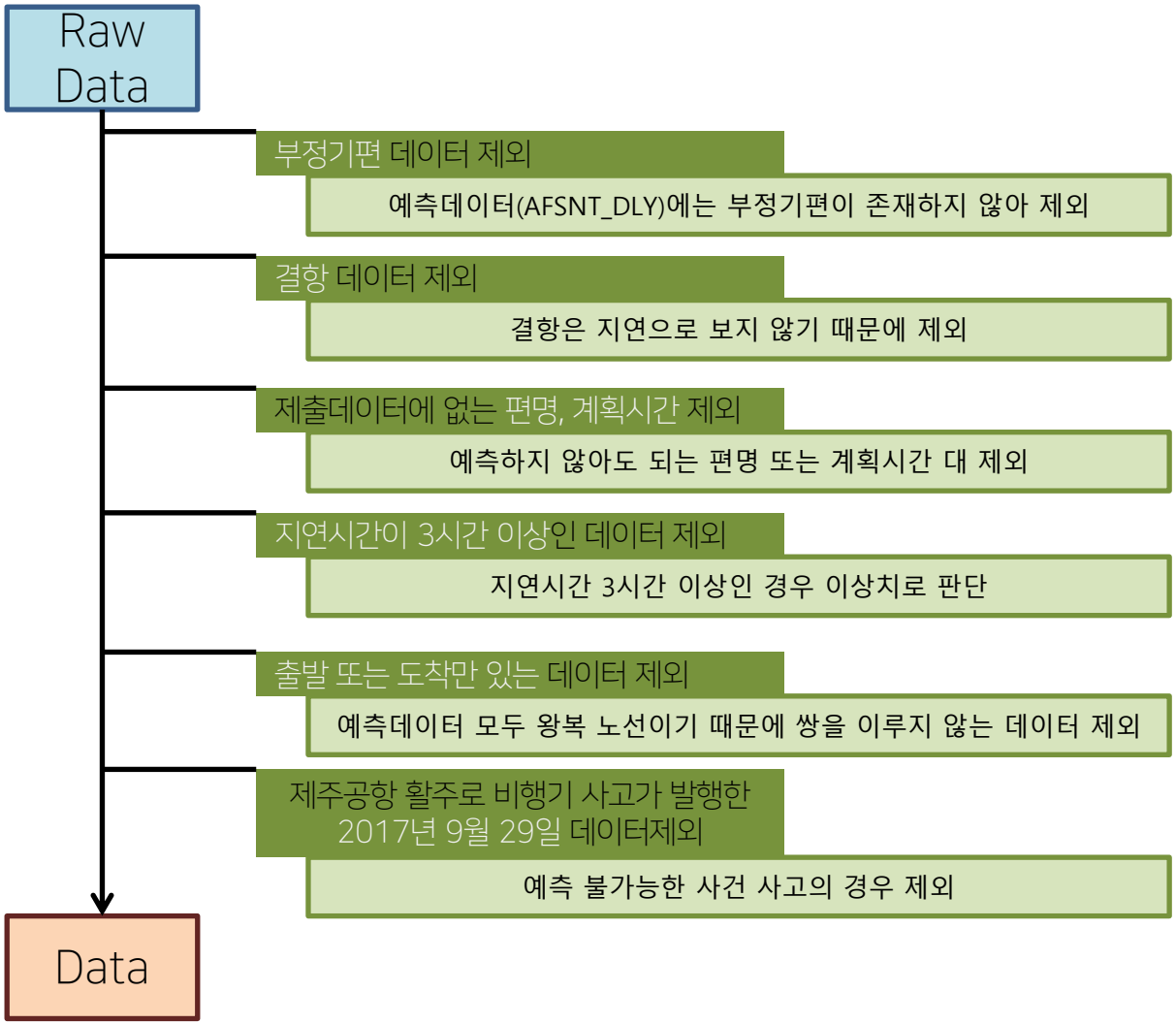
2 전처리 및 변수생성과정

3 모형 비교

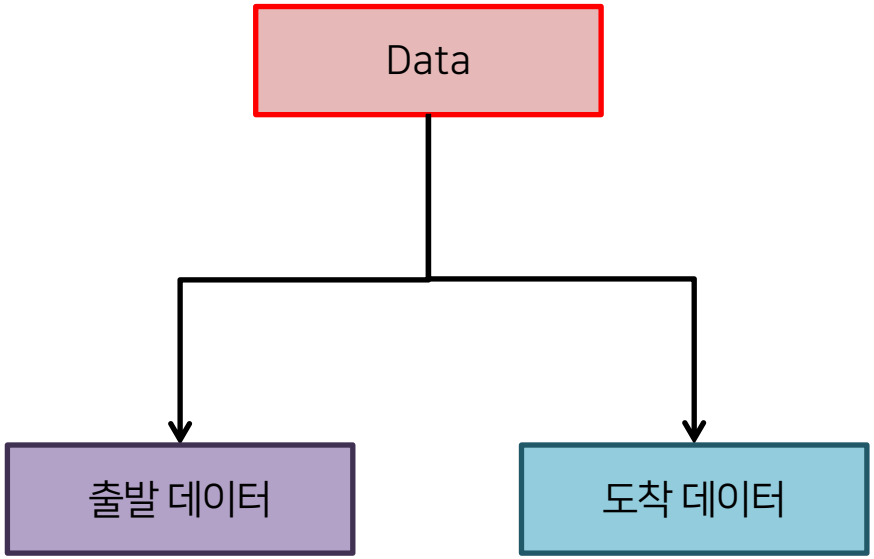
4 최종 모형

2-1. 전처리 과정

데이터 제거



2-2. 데이터 분리



출,도착 데이터가 성격이 상이 하고
출발 데이터의 지연시간이 도착데이터에 영향을 주어 변수로써 넣기 위해 데이터 셋을 나눔

2-3. 변수생성

YY_score	해당 *년도의 지연율*을 소수 둘째자리에서 반올림
MM_score	해당 *월의 지연율*을 소수 둘째자리에서 반올림
DY_score	해당 *요일의 지연율*을 소수 둘째자리에서 반올림
ARP_score	해당 *공항의 지연율*을 소수 둘째자리에서 반올림
ODP_score	해당 *상대공항의 지연율*을 소수 둘째자리에서 반올림
ARP_ODP_score	해당 *노선의 지연율*을 소수 둘째자리에서 반올림

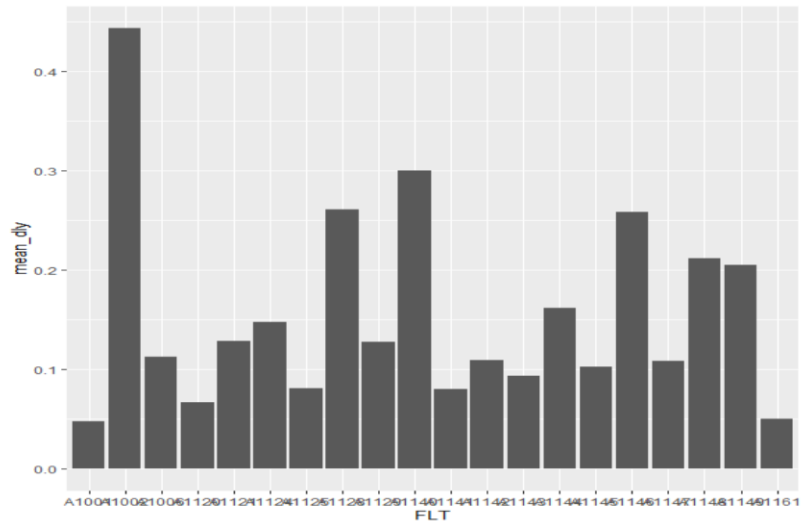
2-3. 변수생성

FLO_score	해당 *항공사의 지연율*을 소수 둘째자리에서 반올림
FLO_REG_n	해당 *항공사의 보유 비행기* 대수
Season_score	해당 *계절의 지연율*을 소수 둘째자리에서 반올림
AM_PM	*계획시간의 오전/오후* 여부
STT_score	*계획시간이 비슷한 시간대*의 지연율
ARP_STT3_n	해당 날짜의 공항에 *시간당 운항개수*

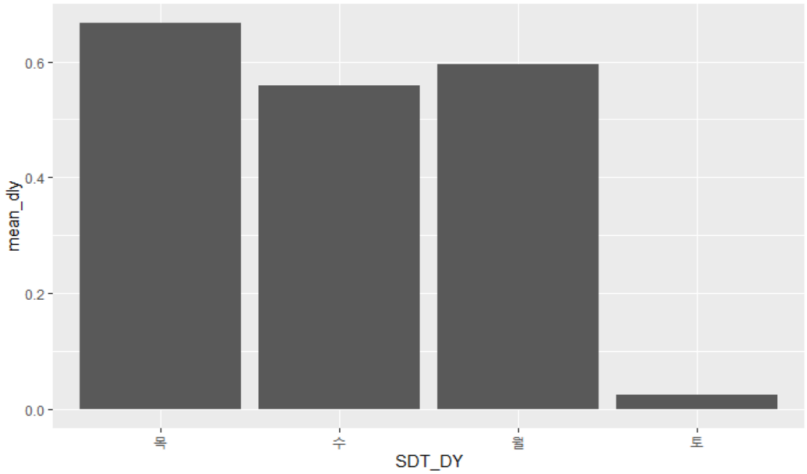
2-3. 변수생성

FLT_rank	해당 요일에 50회 이상 운행한 편명의 경우 *표본 백분위수를 5등분*하여 A/B/C/D/E 등급화
	해당 요일에 50회 미만 운행한 편명의 경우 *표본 백분위수를 3등분*하여 B/C/D 등급화
p_delay_time	위의 모든 변수를 설명변수로 하여 지연시간을 예측하는 XGBoost모형의 *예측 지연시간*
wea	해당 날짜의 *강도 8이상 항공기상* 관측여부 (출처 : 항공기상청 공공데이터 항공통계자료 기사)
DLY	지연 여부

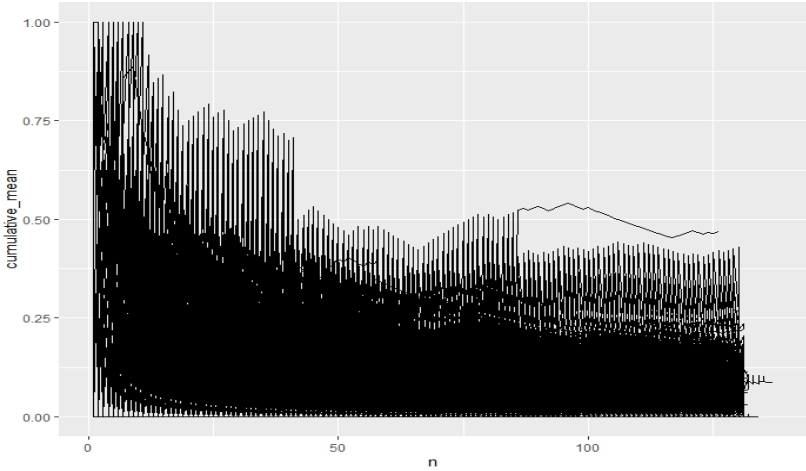
2-3. 변수 생성 – FLT_rank



편명별로 지연율의 차이가 유의하다.



각 편명,요일별로도 지연율의 차이가 유의하다. 위의 예시의 경우 A1002



편명,요일별 지연율이 운행횟수 50회를 기준으로 수렴함을 알 수 있다.

각 요일에 50회 이상 운행한
편명의 경우 표본 백분위수를 5등분하여
A/B/C/D/E 등급화하고
50회 미만 운행한 편명의 경우
표본 백분위수를 3등분하여
B/C/D 등급화 한 FLT_rank 변수 생성

3. 모형비교

1 탐색적 자료분석

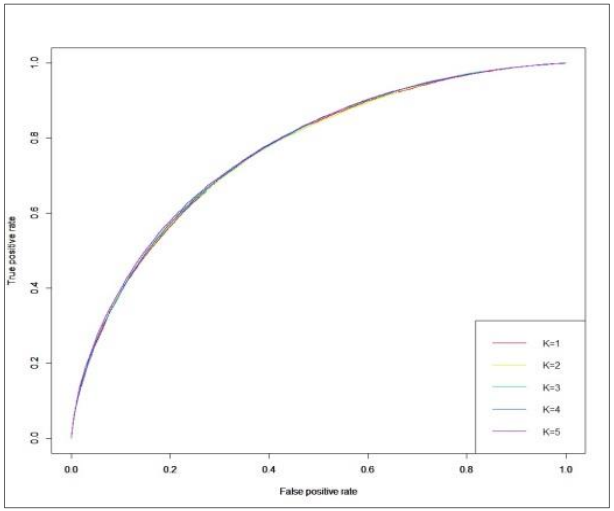
2 전처리 및 변수생성과정

3 모형 비교

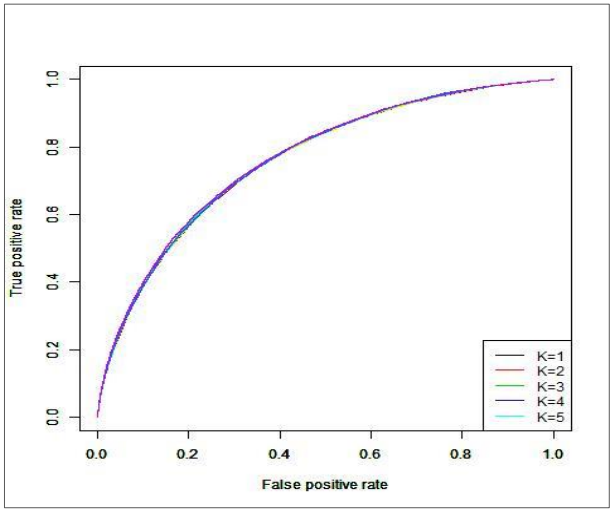
4 최종 모형

3. 모형비교 - 출발

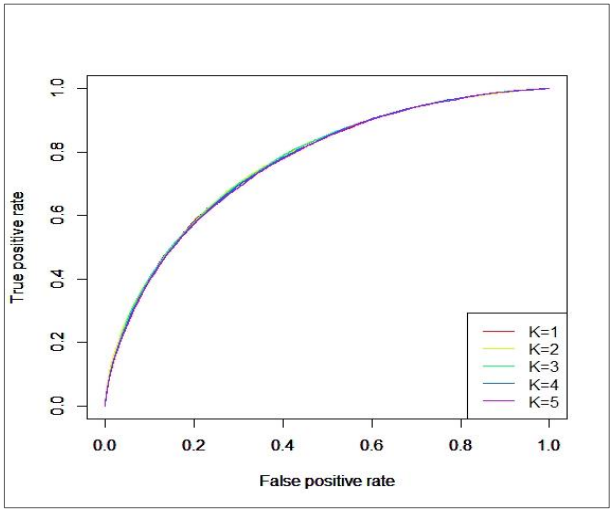
GLM



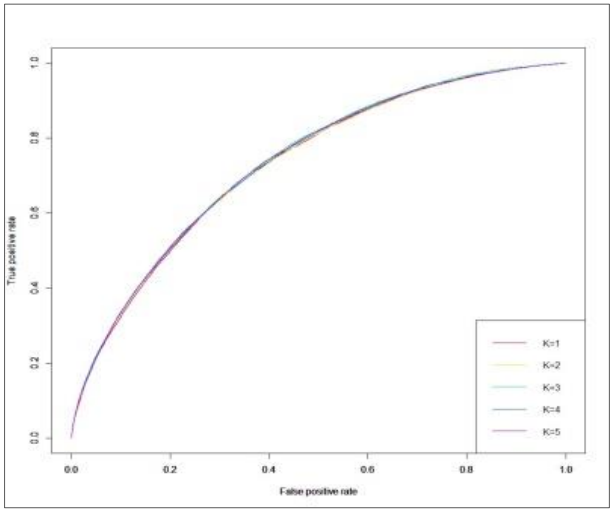
LDA



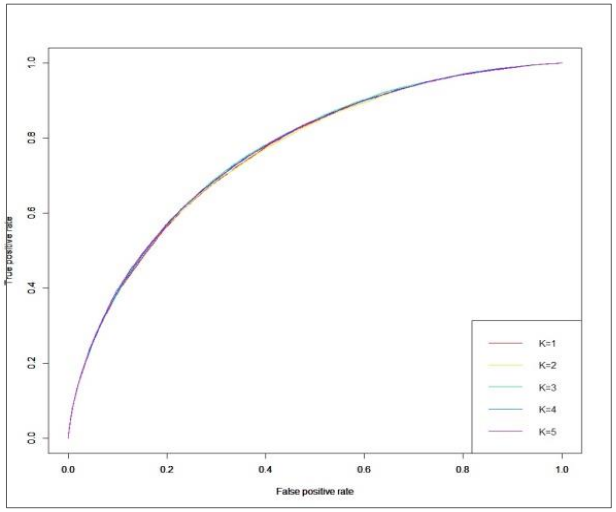
XGBoost



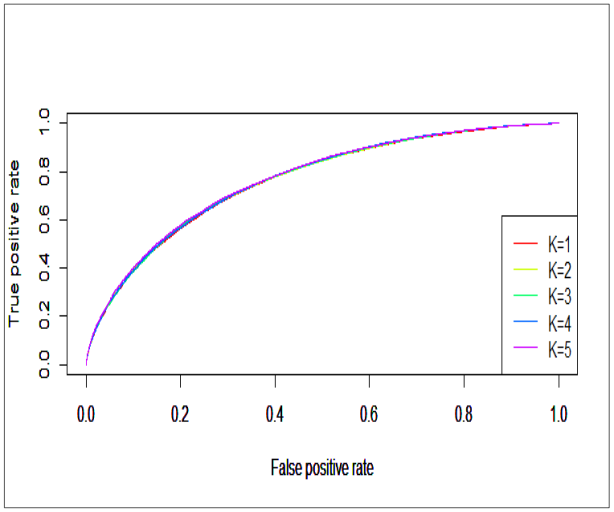
naive bayes



nnet

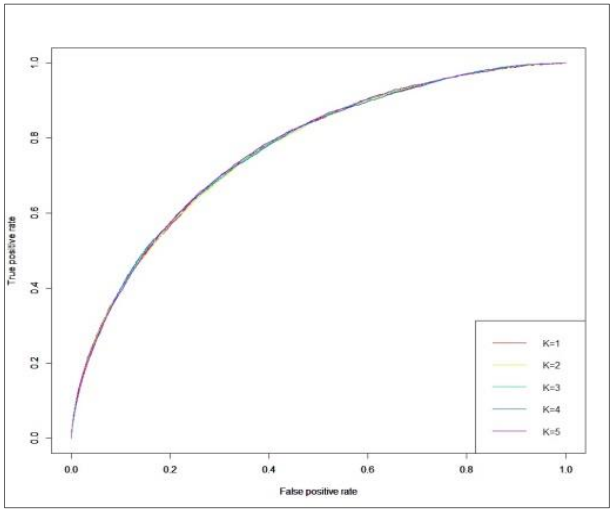


DNN

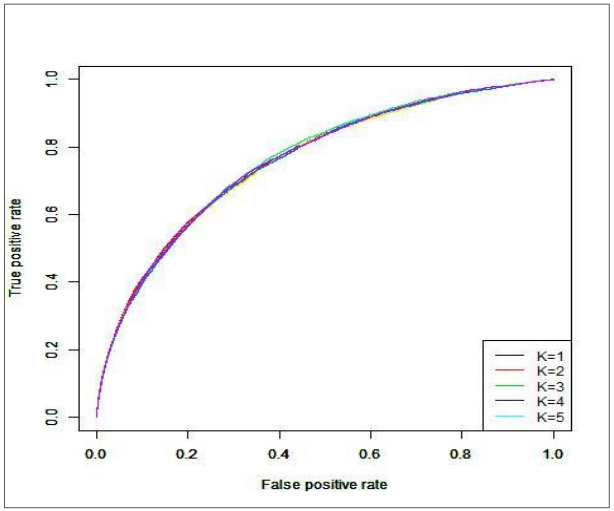


3. 모형비교 - 도착

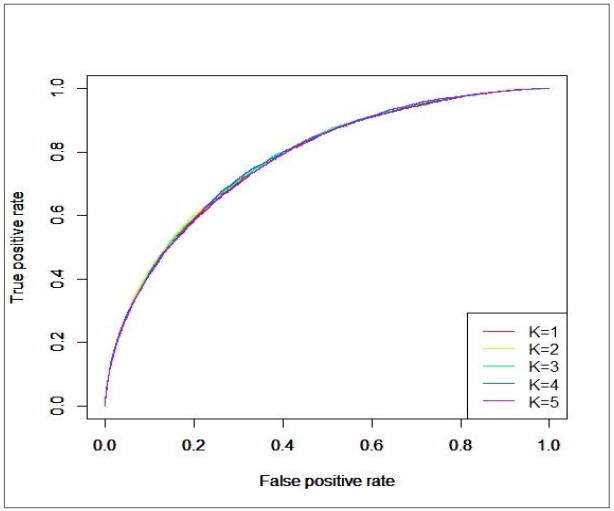
GLM



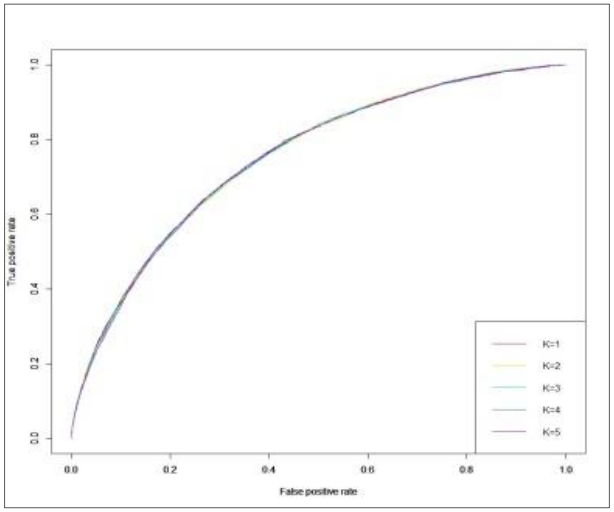
LDA



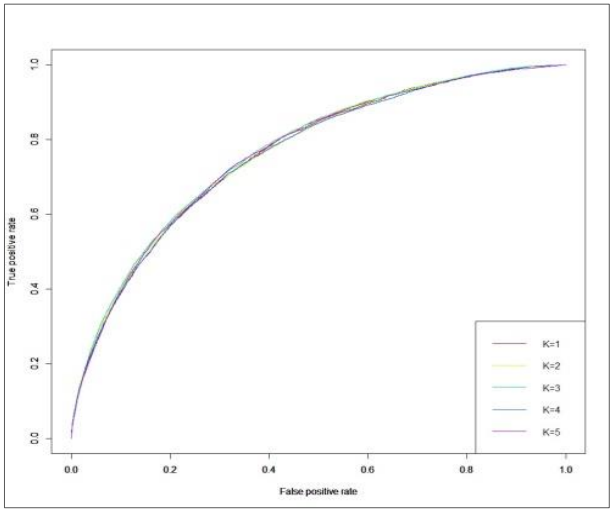
XGBoost



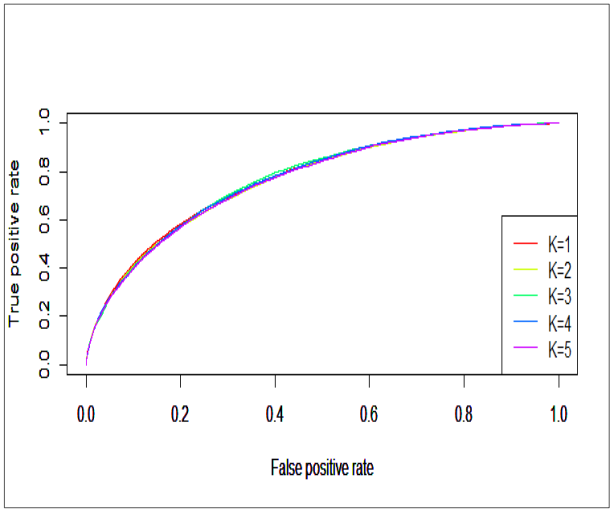
naive bayes



nnet

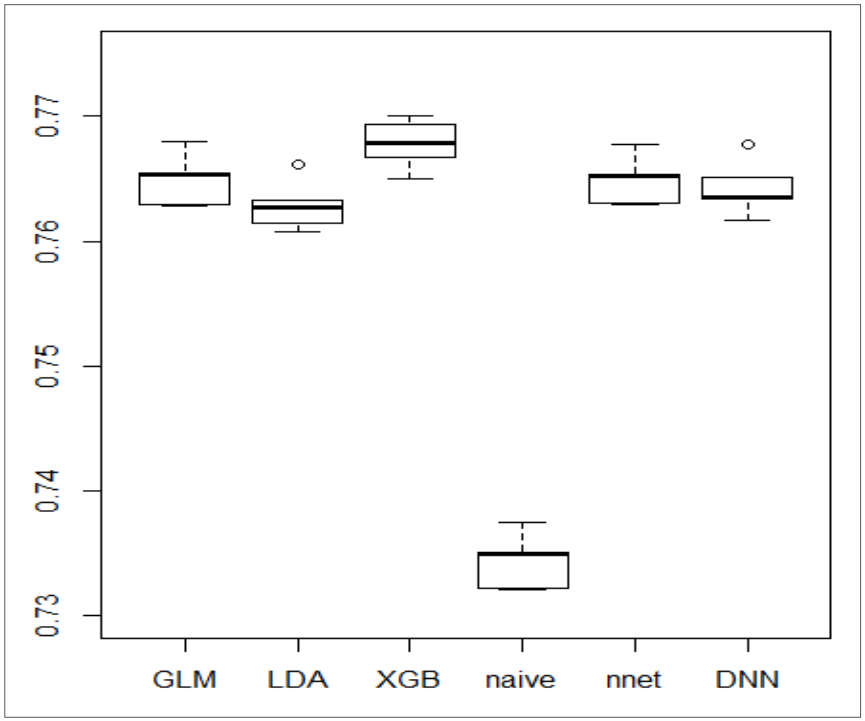


DNN

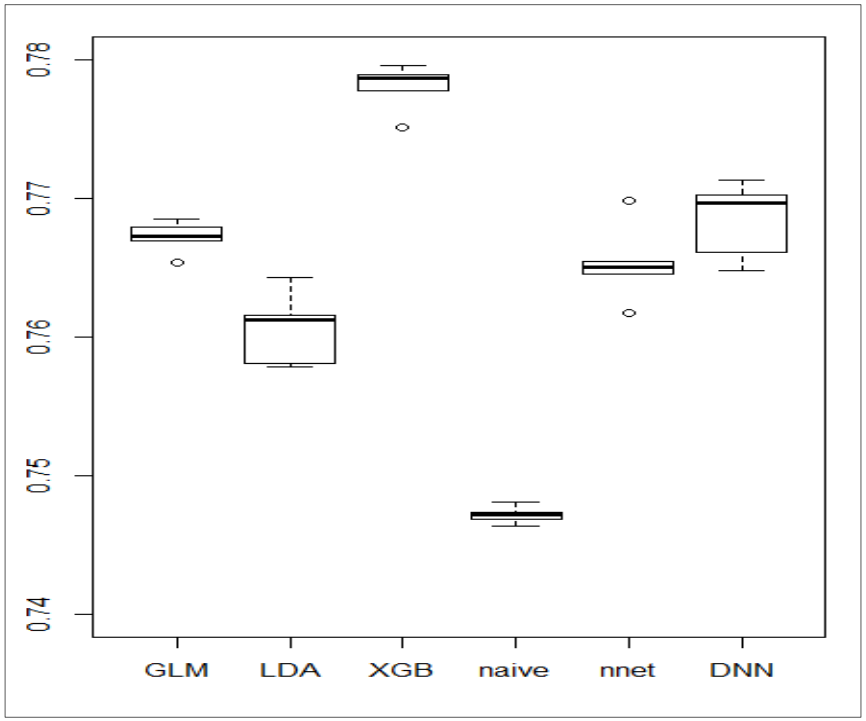


3. 모형비교

출발모형 AUC BoxPlot



도착모형 AUC BoxPlot



출발모형과 도착모형에서 5 fold에 validation set의 Box Plot을 그려보면,
XGBoost의 AUC가 가장 높게 나타나고 분산 또한 작은 것을 볼 수 있다.

4. 최종모형

1 탐색적 자료분석

2 전처리 및 변수생성과정

3 모형 비교

4 최종 모형

4-1. XGBoost

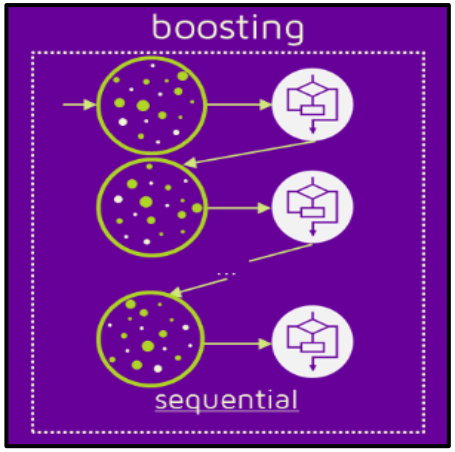
의사결정나무(Dicision Tree)

의사결정나무는 각각의 내부 노드에 존재하는 개별 속성의 *비동질성*을 평가하는 이진 트리로서, 각각의 잎 노드는 의사결정의 경로에 따라 나타나는 결과값 또는 클래스에 대응됩니다.

하지만 의사결정나무의 경우 입력노드의 작은 변동에도 트리구성이 크게 달라지게 됩니다. 만약 같은 분류의 데이터가 모여있지 않고 흩어져 있다면 성능에도 큰 영향을 미치게 됩니다.

(왼쪽 그림의 경우 "사각형인가?"로 내부노드의 비동질성을 나누었습니다.)

Boosting



이러한 단점을 보완 하기 위해, 성능이 약한 학습기 여러 개를 연결하여 예측하는 방법이 개발되었습니다. 이 방법을 Boosting이라 하는데, 하나의 학습기 결과가 또 다른 학습기가 학습될 때 도움을 주는 기법입니다.

대표 적으로 AdaBoost와 Gradient Boost 기법이 있습니다.

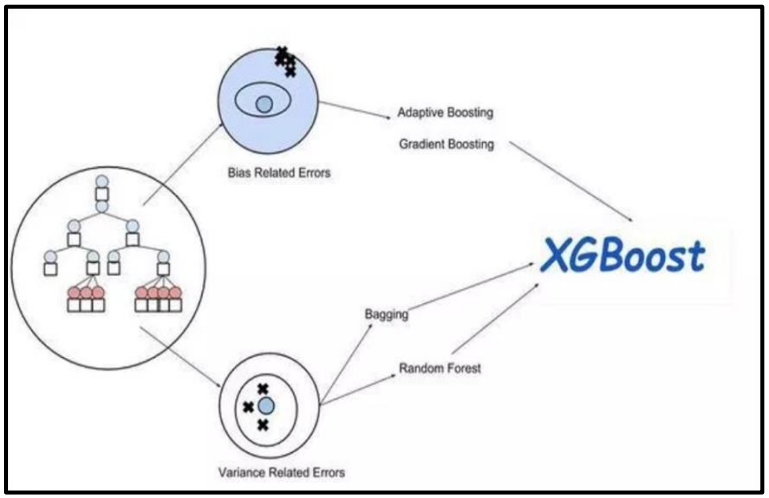
(Gradient Boost는 '경사하강법'을 이용하여 AdaBoost 보다 성능을 개선시킨 Boosting 기법입니다.)

4-1. XGBoost

XGBoost

Gradient Boost의 학습 성능은 좋지만, 수행시간/연산시간이 많이 걸린다는 단점이 있습니다. 이러한 단점을 획기적으로 개선한 방법이 바로 XGBoost입니다.

XGBoost는 의사결정트리를 구성할 때 **병렬 처리 기법**을 사용하여, 수행시간 측면에서 Gradient Boost보다 비약적인 상승을 이루었습니다.

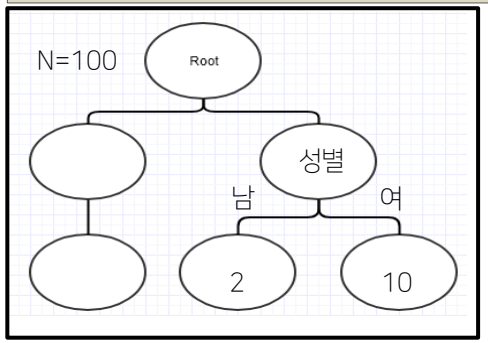


XGBoost의 장점

1. 병렬 처리를 사용하기에 **학습과 분류 속도가 높음**
2. 일반화 오차의 편향과 분산을 모두 **조절하여 낮출 수 있는 가변적인 모델**
3. 정규화 변수를 넣을 수 있어 **과적합이 잘 일어나지 않음**

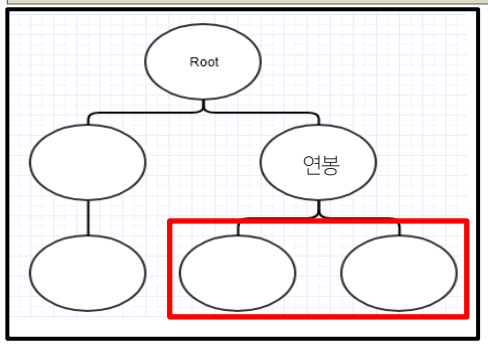
4-2. Importance

Cover



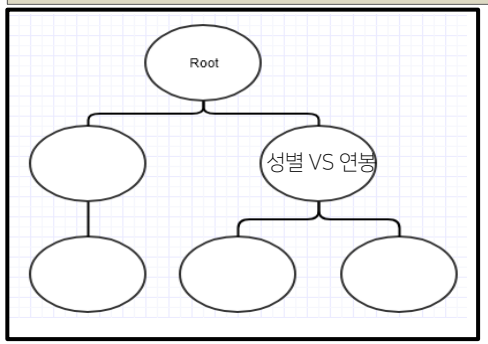
변수와 관련된 **관찰값의 상대적인 수**를 의미합니다. 예를들어 3개의 변수, 3개의 트리, 100개의 관측치가 있는 XGBoost모델이라 생각해 봅시다. 이때 하나의 트리에서 하나의 변수(성별)로 인해 구분되어지는 관측치의 개수가 2개 10개 일때, 성별의 Cover값은 12/100입니다.

Frequency



변수를 사용하여 잎 노드를 만들때 걸리는 **상대적인 시간**을 의미합니다. 예를들어 연봉 변수를 사용 하였을때 만들어진 잎 노드의 수가 2개라면 Frequency의 값은 2/3입니다.

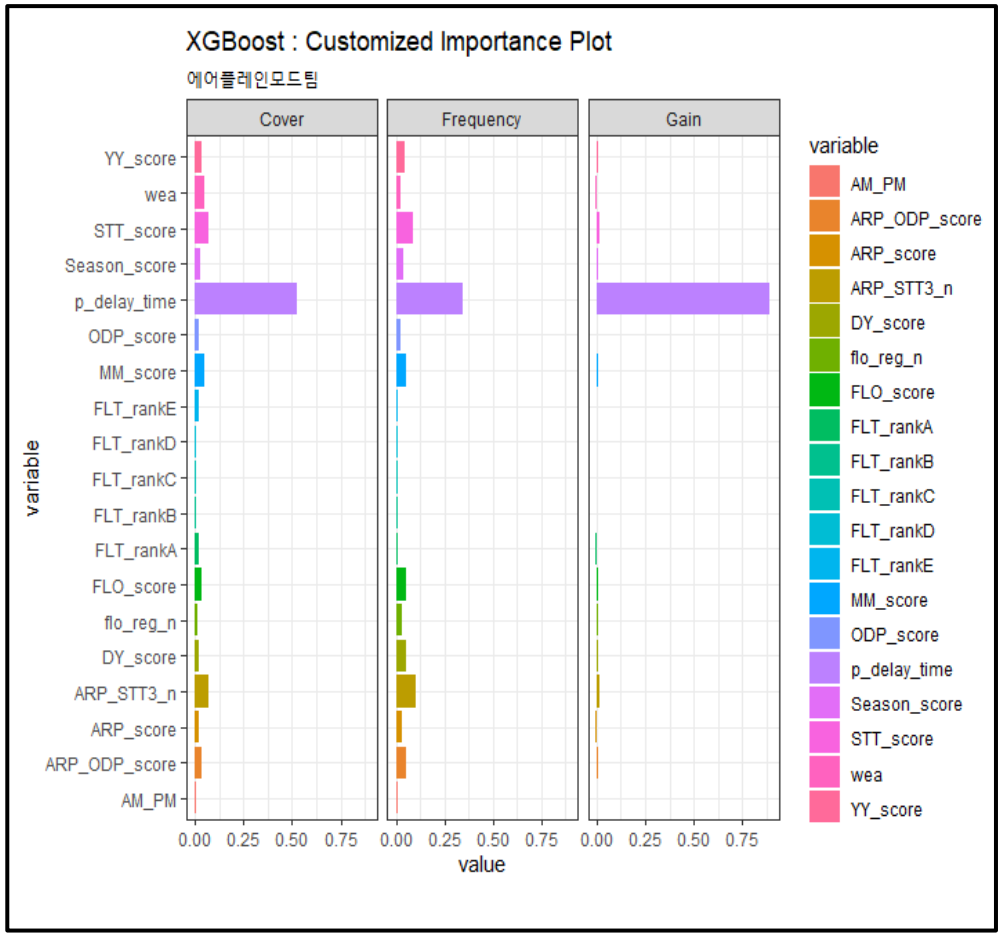
Gain



변수가 각각의 트리에서 **기여한 상대적인 정도**를 말합니다. 기여도는 분할에 각 변수를 사용할 때마다 감소한 평균손실을 기준으로 합니다.

4-2. Importance

XGBoost Importance



Cover, Frequency, Gain 세가지 기준에서 모두 변수의 중요도 순위가 동일하게 나타났습니다.

p_delay_time의 중요도가 세가지 기준에서 **압도적으로 높게** 나타났습니다.

4. 최종모형

Threshold 지정

2017~2018
출발 데이터의 9월 지연율
0.1756

2017~2018
도착 데이터의 9월 지연율
0.0586

예측해야하는
2019년 9월의
지연율이
과거 지연율과
비슷하게
Threshold를 지정

출발 데이터의
Threshold
: 0.23

출발 데이터의
Threshold
: 0.12

감사합니다.