데이터마이닝 9장

# 9.2

**연착 항공편 예측 FlightDelays.xls은 2004년 1월 동안 워싱텅 D.C.에서 출발하여 뉴욕에 도착하는 모든 민간 항공기들에 대한 정보를 담고 있다. 데이터에는 각 항공편에 대한 출발지 및 도착지 공항, 운항 경로 거리, 항공편 예정 날짜 및 시간 등에 관한 정보가 있다. 여기서 예측하고자 하는 변수는 항공편이 연착되는지 아닌지에 대한 여부이다. 연착은 예정된 시간보다 최소 15분 이상 늦게 도착하는 것으로 정의한다.**

**데이터 전처리: 요일, 항공기, 출발 공항, 도착 공항에 대한 가변수들을 만들어 데이터에 추가된 가변수의 총 개수를 17개가 되게 하시오. 항공편 예정 출발 시간을 두 시간 단위로 구간화하시오. 출발시간(DEP\_TIME)을 8구간으로 구분한 후, 이 새로운 변수를 다시 7개의 가변수들로 세분하시오(왜냐하면, 아침저녁 출퇴근 교통 혼잡시간 때문에 변수의 영향이 선형이 아니기 때문이다). 이렇게 하면, 출발 시간을 연속적 예측변수로 다루지 않아도 된다(연착은 교통 혼잡시간과 연과되었다고 여겨지므로 이를 연속형 변수로 다루지 않는 게 더 타당하다). 데이터를 학습 세트와 검증 세트로 나누시오.**

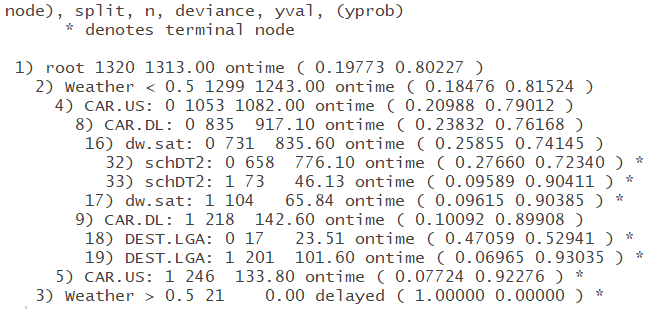
**9.2**

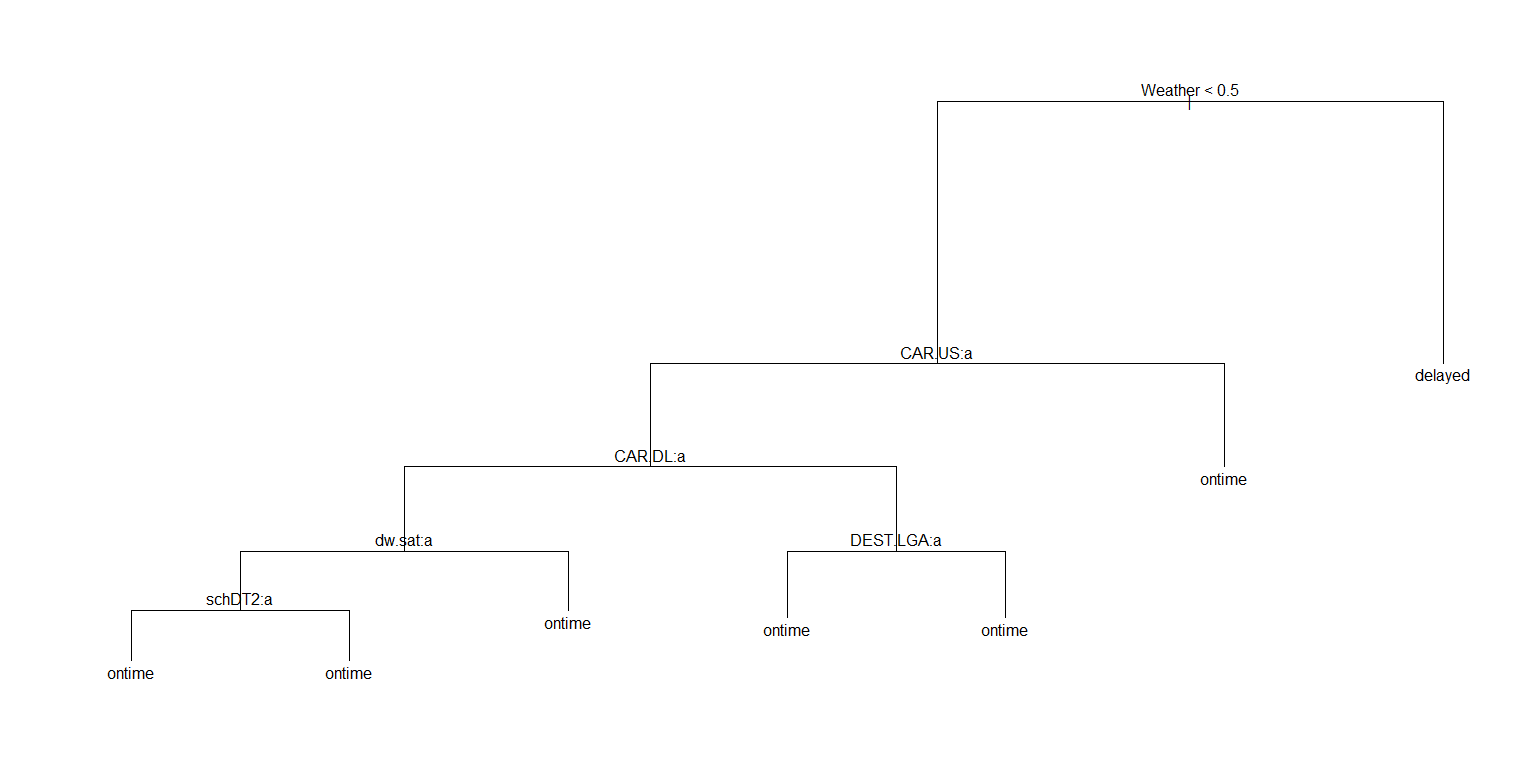
1. **모든 적절한 예측변수들을 사용하여 항공편 연착 변수에 대한 분류나무모델을 만드시오. 예측시점에서는 출발시간(DEP\_TIME)이 알려지지 않으므로(항공기 이륙 후 연착을 예측하는 게 분석이 목적이 아니므로) 모델에는 출발 시간을 포함하지 마시오. 분류 나무 메뉴의 세 번째 단계에서 최대 분할 수준(나무 깊이)의 수를 6으로 설정하시오. 잎 노드에 포함되는 최소 관측치의 수에 대한 제한을 제거한 후, 최적으로 가지치기된 나무 모델을 생성하시오. 나무 모델의 결과를 규칙으로 표현하시오.**

전처리 후 모델 적합을 위한 예측변수는 DISTANCE, Weather, Flight.Status, CARRIER(가변수), DEST(가변수), ORIGIN(가변수), CRS\_DEP\_TIME(예정된 출발시간 가변수) 들을 사용하였다.

문제에서 최대 분할 수준=6이라는 조건은 중간 노드가 6개 즉 잎 노드가 7개라는 것을 뜻한다.

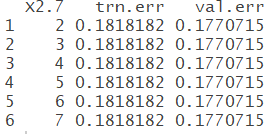
잎 노드에 포함되는 최소 관측치의 수에 대한 제한을 제거하라는 것은 잎 노드의 관측치가 최소 1개가 될 수 있다는 것이고 이는 나무를 완전 성장시키는 것을 의미한다. 즉 나무를 완전 성장시키고, 잎 노드가 7개가 되게 가지를 치는 것과 같다.



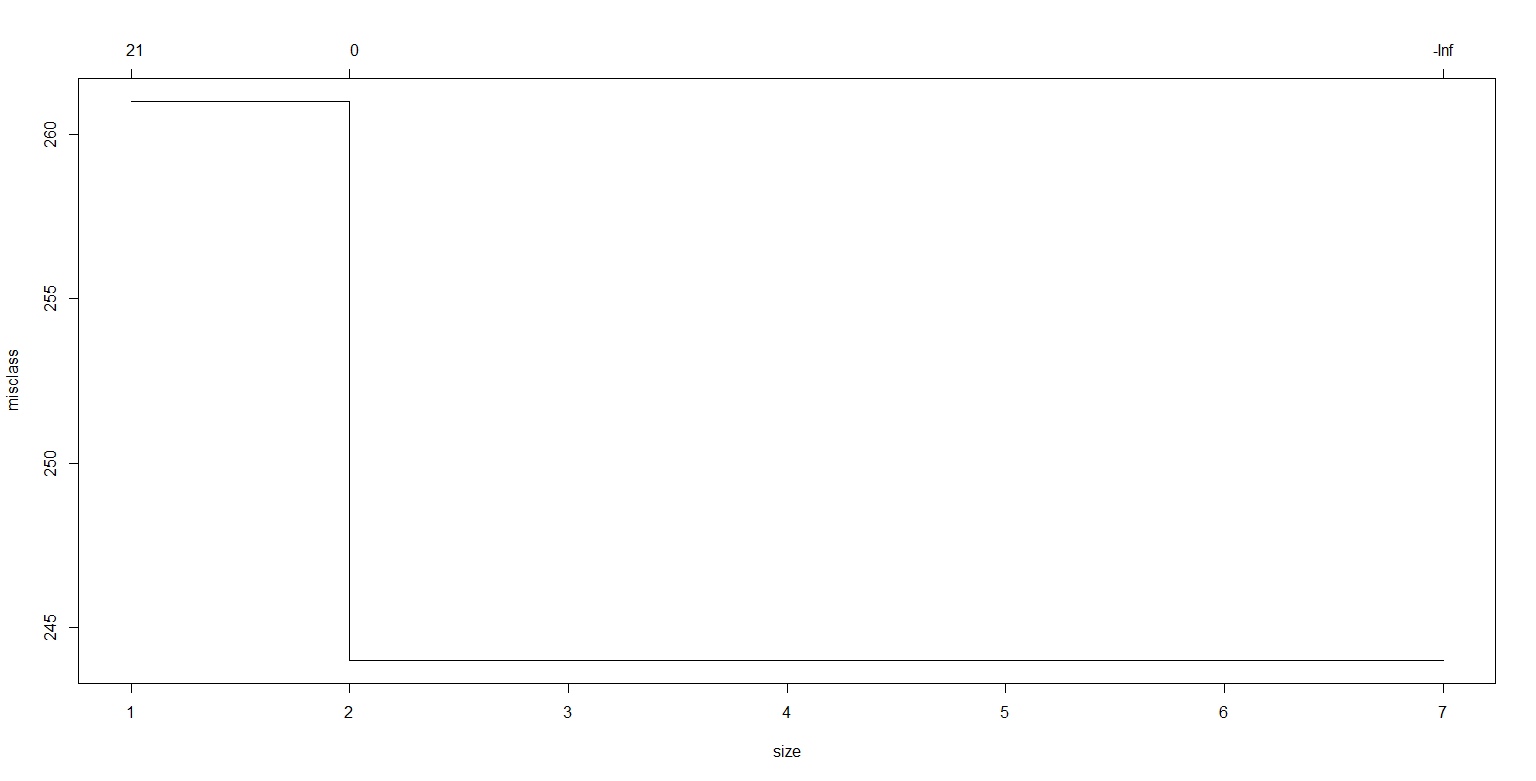


위의 분류는 완전성장한 나무에서 중간 노드가 6개가 되게 가지를 친 나무모델이다. 중요한 변수들은 날씨, 비행기 기종이 US인지 아닌지, 비행기 기종이 DL인지 아닌지 등이 있다.

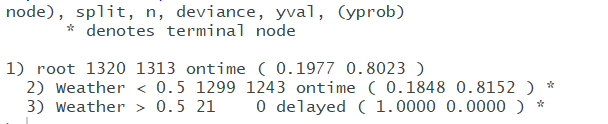
나무 깊이가 6인 위의 나무 모델에서 최적으로 가지치기 된 나무모델을 구하기 위해서 나무의 깊이 별로 검증 데이터에서의 오분류율을 구해보면



와 같이 나타난다 즉 잎 노드의 수가 2개~7개로 증가함에 따라 검증 데이터에 대한 오분류율을 나타내고 있다. 오분류율이 가장 작은 0.1770715에 표준오차(=0.01286078)를 더 해준 값을 구하면 0.1899323 이라는 값을 얻을 수 있고 이 값보다 검증데이터에서의 오분류율이 작고 분할 수가 가장 적은 나무 모델을 선택해야한다. 즉 최적의 가지치기 된 나무 모델은 분할을 1번한 결정 노드가 1개인 나무 모델이라고 할 수 있다. 마찬가지로 cv.tree 함수를 사용해 최적의 가지치기 된 나무 모델을 보면

와 같은 그래프를 구할 수 있는데 이는 2일 때 즉 분할을 1번 했을 때가 최적의 가지치기 된 모델이라는 것을 알려준다.

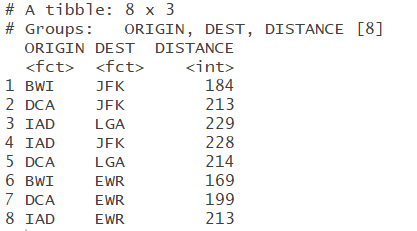
이때의 나무 모델의 규칙은

와 같다.

**b. 월요일 오전 7시에 DCA에서 EWR을 비행해야 한다면, 이 나무 모델을 사용할 수 있겠는가? 필요한 다른 정보는 무엇인가? 이 모델은 실제로 사용할 수 있는 모델인가? 중복된 정보는 무엇인가.**

월요일 오전 7시에 DCA에서 EWR로 비행을 한다는 레코드는 이 모델에 사용할 수 없다. 나무 깊이가 6인 모델에 사용하기 위해 필요한 변수인 Weather, CARRIER 에 대한 변수가 없기 때문에 나무 모델을 적합 시킬 수 없다. 마찬가지로 최적의 가지치기 된 나무 모델에서도 Weather 변수가 없기 때문에 사용할 수 없다.

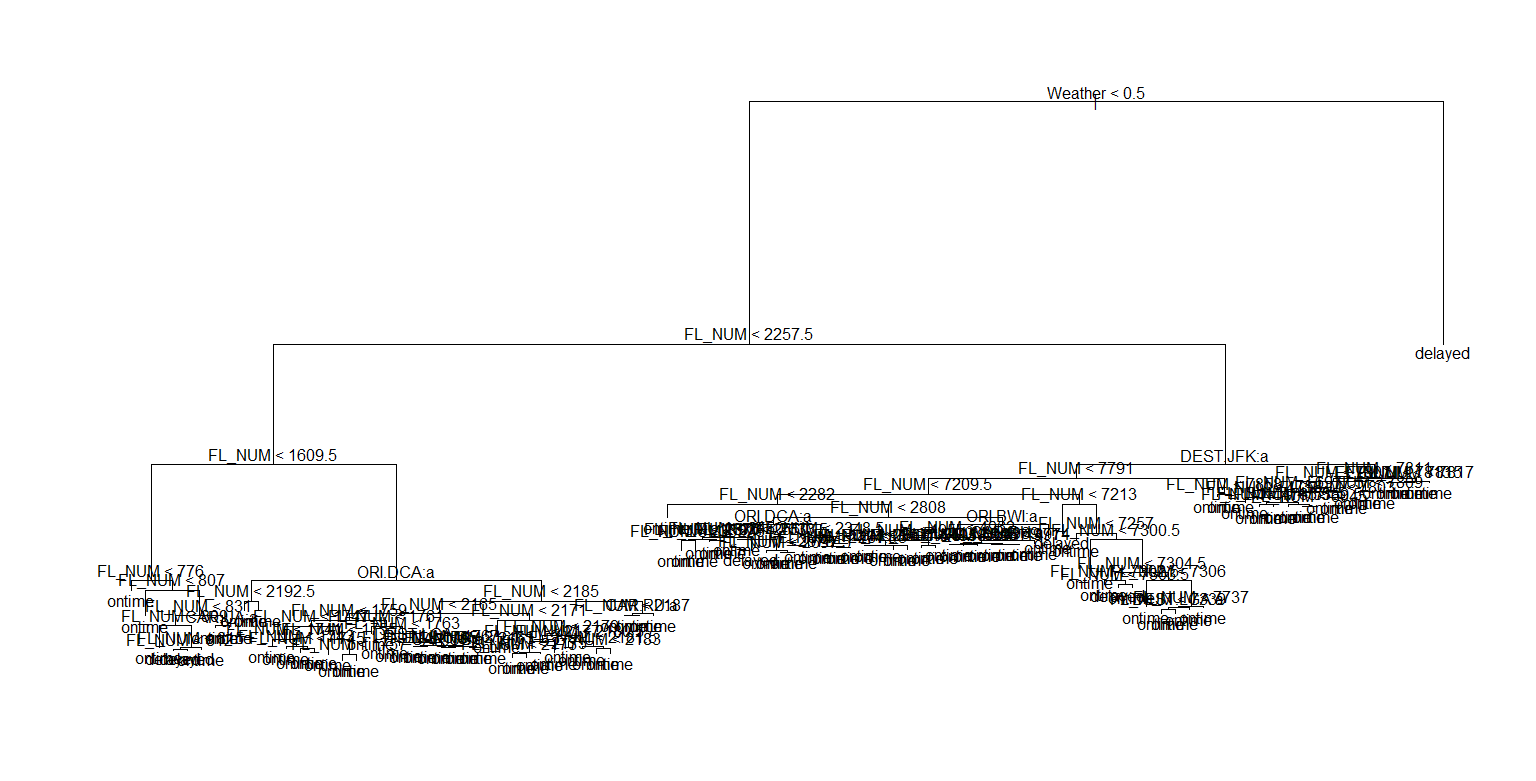
a.에서 생성한 나무 모델에 사용하기 위해서 필요한 정보는 이 모델에서의 결정 노드에 존재하는 변수들이다. 나무 깊이가 6인 모델에서 중복되는 정보는 ORIGIN, DEST, DISTANCE이다.

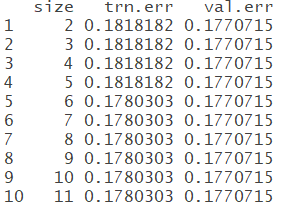
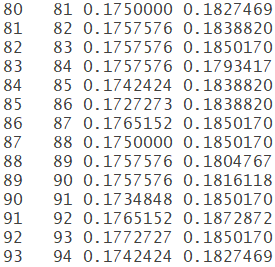


위의 테이블과 같이 출발지와 목적지의 수준 조합에 따른 거리는 항상 같고 거리가 중복되지 않기 때문에, 거리 변수만 사용하거나, 출발지, 목적지 두 변수를 사용해야 정보가 중복되지 않는다

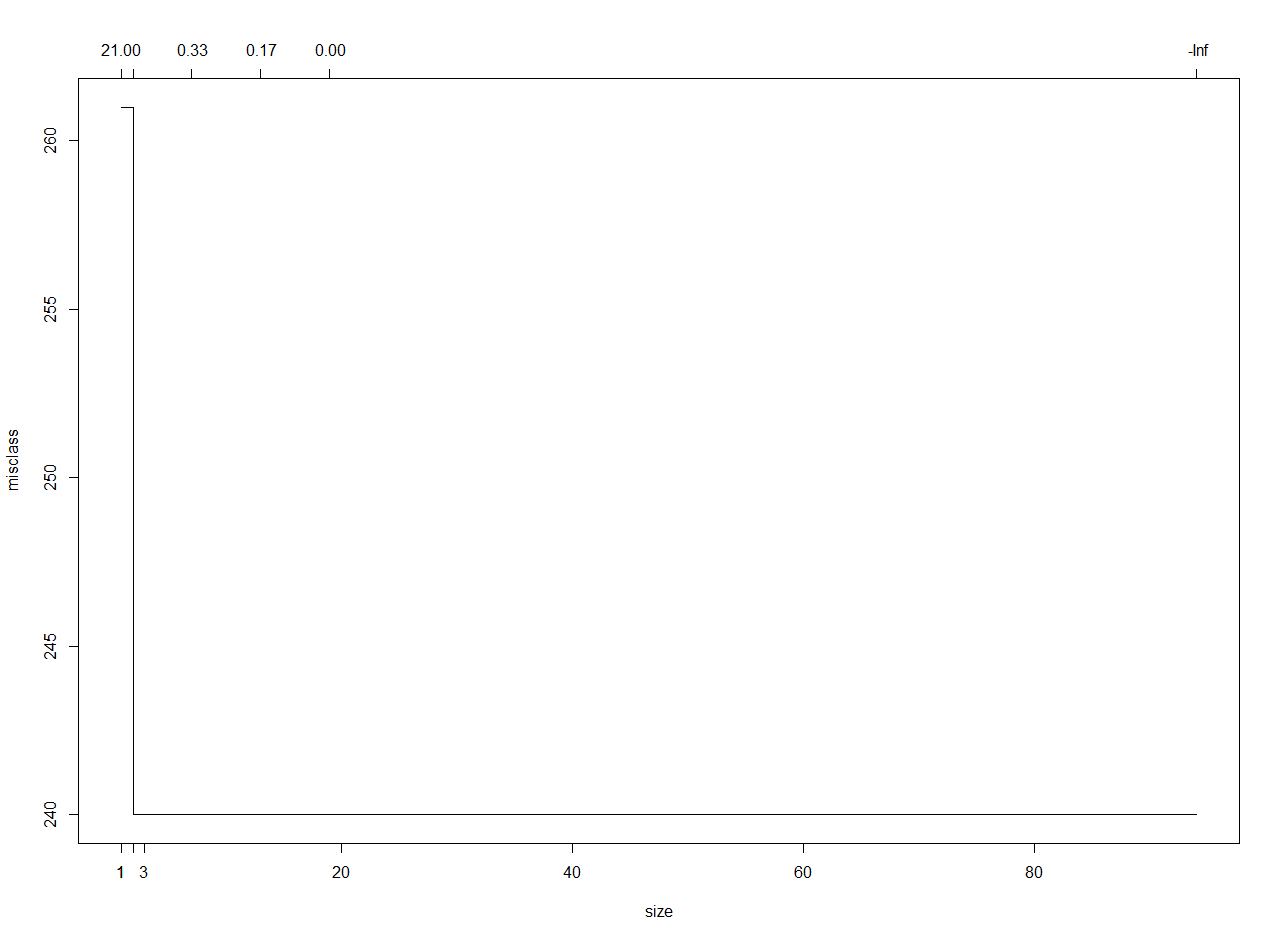
**c. 항공편 예정 날짜에 관한 예측변수를 제외하고 분류나무 모델을 하나 더 만드시오. 메뉴에서 완전히 성장한 나무 모델과 최적으로 가지치기 된 나무 모델을 보여주는 선택사항을 택하시오. 최적의 가치치기 된 나무 모델이 하나의 종점 노드를 갖게 된 것을 확인하시오.**

항공편 예정 날짜에 관한 예측변수와 앞서 생성했던 모델에서 중복되는 정보를 제외하고, 예측변수 (FL\_NUM, Weather, Flight.Status, CAR.OH, CAR.DH, CAR.DL, CAR.MQ, CAR.UA, CAR.US, CAR.RU, DISTANCE) 들을 사용하여 완전 성장한 나무 모델을 적합하면



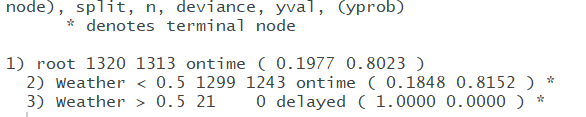
와 같은 모델을 얻을 수 있고 이 모델을 가지치기 하기 위해 검증 데이터에서 오분류율을 구하면,  와 같이 구할 수 있다.

이 때 가장 낮은 검증 오차율은 0.1748014이고 이 때의 표준 오차(0.01279565)을 더해주면 0.1875957이라는 값을 구할 수 있다. 이 값보다 검증 오차율이 낮고 분할 수가 가장 작은 나무 모델은 선택하는 것이 최적의 가지치기 된 모델이 된다. 즉 잎 노드의 수가 2개인 결정 노드가 1개만 존재하는 모델이 최적의 가지치기 된 모델이된다. 마찬가지로 이때의 cv.tree를 이용한 최적의 가지치기 된 모델을 구하면



와 같이 얻을 수 있고 잎 노드의 수가 2개일 때 최적의 가지치기 된 나무 모델인 것을 알 수 있다.

**i. 이 나무 모델은 분류를 위하여 어떻게 사용되는가? (분류규칙은 무엇인가)**

가 분류 규칙이 된다.

**ii. 이 규칙에 상응하는 것은 무엇인가?**

이러한 분류 규칙은 나이브 규칙을 의미한다. 직사각형 내에서 가장 높은 투표율을 갖는 클래스 즉, 과반수의 클래스로 분류하기 때문이다.

**iii. 완전히 성장한 나무 모델을 검토하시오. 이 나무 모델에서 가장 좋은 세 개의 예측변수는 어떤 것들인가?**

Weather, FL\_NUM, ORIGIN=DCA인지 여부 변수

**iv. 기술적으로 볼 때, 가지치기 된 나무 모델의 결과가 왜 단일 노드를 가진 나무 모델이 되었다고 생각되는가?**

cv 그래프를 통해 보면 size=2일 때 이후로 즉 size=2보다 커진다고 해도 오분류율이 더 작아지지 않기 때문에 나무가 더 이상 성장하지 않는 것이다. 마찬가지로 검증 오차율을 통해 살펴보면 검증 오차율의 최소값에 표준 오차를 합해준 값보다 size=2일 때 검증 오차율이 적고 리프 노드의 수가 가장 적기 때문에 단일 노드를 가진 나무 모델이 선택되는 것이다.

**v. 최적으로 가지치기 된 나무 모델(여기서는 단일 노드를 갖게 되었음)을 사용하지 않는 대신 완전히 성장한 나무 모델의 최상위 수준을 사용하는 것이 갖는 단점은 무엇인가?**

학습 데이터에 대해 완전 성장한 나무는 학습 데이터에 대한 성능은 좋지만 학습 데이터에 과적합한 것이고, 이러한 과적합 된 모델을 새로운 레코드 예측에 사용하면 오히려 예측 정확도가 낮기 때문에 학습 데이터에 대한 성능은 낮지만 새로운 레코드에 대한 예측 성능이 좋은 최적으로 가지치기 된 나무 모델을 사용해야한다.