Data Mining for Business Analytics

Assignment. 02

2023.05.15.

9.2 Predicting Delayed Flights

연착 항공편 예측

- [FlightDelays.csv]는 2004년 1월 동안 워싱턴 D.C.에서 출발하여 뉴욕에 도착하는 모든 민간 항공기들에 대한 정보를 담고 있다. 데이터에는 각 항공편에 대한 출발지 및 도착지 공항, 운항경로 거리, 항공편 예정 날짜 및 시간 등에 관한 정보가 있다. 예측하고자 하는 변수는 항공편이 연착되는지 아닌지에 대한 여부이다. 연착은 예정된 시간보다 최소 15분 이상 늦게 도착하는 것으로 정의한다.
- 데이터 전처리: DAY_WEEK 변수를 범주형 변수로 바꾸시오. 항공편 예정 출발시간을 8구간으로 구간화하시오. 이 변수들과 다른 모든 열(column)들을 예측변수로 사용하시오(DAY_OF_MONTH) 변수 제외). 데이터를 학습 셋(60%)과 검증 셋(40%)으로 나누시오.

a. 모든 적절한 예측변수들을 사용하여 항공편 연착 변수에 대한 분류나무 모델을 만드시오. 예측 시점에서는 DEP_TIME(실제 출발시간)이 알려져 있지 않으므로 (항공기 이륙 후 연착을 예측하는 게 분석의 목적이 아니므로) 모델에는 DEP_TIME을 포함하지 마시오. 최대 깊이(depth)=8과 최소 불순도 감소(impurity decrease)=0.01 수준으로 나무모델을 사용하시오. 나무모델의 결과를 규칙으로 표현하시오.

9.2 Predicting Delayed Flights

연착 항공편 예측

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

c. '날씨(weather)'를 예측변수에서 제외하고 (a)와 동일한 분류나무 모델을 하나 더 만드시오. 가지치기 된 나무모델과 가지치기 되지 않은 나무모델을 함께 보이시오. 최적의 가지치기 된 나무모델이 하나의 단말 노드를 갖게 된 것을 확인하시오.

   I. 가지치기 된 나무모델은 분류를 위하여 어떻게 사용되는가? (분류 규칙은 무엇인가?)

   II. 이 규칙은 무엇과 동일한가?

   III. 가지치기 되지 않은 나무모델을 검토하시오. 이 나무모델에서 가장 좋은 3개의 예측변수는 어떠한 것들인가?

   IV. 가지치기 된 나무모델을 사용하지 않고, 가지치기 되지 않은 나무모델의 최상위 수준을 사용한다면 어떠한 단점이 있겠는가?

4.4 Chemical Features of Wine

sol) 연착 항공편 예측

a.

```
|--- Flight Status_ontime <= 0.50
|   |--- class: 1
|--- Flight Status_ontime >  0.50
|   |--- class: 0
```

b. 월요일 오전 7시에 DCA에서 EWR을 비행해야 한다면, 이 나무모델을 사용할 수 있겠는가?

=> Flight Status_ontime이라는 변수 자체가 예측하려는 대상이기 때문에 모델을 실제 상황에서 사용하기는 어려울 것으로 보임

필요한 다른 정보는 무엇인가?

=> 항공사, 비행기의 유형, 날씨, 공항의 교통량 등의 정보가 필요함

중복된 정보는 무엇인가?

=> Flight Status_ontime변수와 isDelayed변수가 중복됨

4.4 Chemical Features of Wine

sol) 연착 항공편 예측

C.

가지치기 된 나무모델은 분류를 위하여 어떻게 사용되는가?

→ 분류문제를 해결하는데 사용. 나무모델은 if-else 질문을 통해 데이터를 분류하고 if-else는 특성 값을 기준으로 데이터 분할.

이 규칙은 무엇과 동일한가?

→ Flight Status_ontime변수에 대한 임계값을 설정하여 항공 지연 여부를 결정하는 규칙과 동일함

가지치기 된 나무모델을 사용하지 않고, 가지치기 되지 않은 나무모델의 최상위 수준을 사용한다면 어떠한 단점이 있겠는가?

→ 과적합 및 해석이 어려움(불가)