Assignment #4

Due date: 5월 4일 (목)

Predicting Delayed Flights

항공기의 연착(delay) 여부를 예측하는 것은 항공사와 공항 등 항공기 운항과 관련된 주체들에게 매우 중요하다. 항공기의 연착에 따라 대체 항공기 이용료, 숙박 비용, 공항 사용료 등의 비용 발생이 매우 크기 때문이다. **FlightRecords.csv** 파일은 2004년 1월동안 Washington, DC 지역으로부터 New York City로 운행한 2201개의 항공기 운행 기록을 포함한다. 본 문제에서는 다음 7개의 변수를 사용하여 항공기의 연착 여부를 예측해 본다.

- **dayweek**: 운행 요일 (**1**: *Mon*, **2**: *Tue*, ..., **7**: *Sun*)
- **deptime**: 출발시각 (예: 1455 = 14시55분, 839: 8시39분)
- **origin**: 출발공항코드(**DCA**: *Reagan Nation*, **IAD**: *Dulles*, **BWI**: *Baltimore-Washington Int'l*)
- **dest**: 도착공항코드(**JFK**: *Kennedy*, **LGA**: *LaGuardia*, **EWR**: *Newark*)
- **carrier**: 항공사코드(**CO**: *Continental*, **DH**: *Atlantic Coast*, **DL**: *Delta*, **MQ**: *American Eagle*, **OH**: *Comair*, **RU**: *Continental Express*, **UA**: *United*, **US**: *USAirways*)
- **weather**: 날씨 (0: *OK*, 1: *Bad*)
- **delay**: 연착여부("*delayed*" or "*ontime*")

1. 다음의 순서로 data preprocessing을 진행하자.

   - 항공기 출발시각(deptime)이 6시 이전이거나 22시 이후인 데이터는 빈도 수가 매우 적으므로 데이터셋에서 제외시킨다.
   - 수치값으로 표현되어 있는 출발시각을 6시부터 22시까지 각 시간대를 나타내는 범주형 변수로 변환한다 (Hint: 원 데이터를 100으로 나눈 후 정수값으로 내림. 그 후 factor로 변환)
   - 수치값으로 표현되어 있는 dayweek와 weather 변수를 factor로 변환한다.
   - factor로 표현되어 있는 delay 변수가 가지는 level의 순서를 "ontime", "delayed" 순으로 변환한다 (logistic regression 수행 시에 연착하는 경우를 $p(X) = 1$로 만들기 위해서).

2. 요일 별 연착비율, 출발 시간대 별 연착 비율, 출발 공항 별 연착비율, 도착 공항 별 연착 비율, 항공사 별 연착 비율, 날씨 별 연착 비율을 각각 그래프로 시각화해보자. 어떤 특성을 관찰할 수 있는가?

3. 7개의 모든 변수들 간의 상관관계를 시각화해보자. 어떤 특성을 관찰할 수 있는가?

4. 데이터셋을 70:30 비율로 training set과 test set으로 분할하자. 이때 stratified sampling을 활용하여 두 set에서 delay 변수의 분포가 크게 차이가 없도록 분할하자.

5. 데이터시각화로부터 weather 변수가 "Bad" 인 경우에는 항상 항공기가 연착되는 것을 관찰할 수 있다. 따라서 weather가 Bad이면 항공기가 연착되고, weather가 OK일 경우 항공기가 연착되지 않는 것으로 예측하는 단순한 모델을 baseline model이라 하자. Training set에 대해 baseline model을 적용했을 때의 confusion matrix를 계산해보자.

6. Training set을 대상으로, 연착여부(delay)를 나머지 모든 변수를 사용하여 예측하기 위한 logistic regression model을 수립해보자.

    1. 변수 deptime19의 regression coefficient에 대한 추정값은 얼마인가? 이 추정값을 바탕으로 출발 시각이 19시대인 항공기에 대해서 어떠한 해석을 할 수 있는가? (Hint: 범주형 변수 deptime을 model에 추가할 때 deptime6을 제외한 deptime7 ~ deptime21에 대한 dummy 변수가 만들어진다.)
    2. 날씨에 문제가 없는 목요일 15시에 IAD에서 출발하여 EWR로 도착한 Delta 항공기가 연착될 확률은 얼마로 예측되는가?
    3. Threshold $k = 0.3, 0.5, 0.7$에 대해서 각각 training set에 대한 confusion matrix를 계산해 보자. 어떠한 경향을 관찰할 수 있는가?
    4. 위의 결과를 바탕으로 Baseline model과 logistic regression model의 성능을 비교해보자.

7. Training set을 대상으로 Lasso regression을 적용하여 logistic regression model을 수립해보자. CV의 결과 바탕으로 모델에 포함되는 feature의 수와 예측정확도를 모두 고려했을 때 가장 적합한 모델을 선택하자.

    1. 어떠한 기준으로 모델을 선택하였으며, 최종적으로 모델에 어떠한 변수들이 포함되었는가?
    2. 기본 logistic regression model과 Lasso를 적용한 logistic regression model의 성능을 나타내는 ROC Curve를 하나의 그래프로 시각화하고, AUC값을 비교해 보자. Lasso regression의 효과가 있다고 말할 수 있는가? (training set과 test set에 대해서 각각 비교해보자.)

8. Training set을 대상으로 *k*-nn을 적용해보자. 이때 cross validation으로 Accuracy가 가장 높은 best k 값을 찾는다. best k 값은 얼마인가?

9. Training set을 대상으로 *SVM*을 적용해보자. RBF Kernel을 활용하고, cross validation으로 Accuracy가 가장 높은 파라미터의 조합을 찾는다. 어떤 파라미터 값을 사용했을 때 RBF Kernel의 CV 성능이 가장 좋은가?

10. 지금까지 찾은 logistic regression, k-nn, svm model들에 대해서, **test set**에 대한 성능을 비교해보자. (최종 모델을 선택하기 위해 test set을 사용하는 것은 아니다. 단순히 세 가지 다른 모델의 성능을 비교하는 것이 목적이다.)

잘 정리된 report를 작성해야 하며, 아래의 파일들을 제출해야 합니다.

- **R markdown 파일 (code, comment, 답안 해설 포함)**
- **생성된 html 파일**