Multivariate Data Analysis Assignment #2

Logistic Regression

[Q1] 본인이 스스로 Logistic Regression 모델을 적용하는데 적합한 데이터셋을 선정하고 선정 이유를 설명하시오. 데이터셋 탐색은 아래에서 제시된 Data Repository를 포함하여 여러 Repository를 검색해서 결정하시오. 보고서에는 데이터를 다운로드할 수 있는 링크를 반드시 제공하시오.

- Kaggle Datasets: https://www.kaggle.com/datasets
- UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/index.php
- 공공 데이터 포털: https://www.data.go.kr/

[Q2] 해당 데이터셋의 종속변수와 설명변수는 어떤 것들이 있는가? 분석 전에 아래 세 가지 질문에 대해서 스스로 생각해보고 답변을 하시오.

1. 이 데이터에서 제공된 설명변수들 중에서 높은 상관관계가 있을 것으로 예상되는 변수들은 어떤 것들이 있는가? 왜 그렇게 생각하는가?
2. 제공된 설명변수들 중에서 종속변수를 예측하는데 필요하지 않을 것으로 예상되는 변수들은 어떤 것들이 있는가? 왜 그렇게 생각하는가?

[Q3] 모든 연속형 숫자 형태를 갖는(명목형 변수 제외) 개별 입력 변수들에 대하여 각각 다음과 같은 단변량 통계량을 계산하고 Box plot을 도시하시오: Mean, Standard deviation, Skewness, Kurtosis. 전체 변수 중에서 정규분포를 따른다고 할 수 있는 변수들은 몇 개인가? 정규분포를 따른다고 가정한 근거는 무엇인가?

[Q4] [Q3]의 Box plot을 근거로 각 변수들에 대한 이상치(너무 값이 크거나 작은 객체) 조건을 정의하고, 해당하는 객체들을 데이터셋에서 제거해 보시오.

다음 각 물음에 대해서는 [Q4]에서 제거된 객체들을 제외시킨 데이터프레임을 사용하여 답하시오.

[Q5] 가능한 모든 두 쌍의 입력 변수 조합에 대한 산점도(scatter plot) 및 Correlation plot (hint: "corrplot" 패키지의 corrplot( ) 함수 사용) 상관관계를 계산해 보시오.

1. 어떤 두 조합의 변수들이 서로 강한 상관관계가 있다고 할 수 있는가?
2. 강한 상관관계가 존재하는 변수 조합들 중에 대표 변수를 하나씩만 선택해서 전체 변수의 개수를 감소시켜 보시오 ([Q7]에서 사용함)

[Q6] 전체 데이터셋을 70%의 학습 데이터와 30%의 테스트 데이터로 분할한 후 모든 변수를 사용하여

Logistic Regression 모델을 학습해 보시오. 이 때 70:30으로 구분하는 random seed를 저장하시오([Q7]

1. 유의수준 0.05에서 유효한 변수의 수는 몇 개인지 확인하고 각 변수들이 본인의 상식 선에서 실제로 유효하다고 할 수 있는지 판단해 보시오.
2. [Q2-2]에서 정성적으로 선택했던 변수들의 P-value를 확인하고 해당 변수가 모델링 측면에서 실제로 유효하지 않는 것인지 확인해 보시오.
3. 학습 데이터와 테스트 데이터에 대한 Confusion Matrix를 생성하고 Simple Accuracy, Balanced Correction Rate, F1-Measure를 산출하여 비교해 보시오
4. 학습 데이터와 테스트 데이터에 대한 AUROC를 산출하는 함수를 직접 작성하고 이를 사용하여 학습/테스트 데이터셋에 대한 AUROC를 비교해 보시오.

[Q7] [Q5]에서 변수 간 상관관계를 기준으로 선택한 변수들만을 사용하여 [Q6]에서 사용한 학습/테스트 70:30분할 데이터으로 Logistic Regression 모델을 학습해 보시오.

1. 유의수준 0.05에서 유효한 변수의 수는 몇 개인지 확인하고 [Q6-1]의 결과와 비교하시오.
2. 학습 데이터와 테스트 데이터에 대한 Confusion Matrix를 생성하고 생성하고 Simple Accuracy, Balanced Correction Rate, F1-Measure를 산출한 뒤, [Q6-3]의 결과와 비교해 보시오.
3. 학습/테스트 데이터셋에 대한 AUROC를 산출하여 [Q6-4]의 결과와 비교해 보시오.

(Optional) [Q8] 이 외 본인이 선택한 데이터셋에 Logistic Regression을 통해 분석/검증해볼 수 있는 아이디어를 제시하고 이에 대한 절차와 분석 결과를 설명하시오.