Multiple Data Analysis Assignment #1

Multiple Linear Regression (MLR)

[Q1] 본인이 스스로 Multiple Linear 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 데이터는 종속변수와 설명변수들 사이에 실제로 "선형 관계"가 있다고 가정할 수 있겠는가? 가정할 수 있음/없음 판단에 대한 본인의 생각을 서술하시오.
2. 이 데이터에서 제공된 설명변수들 중에서 높은 상관관계가 있을 것으로 예상되는 변수들은 어떤 것들이 있는가? 왜 그렇게 생각하는가?
3. 제공된 설명변수들 중에서 종속변수를 예측하는데 필요하지 않을 것으로 예상되는 변수들은 어떤 것들이 있는가? 왜 그렇게 생각하는가?

[Q3] 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) 등을 도시하여 입력변수간 상관성에 대한 분석을 수행해 보시오. 어떤 두 조합의 변수들이 서로 강한 상관관계가 있다고 할 수 있는가? 이렇게 강한 상관관계가 발생한 변수들은 상식적으로도 상관관계가 높은 변수들이라고 할 수 있는가?

[Q6] 전체 데이터셋을 70%의 학습 데이터와 30%의 테스트 데이터로 분할한 후 모든 변수를 사용하여 MLR 모델을 학습한 뒤, Adjusted R^2값을 통해 데이터의 선형성(linearity)을 판단해 보시오. Residual plot과 Q-Q Plot을 도시하고 Ordinary Least Square 방식의 Solution이 만족해야 하는 가정들이 만족될만한 수준인지 정성적으로 판단해 보시오.

[Q7] 유의수준 0.01에서 모형 구축에 통계적으로 유의미한 변수들은 어떤 것들이 있는가? 해당 변수들은 종속변수와 양/음 중에서 어떤 상관관계를 갖고 있는가?

[Q8] Test 데이터셋에 대하여 MAE, MAPE, RMSE를 계산하고 그에 대한 해석을 해 보시오.

[Q9] 만약 원래 변수 수의 절반 이하로 입력 변수를 사용하여 모델을 구축해야 할 경우 어떤 변수들을 선택하겠는가? [Q5]와 [Q7]의 답변을 바탕으로 본인이 선택한 변수들에 대한 근거를 제시하시오.

[Q10] [Q9]에서 선택한 변수들만을 사용하여 MLR 모형을 다시 학습하고 Adjusted R^2, Test 데이터셋에 대한 MAE, MAPE, RMSE를 산출한 뒤, 두 모형(모든 변수 사용 vs. 선택된 변수만 사용)을 비교해 보시오.

[Extra Question] 이 외 해당 데이터셋을 통해 MLR 관점에서 가능한 추가적인 분석을 웹에서 검색해서 수행하고 그 결과를 해석해 보시오.