web-page-phishing.csv 데이터는 웹 페이지가 phishing 사이트인지 아닌지를 구분하는 데이터이며, 웹 페이지에 포함된 속성 정보를 포함하고 있다. 목표변수는 phishing이며 1이면 phishing 사이트이고, 0이면 phishing 사이트가 아니다. 각 속성은 다음과 같다.

컬럼	내용
url_length	The length of the URL
n_dots	The count of '.' characters in the URL
n_hypens	The count of '-' characters in the URL
n_underline	The count of '_' characters in the URL
n_questionmark	The count of '?' characters in the URL
n_equal	The count of '=' characters in the URL
n_at	The count of '@' characters in the URL
n_and	The count of '&' characters in the URL
n_exclamation	The count of '!' characters in the URL
n_space	The count of ' ' characters in the URL
n_tilde	The count of '~' characters in the URL
n_comma	The count of ',' characters in the URL
n_plus	The count of '+' characters in the URL
n_asterisk	The count of '*' characters in the URL
n_hastag	The count of '#' characters in the URL
n_dollar	The count of '$' characters in the URL
n_percent	The count of '%' characters in the URL
n_redirection	The count of redirections in the URL
phishing	The Labels of the URL. 1 is phishing and 0 is legitimate

다음과 같은 과정을 거쳐 phishing 인지 아닌지를 분류하는 모형을 만들어보자.

문제 1. 로지스틱 회귀분석을 통해 사이트가 phishing 인지 아닌지를 분류하는 모형을 만들고자한다. (40점)

- (1) 입력변수로는 목표변수인 phishing을 제외한 모든 변수를 사용하며, 모형을 만들기 위해 필요한 데이터 전처리 과정을 수행하라.
- (2) 로지스틱 회귀분석 모형을 생성하기 위해 전체 데이터를 학습집합과 테스트집합으로 나누어라. 무작위로 데이터를 나누며 테스트집합에는 전체집합의 20%가 포함되어야 한다.
- (3) 학습데이터로 로지스틱 회귀분석 모형을 학습하여라. 생성한 로지스틱 회귀분석 모형을 활용하여 테스트데이터에 대해 분류하고, 정확도를 f1 score로 측정하여라.
- (4) 5겹 교차검증을 활용하여 로지스틱 회귀분석 모형의 성과를 측정하고자 한다. 성과는 f1 score로 측정하며, 5겹 교차검증 성과의 평균치를 보여라. 로지스틱 회귀분석 모형에서 solver는 newton-cholesky로 지정하라.
- (5) 로지스틱 회귀분석 모형의 규제정도를 조절하는 하이퍼 파라미터인 C에 대해서 파라미터 튜닝을 실시하여라. 튜닝 방안은 적절이 선택하여라. C의 default 값은 1.0이며, 성과는 ROC_AUC로 측정하며 10겹 교차 검증을 통해 가장 성과가 좋은 C의 값을 선택하여라. (4)번과 마찬가지로 로지스틱 회귀분석 모형에서 solver는 newton-cholesky로 지정하라.
- C: 정규화 강도의 역수; 양의 실수여야 합니다. 서포트 벡터 머신(SVM)과 마찬가지로, 더 작은 값은 더 강한 정규화를 의미합니다. (Inverse of regularization strength; must be a positive float. Like in support vector machines, smaller values specify stronger regularization.)

In [ ]:

문제 2. 의사결정나무를 활용하여 사이트가 phishing 인지 아닌지를 분류하는 모형을 만들고자한다. (20점)

- (1) 문제 1에서 전처리된 학습집합을 활용하여 최대 나무 의사결정나무 모형을 만들어라. 테스트집합을 대상으로 예측하여 성과를 f1 score로 측정하라. 불순도지표는 default 값을 사용한다.
- (2) 의사결정나무 모형 학습 시 max_depth의 값을 변화시켜 가면서 성과가 가장 좋은 의사결정나무 모형을 찾고자 한다. max_depth의 값을 2에서 20까지 변화시켜 가면서 모형의 성과를 비교하라. 성과지표는 f1 score를 사용하며, (1)에서 사용된 학습집합으로 모형을 학습하고, 테스트집합으로 성과를 측정하라.
- (3) (2)에서 찾아진 가장 성과가 좋은 의사결정 나무 모형에서, 입력 변수 중요도를 파악하라.

In [ ]:

문제 3. (수치 예측) (35점)

insurance.csv는 미국의 의료보험 회사에서 고객에게 청구한 의료비용과 고객 정보를 포함하고 있다. 각 속성은 다음과 같다.

컬럼	내용
age	고객연령
sex	성별(male/female)
bmi	BMI 지수
children	자녀수
smoker	흡연여부(yes/no)
region	주거지역(northeast/southeast/southwest/northwest)
charges	보험사 청구비용

보험사 청구비용을 예측하는 모형을 다음과 같은 절차를 거쳐 만들려고 한다.

- (1) 종속변수는 charges이며, 나머지는 모두 독립변수로 활용한다. 범주형 변수는 원핫인코딩을 수행하고, 수치형 변수는 표준화를 수행하라.
- (2) 전체집합을 무작위로 20%를 테스트집합에 나머지를 학습집합으로 나누고, 학습집합을 활용하여 회귀분석 모형을 학습하라. 청구비용에 유의미한 영향을 미치는 변수는 무엇인가?
- (3) 회귀분석 모형으로 테스트집합을 예측하고, 성과를 MAPE로 측정하라.
- (4) Lasso 회귀분석 모형을 활용하여 청구비용을 예측하려고 한다. (2)에서 만든 학습집합만을 사용하여 alpha에 대한 하이퍼 파라미터 튜닝을 수행하라. 튜닝 방안은 적합한 방안을 사용하고, 10겹 교차검증을 실시하며 성과는 MSE로 측정하여 적합한 alpha를 찾아라. alpha는 L1 규제의 파라미터이다.
- (5) (4)에서 찾은 최적의 alpha 값을 활용하여 Lasso 모형을 학습집합을 활용하여 학습하고, 테스트집합에 대해 예측을 수행하여 성과를 측정하라. 성과는 MAPE로 측정한다.

In [ ]:

문제 4. 이 수업에서 개선되어야하는 사항은 무엇인가? (작성 시 5점, 미작성 시 0점)

In [ ]: