인사말) 1슬라이드

안녕하세요. 팀 maximum likelihood function의 김성연입니다.

목차) 2슬라이드

목차는 데이터 전처리 방식을 시작으로 어떻게 모델을 구성했는지, 결과는 어떻게 나왔는지 순으로 하겠습니다.

변수처리결과) 3슬라이드

신고번호와 신고일자는 단순 인덱스와 날짜변수임으로 제외하였습니다. 그 외에 줄이 그어진 반입보세구역부호등은 뒤에서 설명해드리겠습니다.

원-핫 인코딩) 4슬라이드

범주형 데이터의 전처리 방식으로 원-핫 인코딩과 레이블 인코딩이 있습니다. 원-핫 인코딩은 해당하는 단어일 때는 1을 넣고, 해당하지 않는 단어에서는 0이 들어가는 방식입니다. 단어의 종류에 따라 열 개수가 늘어납니다. 그러므로 단어의 종류가 많다면 열 개수가 많아집니다. 때문에 메모리 크기가 커지고 모델이 과적합 됩니다. 레이블 인코딩은 각각의 단어별로 0,1,2,3 순으로 숫자를 매기는 방식입니다. 그러므로 레이블 인코딩은 학년과 같이 숫자 크기가 의미 있을 때 사용하는 방식입니다. 하지만 분석할 변수가 대부분 부호를 나타내는 변수이기 때문에 데이터 전처리 방식으로 원-핫 인코딩을 사용하게 되었습니다. 다만 부호종류가 많은 변수는 제외를 하거나 줄이는 방향으로 과적합을 피해보겠습니다.

제외한 변수) 5-6슬라이드

제외한 변수는 반입보세구역부호, HS10단위부호, 적출국가코드, 원산지국가코드. 관세율 구분코드 입니다. 이 변수들은 부호종류가 많아 과적합이 됩니다. 실제로 이 변수들을 포함했을 때 f1값이 떨어지기 때문에 이 변수들을 제외하겠습니다.

포함한 변수-우범비율 순) 7슬라이드

다음으로 부호종류를 줄인 변수들입니다. 먼저 통관지세관부호 변수입니다. 맨 위 그림은 각 통관지별로 개수를 보여줍니다. 개수가 1000개 이하인 부호들은 표본 갯수가 너무 적어서 변동성이 심하므로 나머지 변수로 모두 묶어주겠습니다. 하늘색 그래프는 각 통관지별 우범비율을 나타냅니다. 1000개 이하에 부호들을 묶어준 분홍색 그래프를 보면 우범률이 부호별로 차이가 뚜렷히 난다는 것을 알 수 있습니다. 이렇게 개수와 우범률을 고려하여 서울, 인천공항, 인천항, 부산, 평택, 인천우편, 나머지 총 7개에 부호종류로 전처리를 해주었습니다.

포함한 변수-우범비율 순) 8-10슬라이드

다음 변수들도 통관지세관부호 변수와 마찬가지 방법으로 전처리를 하였습니다. 수입통관계획코드와 수입거래구분코드, 수입종류코드와 징수형태코드, 운송수단유형코드 역시 개수와 우범률을 고려해 전처리 해주었고 확실히 분홍색 그래프가 더 깔끔하게 정리된 것을 볼 수 있습니다.

포함한 변수-그대로 사용) 11슬라이드

수입신고구분코드는 애초에 우범비율이나 개수나 모두 잘 나와있던 변수였기 때문에 전처리 없이 그대로 사용했습니다.

포함한 변수-로그변형) 12슬라이드

이번에는 로그변형을 해준 변수들입니다. 먼저 신고인부호 변수입니다. 데이터형태를 보시는 것처럼 영어와 숫자가 섞여있어 종류가 모두 제각각인 형태입니다. 이전처럼 일일이 나누는 전처리를 하면 부호변수가 매우 많아 비효율적입니다. 대신에 몇 번이나 신고했는지 그 횟수를 변수로 취급해주었습니다. 예시를 들면 2o5a2 사용자부호가 1000개 있다면 2o5a2라는 변수이름 대신 1000이라는 숫자를 써 문자를 숫자형으로 바꿔준 것입니다. 이렇게 전처리한 이유는 거래량이 많은 부호일수록 우범비율이 감소하기 때문입니다. 수입자부호도 같은 맥락으로 전처리를 해주었습니다.

포함한 변수-로그변형) 13슬라이드

해외거래처부호와 특송업체부호는 방금 앞의 두 변수와는 공백인 칸이 있다는 점에서 차이점이 있습니다. 공백을 'No'라는 문자로 변경해준 후 'No'의 개수를 세어 'No'대신 그 개수를 집어넣어 전처리를 해주었습니다. 지금까지 설명한 네 변수 모두 거래량이 많은 부호일수록 우범비율이 감소합니다. 다만 거래량이 많다는 것이 2020년 데이터 기준이기 때문에 앞으로 이 모델을 사용한다면 지속적인 업데이트가 필요한 변수이겠습니다.

포함한 변수-로그변형) 14슬라이드

이제부터 설명할 변수들은 연속형 숫자로 값이 주어진 경우입니다. 범위가 그림과 같이 매우 컸기 때문에 로그변형으로 범위를 축소했습니다. 또 자료에서 이상치로 보이는 극단적인 값을 조정해주었습니다. 우측에 그림을 보시면 각 변수별로 위에 그림이 로그변형을 해주기 전, 아래 그림이 로그변형을 해준 후입니다. 원래 값은 아래 쪽에 많이 분포해있었으나, 로그변형 후 값이 이전보다는 골고루 퍼져있는 것을 볼 수 있습니다.

포함한 변수-로그변형) 15슬라이드

관세율 또한 신고중량과 원화금액 변수와 마찬가지로 전처리를 해주었고 역시 로그변형을 해준 후 골고루 값이 퍼져있음을 보실 수 있습니다.

[oversampling 기법] – 1

다음으로 오버샘플링 기법에 대해 설명하겠습니다. 훈련데이터셋에서 우범화물수가 비우범인 화물수 보다 많습니다. 만약 트리 기반 모델을 사용하면, 작은 데이터 셋인 우범화물에 대해서 리프노드를 형성하지 못할 수 있습니다. 그렇게 되면 정확도는 높지만 재현율이 낮은 현상이 발생할 수 있습니다.

[oversampling 기법] - 2

왜냐하면 원 데이터셋에서 비우범수가 많기 때문에 대부분 비우범으로 예측하게 되기 때문입니다. 따라서 우범화물을 정확하게 예측하는 비율인 재현율이 상당히 낮아지는 문제점이 발생합니다.
일반적으로 이와 같은 현상을 방지하기 위해 우범화물인 데이터를 중복하거나 약간의 변동을 주어 데이터 셋을 늘리는 oversampling 기법을 적용합니다.
oversampling 기법 중 많이 사용되는 SMOTE기법은 소수데이터에 변동을 주어 데이터를 근처 데이터를 만들어냅니다. 다만 대부분에 변수를 원핫인코딩하여 열의 갯수가 늘어나 고차원이 된 저희 모델 경우 SMOTE기법이 데이터의 노이즈를 증가시킵니다. 그러므로 단순히 소수 데이터를 복제하는 랜덤 오버 샘플링방법을 사용하겠습니다.

[oversampling 기법] – 3

실제로 데이터에 적용해보아도 랜덤 오버 샘플링을 이용했을 때 F1 값이 더 높은 모습을 보입니다.

[light GBM] - 1

Light GBM은 뛰어난 앙상블 학습으로 평가받는 히스토그램기반 그레이디언트 부스팅 알고리즘입니다. 빠른 속도를 인정받아 최근 경진 대회에서 xgboost와 함께 많이 사용됩니다. ppt에 쓰인대로 다양한 장점이 있기 때문에 저희 모델 분석에 Light GBM방식을 사용하겠습니다.

[light GBM] - 2

Light GBM의 하이퍼 파라미터입니다. 대부분 모델에 과적합을 방지하기 위해 사용합니다. 하이퍼 파라미터를 다양하게 조정한 결과 다음 값에서 F1값이 가장 높게 나와 사용하게 되었습니다.

[결과 및 결언] - 1

훈련 데이터에 대해서 모델을 적용해본 결과 예측도가 0.65로 테스트 데이터에 예측도 0.639와 크게 차이가 나지 않습니다. 이는 모델이 과적합되지 않았다는 것을 의미합니다.

[결과 및 결언] - 2

모델을 적합한 결과 실제 우범비율과 Test 셋에 예측한 우범비율이 비슷할 것이라 생각했습니다. 하지만 저희 모델이 우범을 더 많이 예측하는 것으로 나오게 됩니다. 즉 공격적으로 우범임을 주장하는 모델입니다. 우범화물 후보를 찾아 정밀검사를 하는 것이 모델에 목표인데 잘 부합한다고 할 수 있습니다.