EURON 5주차 예습과제_정윤지

4.9 캐글 산탄데르 고객 만족 예측

캐글 산탄데르 고객 만족 예측
- 370 개의 피처로 주어진 데이터 세트 기반에서 고객 만족 여부를 예측하는 것
- 피처 이름은 익명처리 돼, 어떤 속성인지 추정이 불가함
- 클래스 레이블 명은 TARGET이며, 이 값이 1이면 불만을 가진 고객, 0이면 만족한 고객임
- 모델의 성능 평가는 ROC-AUC로 평가함(대부분 만족이고, 불만족인 데이터는 일부일 것이기 때문에 정확도 수치보다 더 정확함)

데이터 전처리
- 새로운 주피터 노트북 생성
- 내려받은 train_santander.csv 파일을 노트북이 생성된 디렉터리로 이동
- 필요한 모듈을 로딩, 학습 데이터를 데이터프레임으로 로딩
- 111개의 피처가 float 형, 260개의 피처가 int 형으로 모든 피처가 숫자형
- Null 값은 없음
- 대부분이 만족이며, 불만족인 고객은 4%에 불과함
- Var3 칼럼의 경우 min 값이 -999999
- Var 3은 다른 값에 비해 편차가 심하므로 가장 값이 많은 2로 변환
- ID 피처는 단순 식별자에 불과하므로 피처를 드롭
- 클래스 데이터 세트와 피처 데이터 세트를 분리해 별도의 데이터 세트로 저장
- 학습과 성능 평가를 위해 원본 데이터 세트에서 학습 데이터 세트와 테스트 데이터 세트를 분리
- 비대칭 데이터 세트이므로 target 값 분포도가 학습 데이터와 테스트 데이터 세트에 비슷하게 추출됐는지 확인

XGBoost 모델 학습과 하이퍼 파라미터 튜닝
- XGBoost의 학습 모델을 생성, 예측 결과를 평가
- 성능 평가 기준이 ROC-AUC 이므로 eval_metric = 'auc'
- HyperOpt를 이용해 베이지안 최적화 기반으로 XGBoost의 하이퍼 파라미터 튜닝을 수행
- 목적 함수는 3 Fold 교차 검증을 이용해 평균 ROC-AUC 값을 반환하되, -1을 곱해주어 최소 반환값이 되도록 함
- max_eval = 50만큼 반복하면서 최적화 하이퍼 파라미터를 도출
- 도출된 하이퍼 파라미터를 기반으로 XGBClassifier를 재학습, 테스트 데이터 세트에서 ROC-AUC를 측정
- XGBoost는 GBM을 기반으로 하기 때문에, 수행 시간이 상당이 많이 요구됨
- XGBoost의 예측 성능을 좌우하는 가장 중요한 피처는 var38, var15

LightGBM 모델 학습과 하이퍼 파라미터 튜닝
- LightGBM으로 학습을 수행하고, ROC-AUC를 측정
- XGBoost보다 학습에 걸리는 시간이 단축됨
- HyperOpt를 이용하여 다양한 파라미터에 대한 튜닝을 수행
- LightGBM의 경우 학습 시간이 상대적으로 빠름

4.10 분류 실습 - 캐글 신용카드 사기 검출

캐글 신용카드 사기 검출 분류 실습
- 데이터 세트의 레이블인 Class 속성은 매우 불균형한 분포를 가짐
- 전체 데이터의 약 0.172% 만이 레이블 값이 사기 트랜잭션임
- 일반적으로 사기 검출이나 이상 검출과 같은 데이터 세트는 이처럼 레이블 값이 극도로 불균형한 분포를 가지기 쉬움(사기와 같은 이상 현상은 전체 데이터에서 차지하는 비중이 매우 적기 때문)

언더 샘플링과 오버 샘플링의 이해
- 레이블이 불균형한 분포를 가진 데이터 세트를 학습시킬 때 예측 성능의 문제가 발생할 수 있음
- 지도학습에서 극도로 불균형한 레이블 값 분포로 인한 문제점을 해결하기 위해 적절한 학습 데이터를 확보하는 방안이 필요: 오버 샘플링, 언더 샘플링 방법이 있음
- 오버 샘플링 방식이 예측 성능 상 유리한 경우가 많아 상대적으로 더 많이 사용됨

- 언더 샘플링: 많은 데이터 세트를 적은 데이터 세트 수준으로 감소시키는 방식, 정상 레이블의 경우 제대로 된 학습을 수행할 수 없는 문제가 발생할 수 있으므로 유의해야 함
- 오버 샘플링: 이상 데이터와 같이 적은 데이터 세트를 증식하여 학습을 위한 충분한 데이터를 확보하는 방법, 동일한 데이터의 단순한 증식은 과적합이 되기 때문에 원본 데이터의 피처 값들을 아주 약간 변경하여 증식, 대표적으로 SMOTE 방법이 있음
- SMOTE: 적은 데이터 세트에 있는 개별 데이터들의 K 최근접 이웃을 찾아, 이 데이터와 K개 이웃들의 차이를 일정 값으로 만들어, 기존 데이터와 약간 차이가 나는 새로운 데이터를 생성하는 방식
- SMOTE 방식은 imbalanced-learn 패키지로 구현이 가능

데이터 일차 가공 및 모델 학습/예측/평가
- 새로운 주피처 노트북을 생성, 다운로드 받은 파일을 동일한 디렉터리로 이동시킨 후, 데이터프레임으로 로딩
- Time 피처의 경우 데이터 생성 관련 작업 용 속성으로 의미가 없기 때문에, 제거함
- 다양한 데이터 사전 가공을 수행하고, 이에 따른 예측 성능도 비교
- get_train_test_dataset()를 통해 학습 피처/레이블 데이터 세트, 테스트 피처/레이블 데이터 세트를 반환
- 생성한 학습 데이터 세트와 테스트 데이터 세트의 레이블 값 비율을 백분율로 환산하여 분할 정도를 확인
- 모델 생성 시, 로지스틱 회귀와 LightGBM 기반의 모델이 데이터 가공을 수행하면서 예측 성능의 변화를 알아볼 것
- 로지스틱 회귀를 적용하면 재현율이 0.6216, ROC-AUC가 0.9702
- LightGBM을 적용하면 재현율이 0.7568, ROC-AUC가 0.9790으로 로지스틱 회귀보다 높은 수치를 나타냄

데이터 분포도 변환 후 모델 학습/예측/평가
- 왜곡된 분포도를 가지는 데이터를 재가공한 뒤, 모델을 재 테스트함
- Creditcard.csv의 중요 피처 값의 분포도
- 로지스틱 회귀는 선형 모델이며, 선형 모델은 중요 피처들의 값이 정규 분포 형태를 유지하는 것을 선호함
- Amount 피처의 분포도는 꼬리가 긴 형태의 분포 곡선을 가지고 있음
- Amount를 표준 정규 분포 형태로 반환한 뒤, 로지스틱 회귀의 예측 성능을 측정
- 정규 분포 형태로 amount 피처 값을 변환한 후 테스트 데이터 세트에 적용한 로지스틱 회귀의 경우, 정밀도와 재현율이 저하되었고, LightGBM의 경우 약간의 저하가 있지만 성능 상의 큰 차이는 없음
- 로그 변환을 수행
- 로그 변환은 데이터 분포도가 심하게 왜곡되어 있을 경우 적용하는 중요 기법
- 기존 값을 log 값으로 변환해 원래 큰 값을 상대적으로 작은 값으로 변환하기 때문에 데이터 분포도의 왜곡을 상당 수준 개선해줌
- 이후 다시 로지스틱 회귀를 적용하면 원본 데이터 대비 정밀도는 향상되었지만 재현율은 저하됨
- LightGBM의 경우 재현율이 향상됨
- 레이블이 극도로 불균일한 데이터 세트에서 로지스틱 회귀는 데이터 변환 시 불안정한 성능 결과를 보여줌

이상치 데이터 제거 후 모델 학습/예측/평가
- 이상치 데이터는 전체 데이터의 패턴에서 벗어난 이상 값을 가진 데이터이며, 아웃라이어라고도 불림
- 이상치를 찾는 방법 중 하나는 IQR(Inter Quantile Range) 방식
- IQR은 사분위 값의 편차를 이용하는 기법으로 Box Plot 방식으로 시각화가 가능
- 사분위: 전체 데이터를 값이 높은 순으로 정렬하고, 이를 35%씩 구간을 분할하는 것(Q1, Q2, Q3, Q4)
- IQR을 이용해 이상치 데이터를 검출하는 방식은 IQR에 1.5를 곱해, 생성된 범위를 이용하여 최댓값과 최솟값을 결정한 뒤, 최댓값을 초과하거나 최솟값에 미달하는 데이터를 이상치로 간주함
- 박스 플롯은 사분위의 편차와 IQR, 이상치를 나타냄
- 이상치 데이터를 IQR을 이용해 제거하려면, 어떤 피처의 이상치 데이터를 검출할 것 인지 선택이 필요
- 결정값과 가장 상관성이 높은 피처들을 위주로 이상치를 검출하는 것이 좋음
- 넘파이의 percentile()을 이용해 1/4 분위와 3/4 분위를 구하고, 이에 기반해 IQR을 계산, 1.5를 곱해서 최댓값과 최솟값 지점을 구한 뒤, 최댓값보다 크거나 최솟값보다 작은 값을 이상치로 설정, 이상치가 있는 dataframe index를 반환
- 이상치 데이터를 제거한 뒤, 로지스틱 회귀와 LightGBM 모두 예측 성능이 크게 향상됨

SMOTE 오버 샘플링 적용 후 모델 학습/예측/평가
- SMOTE를 적용할 때, 학습 데이터 세트만 오버 샘플링을 진행해야 함
- SMOTE 적용 후 2배 정도로 데이터가 증식됨
- 생성된 학습 데이터 세트를 기반으로 로지스틱 회귀 모델을 학습한 뒤 성능을 평가하면, 재현율이 증가하지만, 정밀도가 저하됨

- 분류 결정 임곗값에 따른 정밀도와 재현율 곡선을 통해 SMOTE로 학습된 로지스틱 회귀 모델에 어떠한 문제가 발생하고 있는지 시각적으로 확인
- 분류 결정 임계값을 조정하더라도 임계값의 민감도가 너무 심해, 올바른 재현율/정밀도 성능을 얻을 수 없으므로 로지스틱 회귀 모델의 경우 SMOTE 적용 후 올바른 예측 모델이 생성되지 못함
- LightGBM의 경우, 재현율은 높아지나, 정밀도는 낮아짐
- 재현율 지표를 높이는 것이 머신러닝 모델의 주요한 목표인 경우 SMOTE를 적용하면 좋음