Team 이지은\_조성혁\_김효정 활동 소개

|  |  |  |  |
| --- | --- | --- | --- |
| 일시 | 2020년 4월 13일 (월) ~ 2020년 4월 22일 (수) | | |
| 작성일 | 2020년 4월 23일 (목) | 작성자 | 조성혁 |
| 장소 | 더조은컴퓨터아카데미 강남 별관 1103호 | | |
| 팀원 | 이지은, 조성혁, 김효정 | 과정명 | 응용 SW 기초 기술활용 |
| 구분 | H백화점에 방문하는 고객의 성별을 예측하는 프로그램 만들기 | | |
| 활동사항 | 1. 데이터 확인  * csv형식으로 된 train.csv,test.csv 을 다같이 확인 해봤을 때 우선 계절별 시긴대별 구매건수열에 결측값(None Of Number)이 다수 보였고, jupyterlab을 활용해 데이터의 자세한 정보를 시각화 해본 결과, 남녀 셩별간의 데이터양이 여성쪽으로 두 배 이상 치우쳐 있었고, 아침에 구매한 건수, 주 구매코너를 제외한 모든면에서 + - 5% 이하의 일치율으로 남녀간의 상반된 데이터를 도출하기에 애를 먹었다.  1. 데이터 전처리(결측값 처리,모델링)  * 이와 같은 문제점을 해결하기 위해 1) 계절별,시간대별 구매건수의 결측값을 구매하지않은(0건) 경우로 판단하고 모두 0으로 대체하여 채웠다. 2) 여성쪽으로 두 배 이상 과적합된 데이터를 보완하기 위해 SMOTETomek(under sampling + over sampling)모듈의 복합샘플링기법을 이용해 남녀간의 데이터 비율 차이를 좁혀줬다. 3) 상반된 데이터를 도출해내기 위해 matplotlib 에 내장되어 있는 그래프를 활용해 아침 구매건수에서 약간의 빈도 차이를 발견하고 주구매코너에서 화장품(남:25.5%,여:74.5%),일반식품(남:34.6%,여:65.4%),스포츠(남:40.2%,여:59.8%)으로 남녀비율차이(약2.3배)를 고려 했을 때 화장품은 여성쪽으로 치우치고, 일반식품과, 스포츠는 남성쪽으로 조금 더 치우친 상반된 데이터를 발견했다.  1. Feature Engineering  * 여기에선 모델의 성능을 올리기 위해 다양한 파생변수를 만들어 봤지만 아침 구매건수를 활용하여 기존의 데이터들과의 관계를 사칙연산을 통해 만들어 냈을 때 제일 좋은 성능을 보였다.  1. 아침구매가격(평균) : 아침구매건수 X 평균구매가격 = 아침에 구매한 총 건수에 대한 평균적인 구매가격(ex: 총 구매액:270,000원 총 구매건수:9건,아침구매건수:3건, 평균구매가격:30,000원, 3 X 30000 = 90000, 하지만 한 번 구매할 때마다 정확히 비용이 얼마나 들었는지는 자료에 없기 때문에 알 수 없지만 예측 알고리즘이 조금 더 예측에 도움이 될 수 있도록 여기서는 평균적인 값으로 대체하여 파생변수를 만들었다.) # 만약 아침구매건수 값이 0이면 아침구매가격을 0으로대체함 2. 아침구매비율 : 아침구매건수/ (총)구매건수 = 총 구매건수에 대한 아침구매 비율(ex: 아침구매건수:6건, (총)구매건수:10건, 6 / 10 = 0.6, 총 구매건수에 대한 아침구매비율은 60%이다.) # 만약 아침구매건수 값이 0이면 아침구매비율을 0으로대체함 3. 아침-점심 구매 가격(평균) : (아침구매건수 + 점심구매건수) X 평균구매가격 = 아침과 점심에 구매한 건수에 대한 평균적인 구매 가격(아침구매가격과 마찬가지로 정확한 데이터가 없기 때문에 평균적인 값으로 대체함, 아침구매x건-점심구매0건인 경우는 아침에 구매한 평균가격, 아침구매0건-점심구매x건인 경우는 점심에 구매한 평균가격이 된다. 아침구매가격과 겹치는 값이 존재하지만 점심구매가격,아침과 점심동시에 구매한 가격을 만들어내는 대에 의미를 두고 겹치는 값은 무시했음) # 아침구매&점심구매 건수 값이 둘다 0이면 아침-점심 구매 가격을 0으로 대체함 4. 아침-점심 구매 건수 : 아침구매건수 + 점심구매건수 = 아침과 점심에 구매한 건수 # 아침구매&점심구매 건수 값이 둘다 0이면 아침-점심 구매건수를 0으로 대체함 5. 아침-점심 구매 비율 : (아침구매건수 + 점심구매건수) / (총)구매건수 = 총 구매건수에 대한 아침과 점심에 구매한 비율(아침구매비율과 원리는 같다.) # 아침구매&점심구매 건수 값이 둘다 0이면 아침-점심 구매비율 값을 0으로 대체 6. 아침한정 내점구매율 : (구매건수/내점일수) / 구매건수 X 아침구매건수= 내점구매율에 대하여 아침에 구매한 건수의 비율(ex: (총)구매건수:9,(총)내점일수:8일, 아침구매건수:3건, (9/8)/9X3 = 0.375, 하루 내점 할 때마다 아침에 구매하는 비율이 37.5%이다.) # 아침구매건수 값이 0이면 아침 내점구매율 값을 0으로 대체함 | | |
| 종합 결과 | 모델: LightGBM, <classification\_report - accuracy : 0.786, precision : 0.77,0.80, recall : 0.81,0.76, f1 score : 0.79,0.78> | | |