프로젝트 결과

1. 주제

실제 고객 데이터 기반 은행 마케팅 비용 감축 최적 모델 선정

* 필요성 제시에서 은행의 마케팅 방식에 ~가 있다. (소개)를 진행하고 여기서 비용 감축의 중요성을 환기시킨다.
* 데이터 전처리 부분 소개
* 대략적 데이터 분포 및 데이터 선정 과정 소개
* 편향된 데이터라는 것은 밝히고 시작
* 어떤 모델을 선택해서 파라미터를 어떻게 결정했는지 과정 소개
* 최소의 비용으로 최대의 효율을 만들어 내기 위해서
* 회계적 비용, 고객의 부정적 경험으로 발생하는 비용이 줄일 수 있었다. 하지만
* 비용 감축에는 효과적이었다.
* 한계점 제시를 통해서 최대한 데미지를 줄여봅시다.
* 낮은 f1\_score, 재현율: 원데이터의 편향 때문에 각종 샘플링 방식을 적용해도 근본적인 해결 불가

1. 요약

Kaggle의 실제 포르투갈 은행 마케팅 데이터(정확히는 UCI의 머신러닝 데이터 셋) 41187개를 기반하여

decision tree, random forest, DNN, XGBoost, LightGBM, KNN 방식을 이용하여 최적의 예측 모델을 선정하고자 했다. 해당 데이터에 대해 ‘unknown’으로 표시된 값에 대해 결측치 처리를 한 경우와 일종의 데이터로 취급한 경우의 2가지 종류로 나눠서 결과를 확인했다.

결과적으로 두 데이터에서 학습 결과가 눈에 띄는 차이를 보여주지 않았다. 샘플링을 진행하지 않은 모델의 성능은 전반적으로 정확도는 86~89%, 정밀도는 random forest와 decision tree ,knn 방식을 제외한 경우 59~64% 수준을 보여 주었다. 하지만 재현율 부문에서 18~20% 수준을 보여주었다.

해당 현상에 대해서 실제 데이터의 종속변수의 편향이 심하기 때문이라 판단하여(성공: 약 12%, 실패: 88%) 다양한 샘플링 기법을 적용하여 결과치를 확인했다. under sampling 방식에서는 재현율이 39~58% 사이의 수치를 보여주었지만 정확도가 전체적으로 61~80% 사이의 수치로 낮아졌다. 또한 정밀도가 16~27% 사이 수준으로 크게 하락하여 해당 방식을 기각하였다.

Over sampling 방식을 적용한 결과 정확도는 70~80% 수준을 보여주고 재현율도 30~50% 대로 향상되었지만 정작 정밀도가 17~30% 수준으로 크게 하락하여 해당 방식을 기각하였다.

Over sampling 방식 중 smote 방식을 적용한 경우 정확도가 87~89% 수준을 보여주었고 random forest와 decision tree ,knn 방식을 제외한 경우 정밀도가 61~65% 사이의 값을 보여주었고 randomforest, decisontree, knn에서 정밀도는 37~43% 수준을 보여주었다. 해당 방식에서 재현율은 18~30% 사이의 값을 보여주었다.

해당 결과에 대해서 관련 논문을 다시 확인한 결과 해당 논문에서는 원 데이터에 대해서 일종의 언더 샘플링 기법을 적용하여 종속변수의 비율을 1:1 수준으로 맞춘 후 해당 데이터를 기반으로 모델의 성능을 평가하였다. 해당 방식은 좋은 성능을 보여주었다. 하지만 프로젝트의 목표인 실제 고객 데이터 기반의 마케팅 성공 예측이라고 볼 수 없었기에 해당 방식은 기각했다.

모델의 성능 평가 지표로 정밀도로 선정했다. 은행의 고객 마케팅에서 실제 가입하지 않을 것인데 마케팅을 진행해 발생하는 고객의 부정적 경험이 가입할 가능성이 있지만 모델에서 걸러져서 홍보하지 못해 발생하는 손실보다 크다고 판단하였다.

이런 기준에 따라서 모델의 평가 결과 ‘unknown’ 값을 데이터로 처리한 경우에서 샘플링을 거치지 않은데이터를 기반으로 훈련한 최적 파라미터를 적용한 XGBoost모델의 성공 예측 분류 모델이 최적의 모델로 선정되었다.

해당 프로젝트 결과는 높은 정확도와 기존 마케팅 성공률 대비 3배 가까이 상승시켜주지만 재현율이 너무 낮아서 일반적인 마케팅 상황에 적용하기 힘들 것으로 생각된다. 이는 학습 데이터 종속변수의 편향성으로 인해 발생한 문제로 예상된다. 해당 문제를 해결하기 위해 다양한 샘플링 방식을 적용하여 학습한 모델에 대해서 실제 데이터를 예측한 결과 under sampling에서 재현율이 높아지긴 했지만 모델 성능 평가의 주요 지표인 정밀도가 큰 폭으로 감소하여 모델 예측 기반 마케팅 성공율이 이전보다 떨어지는 현상이 발생했다.

1. 데이터 선정

프로젝트의 데이터는 어바인 캘리포니아 대학교(UCI) 머신러닝 연구소의 포르투갈 은행 마케팅 데이터 선정했다. 해당 데이터는 고객 데이터, 실제 마케팅 데이터, 사회경제적 속성 데이터, 기타 속성, 실제 가입 여부로 데이터를 분류하고 있다. 여기에서 우리 프로젝트는 고객 데이터 기반 은행 마케팅 비용 감축 최적 모델 선정이기 때문에 실제 마케팅 이후에 알 수 있는 데이터인 실제 마케팅 데이터를 제외했다. 추가적으로 사회 경제적 속성의 경우 고객 데이터와 거리가 있기 때문에 제외하고 모델 학습을 진행하고자 한다.

1. EDA

데이터의 탐색적 분석 결과 다음과 같은 사항에 대해서 알 수 있었다.

* 고객 연령

은행 마케팅은 30대, 40대 20대 50대 순으로 많이 진행했고 이는 은행 마케팅이 경제활동이 활발한 인구층을 대상으로 진행된다는 사실을 알 수 있었다.

실제 가입 수는 30대, 20대, 40대, 50대 순서로 많았고

마케팅 대비 가입 확률은 주력 마케팅층에서 가장 적었다.

* 직업

직업별 마케팅은 관리직, 숙련노동자, 기술자, 서비스업 종사자 순서로 많이 진행했다.

실제 가입자는 관리직, 기술자, 숙련노동자, 은퇴자 순서로 나타났다

가입 확률은 학생인 경우 45% 수준을 보이고 은퇴자인 경우 33%, 나머지 경우는 15~7% 수준을 보였다.

* 혼인

혼인 여부에 따른 마케팅을 분류한 경우 기혼자를 대상으로 한 마케팅이 전체 60% 수준을 차지했다.

실제 가입자는 기혼자인 경우가 가장 많았고 미혼, 이혼, 알 수 없음 순서로 나타났다.

실제 가입 확률은 결혼 여부를 밝히지 않은 고객 17%로 가장 많았고 그 다음 미혼인 경우가 16%로 나타났다. 이혼인 경우와 혼인한 경우 11%대의 가입 확률을 보였다.

알 수 없음은 모수 자체가 매우 적기 때문에 가입 확률이 눈에 띄게 커 보인다.

* 교육수준

교육수준에 따른 마케팅을 분류한 경우 대졸자 대상의 마케팅이 가장 많았고 고졸자,박사, 중졸자 순으로 많이 나타났다.

실제 가입자는 마케팅 횟수와 거의 비례하여 대졸자, 고졸자, 박사, 중졸자 순서로 많았다.

실제 가입 확률 부문을 보면 대졸자가 7.7%로 나타나고 우리나라로 치면 초등학교 중퇴자의 비율이 7.4%로 두 번째로 나타나는데 이는 해당 모수 자체가 대졸자의 23% 수준으로 나타났기 때문으로 생각된다. 이 경우를 제외하고 고졸자가 6.9%, 박사가 6.3%로 마케팅 성공확률이 교육수준에 거의 비례하는 것을 알 수 있었다.

* 파산 여부

파산 여부에 따른 마케팅을 분류한 결과 전체 4만 여건의 마케팅에서 단 3명만 파산 경험이 있었고 이들은 모두 정기적금에 가입하지 못했다.

실제 가입자는 파산 경험이 없거나 관련 자료가 없는 경우 순서로 많았다.

가입 확률은 파산 경험이 없는 경우 14% 수준, 알 수 없는 경우 5.4% 정도인 것이 확인되었다.

* 주택 담보대출 보유 여부

주담대 보유 여부에 따른 마케팅을 분류한 결과 보유한 경우가 약 2.1만건, 없는 경우가 1.8만건으로 각각 전체의 52%, 45%를 차지하고 있어 주담대 보유 여부는 실제 마케팅 진행에 큰 영향이 없는 것으로 확인되었다.

실제 가입자 수는 위와 같이 주담대 보유, 주담대 없음, 알수 없음 순서로 발생했고 가입 확률은 12~13% 내에 있어 실질적으로 주담대 보유 여부는 은행 마케팅에 영향력이 매우 제한적인 것으로 확인 되었다.

* 개인 신용 대출 보유 여부

신용 대출 보유 여부에 따른 마케팅을 분류한 결과 없는 경우가 약 3.4만, 있는 경우가 6천여건으로 실제 대출 보유에 따른 마케팅 진행에 차이를 보였다.

가입자 수는 해당 결과에 비례하여 나타났다. 그러나 가입 확률을 확인하면 대출 보유 여부와 정보를 알 수 없는 경우를 모두 포함하여 12%대의 가입 확률을 보여주어 개인 신용 대출은 마케팅 성공에 미치는 영향이 매우 제한적인 것을 확인할 수 있었다.

* 마케팅 성공 여부

전체 마케팅 시도 중 성공으로 이어진 경우가 12%, 실패로 이어진 경우가 88%로 실패에 매우 편향된 데이터인 것을 확인 할 수 있었으며 해당 데이터로 분류 모델을 만들 경우 그 성능이 좋지 않을 것으로 예상된다.

1. 데이터 구성

* 전체 데이터 세트에 대해서 ‘unknown’으로 처리된 데이터에 대해서 다음 2가지로 분류하여 모델의 성능을 비교하고자 한다.
* 결측치를 처리하여 해당 값이 포함된 데이터의 열을 제거

-해당 처리 결과 30488행의 데이터 세트가 만들어졌다.

* 일종의 데이터로 취급하여 인코딩을 같이 진행
* 학습 데이터, 검증 데이터의 비율은 7:3의 비율로 진행
* 데이터의 편향이 심하기 때문에 샘플링 방식에 따라 다음 4가지로 학습 데이터를 구성하여 최적의 모델을 선정하고자 한다.
  + - * 샘플링 작업을 하지 않는 원상태의 데이터
      * 랜덤 언더 샘플링
      * 랜덤 오버샘플링
      * SMOTE(Sythetic Minority Over-sampling Technique)

1. 모델 종류

* 최초 계획은 생각 가능한 모든 분류 모델에 대해서 별도 파라미터 조정을 하지 않은 경우와 GridSerarchCV를 통해 얻어진 최적 파라미터로 구성된 모델 중을 최적의 모델을 선정하고자 했으며 그 종류는 다음과 같다.
  + - * Decision Tree
      * Random Forest
      * XGBoost
      * GBM
      * LightGBM
      * KNN
      * 딥러닝 방식
* 여기서 조정 가능한 모든 파라미터를 대상으로 GridSearchCV를 진행하고자 했으나 컴퓨터의 한계로 일부 파라미터에 대해 지정된 값들 중에서 최적의 파라미터를 찾았으며 Decision Tree, Random Forest, GBM, KNN의 경우 컴퓨터 성능상 한계로 GridSearchCV가 작동하지 않아 이를 제외하고 일부 모델에 대해 다음 제시되는 일부 파라미터 리스트를 적용하여 GridSearchCV를 진행했다.
  + - * LightGBM ={ 'max\_depth':[-1,5,10,15,20],

'learning\_rate':[0.01,0.05,0.1,0.3,0.5] }

* + - * XGBoost = { n\_estimators=[100,200,300,400,500],

learning\_rate = :[0.01,0.05,0.1,0.3,0.5] ,

max\_depth =[ 5,10,15,20

] }

* 그 결과 다음과 같은 최적 파라미터가 선정 되었다.
  + - * LightGbm = {'max\_depth':20, ‘learning\_rate’=0.05}
      * XGBoost = {‘n\_estimator’:400, learing\_rate = 0.05, max\_depth=5}
* 딥러닝의 경우 각각 64개의 노드를 갖는 9개의 은닉층을 갖는 구조로 구성되어 있다.

1. 모델 성능 평가

* 모델 성능 평가의 지표로 1차적으로 정밀도를 선정했다.
* 이유:

1. 고객 데이터를 통해서 예측된 결과값을 기반으로 마케팅에 활용되기 때문이다.
2. 가입하지 않을 고객에게 마케팅을 진행해서 고객이 얻는 부정적 경험으로 인한 손실이 가입 할 수 있지만 놓쳐서 발생하는 손실보다 더 크다고 생각하기 때문이다.

* 그렇다고 해서 precision만을 고려하는 경우 모델이 성공으로 예측하는 전체적인 숫자 자체가 매우 적어질 것으로 예상한다.
* 그렇기에 재현율에 56%의 가중치를 부여하여 모델을 평가하고자 한다.
* 최종 평가 지표 = 정밀도(precision) + 재현율(recall)\*56%
* 이 부분은 자료를 통해서 제시 할 수 있으면 해당 내용을 참고하여 설명하고 그렇지 않으면 우리가 임의로 가정했다는 사실을 밝혀야 한다.

1. 모델 결과 분석
2. 데이터 종류(‘unknown’값의 처리 방식에 따른 구분)
   * + - 눈에 띄는 큰 차이를 보이지는 못함
3. 모델에 따른 성능 비교
   * + - 전반적으로 Decision tree, Random forest, knn 방식의 성능이 떨어지는 것이 눈에 띄고 나머지는 큰 차이를 보이지는 않았다.
4. 샘플링 방식에 따른 성능 비교
   * + - 랜덤 샘플링 방식을 적용했을 때, 전반적으로 모델들의 재현율이 상당히 개선되었으나 나머지 지표가 큰 폭으로 하락하는 모습을 보여 주었다.
       - SMOTE 방식에서 일부 모델의 지표의 개선이 보여졌다. 하지만 이것이 전반적인 모델의 성능 향상으로 이어지지는 않았다.
5. 추후 진행

* 해당 선정 모델을 기반으로 고객 마케팅 예측 파일을 만들어서 고객에 대한 raw 데이터를 넣으면 마케팅 성공 예측 고객의 신상을 csv 파일로 뽑아내는 시도를 하고자 한다.