<1>

안녕하십니까? 이번에 고객 데이터 기반 은행 마케팅 비용 감축 최적 모델 선정을 주제로 발표를 하게 된 연방 준배은행 의장 박준배라고 합니다.

<2>

발표 순서는 다음과 같이 진행 될 예정입니다.

<3>

팀 구성 및 역할입니다. 저는 데이터 수집 및 분석, 자료 수집, 최종 발표를 맡았습니다. 이어서 조해원 이사님은 금융 도매인 및 데이터 수집, 김민성 이사님은 경영 도메인 및 회의록, 일지 작성을 맡아주셨습니다. 박희연 이사님은 수학, 통계 도메인 및 ppt 제작을 맡아 주셨고 이명희 이사님은 중간 발표 및 자료 수집을 맡아주셨습니다.

<4>

프로젝트 개요입니다. 최근 은행의 DT가 활발하게 이뤄지면서 이에 따라 은행의 고객 데이터 기반 마케팅이 떠오르고 있는 상황입니다.

<5>

이런 상황에서 무리하게 잠재 고객에게 마케팅을 진행하여 부정적 경험을 주기 보다 가입 확률이 높은 예상 고객들을 기반으로 마케팅을 진행하여 마케팅 비용 감소와 잠재 고객의 부정적 경험을 제거하고자 했습니다.

<6>

그래서 결론적으로 한정된 자원으로 소비자의 니즈 및 성향을 파악하는 것이 마케팅의 경쟁력으로 연결되어서 이번 프로젝트에서

고객 데이터를 이용한 마케팅 비용 감소에 초점을 두어 프로젝트를 진행하게 되었습니다.

<7>

먼저 데이터 선정 단계입니다. 저희는 캐글에서 은행 마케팅 데이터를 얻었습니다.

<8>

해당 데이터는 포르투갈 은행의 마케팅 데이터로 크게 연령이나 직업 같은 고객 데이터, 마케팅 통화 시간, 마케팅 시행 요일 등 실제 마케팅 관련 데이터, 고용률이나 6개월 만기 유로 금리 같은 사회 경제적 속성 데이터, 이전 마케팅의 결과 같은 기타 속성, 실제 가입 여부로 분류되어 있습니다. 해당 데이터에서 실제 마케팅 이전에는 알 수 없는 관련 데이터와 고객 데이터와의 연관성이 조금 떨어지는 사회 경제적 속성을 제외하여 분석 데이터를 구성하였습니다.

<9>

해당 데이터를 기반한 탐색적 데이터 분석을 진행했습니다.

먼저 데이터 분포 측면에서 부문별로 살펴보겠습니다.

연령 부문은 30,40,20대 순서로 마케팅을 진행하고 있으며 이는 현행 은행 마케팅이 경제 활동이 활발한 인구층을 대상으로 발생한다는 사실을 알 수 있습니다.

직군별 분류에 따르면 관리직, 숙련노동자, 기술자, 서비스 업종 순서로 마케팅을 진행하며 상대적으로 학생, 실업자, 가사 도우미의 마케팅 진행 횟수가 낮은 것을 보아 은행 마케팅이 경제적으로 풍족할 것으로 예상되는 직군을 위주로 마케팅을 진행한다는 사실을 알 수 있습니다.

결혼 여부에 따르면 기혼자 대상 마케팅 비중이 60% 이상인 것을 확인 할 수 있었습니다.

교육 수준 부문에서는 박사 학위인 경우를 제외하고 은행 마케팅이 교육 수준에 거의 비례하여 발생한다는 사실을 알 수 있습니다.

파산 여부에 따른 분류에서 전체 데이터 4만여건 중 단 3건만 파산 경험이 있다는 것을 보아 사실상 파산 경험이 있는 사람에게는 은행 마케팅이 진행되지 않는 것을 알 수 있습니다.

주택 담보대출 보유 여부에서는 있는 경우와 없는 경우가 각각 52%와 45%를 차지하고 있어 은행 마케팅 진행에 영향이 거의 없는 것으로 보이지만 개인 신용 대출 보유 여부에 따르면 82%가 없는 것으로 확인되면서 은행 마케팅 진행에 영향력이 큰 것으로 확인되었습니다.

<10>

직전 은행 마케팅 이후 경과 기간의 분류에서 직전 마케팅에서 연락 하지 않은 경우가 전체의 96%에 달했습니다.

이전까지 해당 고객에게 마케팅을 시도한 횟수는 전체의 86%가 시도한 적 없음으로 분류되었고 시도한 적이 있는 14%에서 10%p 가량이 이전 마케팅에 실패했고 약 3%p가량만 성공으로 이어진 것을 확인할 수 있었습니다.

<11>

해당 결과에 대해 pairplot을 통해 상관관계를 시각적으로 확인한 결과 뚜렷한 상관관계가 보이지 않았습니다.

<12>

마지막으로 위에 설명한 독립변수와 종속변수인 가입 여부와의 관계를 비율로 확인하겠습니다.

연령 부문에서 주력 마케팅 연령층에서 실제 마케팅 성공 비율이 적게 나왔는데 이는 해당 고객 층을 제외한 경우 마케팅 진행 횟수 자체가 적어져서 확률에 일종의 왜곡 효과가 나타난 것으로 보입니다.

직업 부문에서도 이와 비슷하게 학생과 은퇴자의 성공 비율이 45%, 33%로 유독 높게 나타나고 나머지는 7~15% 사이의 값을 보여주고 있습니다.

혼인 여부 부문에서는 싱글인 경우와 알수 없음 부문에서 16%, 17%로 같은 현상을 보이고 있으며 11%대의 성공 확률을 보여주고 있습니다.

교육 수준에서는 위와 같은 이유로 문맹인 경우와 unknown의 마케팅 성공 확률이 28%, 17% 수준으로 높게 나오고 해당 사항을 제외하면 마케팅 비중에 비례한 성공 확률을 보여주었습니다.

파산 여부에서는 파산하지 않는 경우 14%, 정보를 알 수 없는 경우 5% 수준의 마케팅 성공 확률을 보여주었습니다.

마지막은 주택담보 대출, 신용 담보 대출의 보유 여부인데 그래프 상에서는 보유 여부에 따른 마케팅 성공 확률이 큰 차이가 있는 것처럼 보이지만 실제 성공 확률은 보유 여부에 상관없이 12~13% 내의 값을 보여주면서 대출 보유 여부는 마케팅 성공 확률에 미치는 영향력이 제한적인 것을 확인 할 수 있었습니다.

<13>

다음은 전처리 과정 입니다. 전체 데이터 세트에 대해 null값은 없었지만 unknown으로 처리된 데이터가 있었습니다. 이에 unknown 값을 인코딩 하는 과정에서 결측치로 취급하여 해당 데이터가 포함된 행을 제거하는 경우와 일종의 데이터로 취급하는 경우의 2가지 데이터 세트에 대해 모델의 성능을 비교하고자 했습니다.

<14>

학습과 검증 데이터를 7:3으로 구분하고 데이터 분석 과정에서 편향을 확인했기 때문에 샘플링 방식에 따라 다음 4가지로 학습 데이터를 구성하여 최적의 모델을 선정하고자 했습니다.

<15>

다음은 모델 종류 선정입니다. 최초 계획에서는 생각 가능한 모든 분류 모델에 대해 별도의 하이퍼 파라미터를 조정하지 않는 경우와 GridSearchCV를 통해 조정 가능한 모든 파라미터를 대상으로 최적 파라미터로 구성된 모델 중에서 최적의 모델을 선정하고자 했으며 그 종류는 다음과 같이 구성되어 있습니다.

하지만 컴퓨터의 한계로 일부 파라미터에 대해서만 지정된 값들 중에서 최적 파라미터를 찾을 수 있었고 DecisionTree, RandomForest, KNN의 경우 컴퓨터 성능상 한계로 GridSearchCV가 작동하지 않아 이를 제외하고 일부 모델에 대해 다음과 같이 파라미터에 대한 리스트를 적용해 GridSearchCV를 진행했습니다.

그 결과 다음과 같은 최적 파라미터가 선정되었습니다.

딥러닝의 경우 은닉층을 1,3,5,7,9, 노드 개수를 4,8,16,32,64개로 조정하면서 최적의 구조를 찾으려고 했고 이에 노드는 32개, 은닉층은 5개일 경우 최적인 결과가 나왔습니다.

<16>

다음은 모델의 성능 평가 지표 부문입니다.

1차적으로 정밀도가 선정되었는데 그 이유는 다음과 같습니다.

첫 번째, 고객 데이터를 통해 예측된 결과값을 기반으로 마케팅이 진행되기 때문입니다.

두 번째, 가입하지 않을 고객에게 마케팅을 진행해서 고객이 얻는 부정적 경험으로 인한 손실이 가입할 수 있지만 모델 판단 과정에 놓쳐서 발생하는 손실보다 더 크다고 생각했기 때문입니다.

그러나 이때, 재현율을 무시할 경우 모델이 성공으로 예측하는 전체적인 숫자 자체가 매우 적어질 것으로 생각되어 다음과 같은 자체 모델 선정 지표를 만들었습니다.

<ppt화면: 정밀도+재현율\*0.56>

해당 지표 선정의 근거로 고객 경험 전문가 blake morgan의 통계 자료를 참고하였습니다.

해당 자료에서 고객은 부정적 경험에 대해 16명의 지인에게 해당 경험을 이야기하고 긍정적 경험에 대해 9명의 지인에게 이야기했습니다.

<17>

여기서 9명을 가입할 고객을 놓쳐서 발생하는 손실, 16명을 고객의 부정적 경험을 수치로 표현한 손실로 생각했습니다. 그래서 부정적 경험으로 잃는 손실을 1로 두어 부정적 경험으로 잃는 손익 대비 긍정적 경험으로 얻는 손익을 산출한 결과 56%가 나와 이 수치를 재현율에 패널티로 적용하여 정밀도와 재현율의 합으로 자체 모델 선정 지표를 구성했습니다.

<18>

다음은 실제 모델 결과치입니다.

전반적인 내용을 요약하여 설명해 드리면

‘unknown’ 데이터를 결측치로 처리한 경우와 데이터 처리한 경우를 비교한 결과 눈에 띄는 큰 차이를 보이지는 않았습니다.

모델에 따른 성능을 비교 했을 때, 전반적으로 Decision tree, Random forest, knn 방식에서 성능이 다른 모델에 비해 떨어지는 것을 확인했고 나머지는 큰 차이를 보이지 않았습니다.

샘플링 방식에 따라 성능을 비교 했을 때, 랜덤 샘플링을 진행한 경우 전반적으로 모델의 재현율이 상당히 개선 되었으나 정밀도가 상당히 떨어지는 것을 확인할 수 있었습니다. SMOTE 방식에서 일부 모델에서 지표의 개선이 보여졌으나 전반적인 모델 성능 향상으로 이어지지 않았습니다.

해당 결과에 대해 세부적으로 말씀 드리면 다음과 같습니다. 여기서는 unknown 값을 데이터로 처리한 경우에 대해서만 보여드리도록 하겠습니다.

첫번째는 샘플링을 하지 않은 경우의 모델 성능입니다. 전반적으로 정확도는 86~89%, 정밀도는 random forest와 decision tree ,knn 방식을 제외한 경우 59~64% 수준을 보여 주었습니다. 하지만 재현율 부문에서 전체적으로 18~20% 수준을 보여주었습니다. 그 결과 자체 성능 지표에서 44~75% 사이의 성능을 보여주었습니다.

해당 현상에 대해서 실제 데이터의 종속변수의 편향이 심하기 때문이라는 판단하여 다양한 샘플링 기법을 적용한 결과치도 확인하였습니다. Under sampling 방식에서 재현율이 42~57% 사이의 수치를 보여주었지만 정확도가 전체적으로 60~80% 사이의 수치로 낮아졌습니다. 또한 정밀도 역시 16~27% 사이 수준으로 크게 하락하였습니다. 그 결과 46~51% 사이의 성능을 보여주었습니다

<19>

Over sampling 방식을 적용한 결과 정확도는 72~83% 수준을 보여주었고 재현율도 31~46% 대의 수준으로 향상되었지만 정작 정밀도가 16~27% 수준으로 크게 하락하여 결과적으로 36~51%(ppt 수정 필요) 사이의 성능을 보여주었습니다.

Over sampling 방식 중 SMOTE 방식을 적용한 경우 정확도가 87~89% 수준을 보여주었습니다. 정밀도 측면에서 Randomforest, Decision Tree, knn 방식은 34~40% 수준의 정밀도를 보여주었고 이를 제외한 경우 정밀도가 60~63% 사이의 값을 보여주었습니다. 해당 방식에서 재현율은 19~26% 사이의 값을 보여주었습니다. 그 결과 45~74% 사이의 성능을 보여 주었습니다.

<20>

최종적으로 unknown을 데이터 처리한 경우 훈련 데이터 자체에 샘플링을 적용하지 않았을 때, 최적 파라미터를 적용한 XGBClassification에서 최적 모델로 선정되었습니다. 그래서 모델을 기반으로 빅데이터를 전혀 모르는 은행 직원이 고객 데이터를 기반으로 마케팅 성공 예상 고객 정보를 뽑아 csv 파일로 저장 할 수 있는 파이썬 파일을 만들었습니다.

<시연… 화면 미리 띄어놓을 것.>

<21>

저희 프로젝트의 결론 및 시사점입니다.

선정 모델 자체가 마케팅 비용 절감 측면 자체만 보면 유의미한 결과를 도출했으나 마케팅 성공 예측 모형에는 부적절했습니다. 이를 해결하기 위해 다양한 샘플링 방식, 그리고 소개되지 않았지만 가중치 부여 방식 등을 시도 했지만 만족할만한 성과가 나오지 않았습니다.

이런 결과가 나온 원인을 분석한 결과 선정된 고객 데이터의 수, 정확하게 컬럼의 개수가 마케팅 성공을 예측하기에 적기 때문이라는 사실을 알게 되었습니다. 실제로 원데이터 전체를 기반으로 분류 모형을 만들었을 때, 그 결과가 다음과 같이 성능이 좋은 분류 모형이 나온 연구 결과를 확인했습니다. 하지만 저희 프로젝트는 고객 데이터를 활용해 마케팅의 성공을 예측하는 것이 목적이기 때문에 해당 방식을 적용할 수 없었습니다.

본 프로젝트에서 충분한 고객 데이터를 확보 하지 못해서 아쉬움이 남습니다. 그렇기에 향후 충분한 고객 데이터가 주어졌을 때, 실제 은행에서 사용 가능한 모형을 도출하는 것을 추후 연구 과제로 남겨두고 싶습니다.

<22>

이상으로 발표를 마치도록 하겠습니다. 감사합니다.

질문 정리

질문 타임

- f1 score가 왜 그 모양?:

모든 모형에서 f1score가 그렇게 나왔고 이 부분을 해결하기 위해 각종 샘플링 방법이나 데이터 전처리 방법을 다르게 하는 등의 방식을 동원했지만 해결 되지 않아서 최종적으로 해당 모형을 선정했다.

- roc-auc score가 왜 그 모양?:

그렇기 때문에 새로운 지표를 만들어서 최적 모델을 선정하고자 했다.

- precision 지표의 근본적 문제가 성공 수를 적게 선정하는 것인데 그러면 최적 모형의 실제 활용에 있어서 제한적이지 않는가?->

그 부분은 우리도 아쉽게 생각하고 있다. 그렇기에 해당 부분이 우리 프로젝트의 한계점이라 생각한다.

- 결과가 그렇게 나오면 결국 데이터 부족 문제가 아니라 해당 데이터 자체가 분류에 적합하지 않은 것 아니냐?

=> 어떤 근거로 그렇게 말하는 것인가?: 결과치 보고 그렇게 말한다.

=> 컬럼이 충분히 주어졌을 경우 모델을 돌려본 이후에 정확하게 대답할 수 있을것 같다. 다만 실제 사회 경제적 속성을 추가했을 때, 성능이 개선되었고 원 데이터 전체에 대해서 분류 돌릴 때, 성능이 좋은 모형이 나왔다는 것은 말씀드릴 수 있다.

- 고객 경험의 비율을 정밀도, 재현율에 그렇게 적용한 이유???

- 정밀도: 재현율 = 부정적 경험: 긍정적 경험 = 16:9 = 1:0.56

재현율(recall): d/(c+d) 실제 성공 중에서 예측이 성공인 비율

정밀도(precision): d/(b+d) 예측이 성공 중에서 실제가 성공인 비율

c: 실질적으로 가입할 생각이 있지만 모델이 놓친 인원 -> 긍정적 경험을 하지 못함

b: 가입할 생각이 없지만 모델이 가입한다고 한 인원 -> 부정적 경험을 함

실질적으로 가입할 생각이 있는 고객을 찾아서 마케팅을 진행한다는 사실 자체를 고객의 긍정적 경험으로

가입할 생각이 없는 고객에 대해서 마케팅을 진행한다는 사실 자체를 고객의 부정적 경험으로 생각했다.

여기서 정밀도는 예측 성공 중에서 실제 성공인 비율로 실제는 가입할 생각이 없지만 모델이 가입할 것으로 예측한 인원을 포함한다. 즉 해당 수치가 높아질수록 고객의 부정적 경험이 줄어든다.

재현율은 실제 성공 중에서 예측이 성공으로 하는 비율로 실제는 가입할 생각이 있지만 모델이 예측을 실패한 인원을 포함한다. 즉 해당 수치가 높아질수록 고객이 긍정적 경험을 놓치는 경우가 줄어든다.

그래서 위에서 설명한 부정적 경험 대 긍정적 경험의 비율을 적용했다.

-라벨 인코딩, 원핫 인코딩을 한 이유는?

라벨 인코딩: 데이터를 수치화하기 위한 목적으로 진행했다.

원핫 인코딩: 라벨 인코딩한 데이터를 분류모형을 돌릴 경우 해당 값의 차이에 따라 일종의 관계성이 있다고 판단 할 수

있기 때문에 분류 모델에는 원핫 인코딩을 진행하였습니다.