1. 서론

은행의 마케팅 캠페인 효율성을 증대시키기 위한 고객 반응 예측 모델 개발을 진행하였다. 기존의 무작위적인 캠페인 방식이 아닌, 특정 상품에 가입할 가능성이 높은 고객을 사전에 예측함으로써 마케팅 자원의 효율적인 배분을 목표로 한다.

은행이 보유한 고객 데이터와 이전 마케팅 캠페인 결과 및 경제 지표를 활용해, 고객의 상품 가입 여부를 예측하는 이진 분류 모델을 구축하고자 한다. 정확한 예측 모델은 은행이 잠재 고객에게 맞춤형 마케팅 전략을 수립하고, 불필요한 마케팅 비용을 절감하며 수익성을 향상시키는 데 기여할 수 있다.

2. 데이터

2.1 데이터 소개

은행 마케팅 캠페인 데이터셋은 고객 정보, 과거 마케팅 캠페인 참여 이력, 경제 지표 등을 포함하고 있다. 종속 변수는 고객이 해당 마케팅 캠페인을 통해 은행 상품에 가입했는지 여부(y)이며, 'yes(1)' 또는 'no(0)'의 이진 값을 가진다.

변수명	설명	데이터 타입 값 종류
age	고객의 나이	Int
Job	직업 유형	category
marital	결혼 여부	Category
education	교육 수준	Category
default	신용 불이행 여부	Category
housing	주택대출 보유 여부	Category
loan	개인대출 보유 여부	Category
contact	연락 방법	Category
month	마지막 연락 월	Category
day_of_week	마지막 연락 요일	Category
duration	마지막 연락 지속 시간(초)	int
campaign	현재 캠페인 중 고객과 연락한 횟수	int
pdays	이전 캠페인에서 고객에게 연락한 후 경과된 날짜 수	int
previous	이전 캠페인에게 고객과 연락한 횟수	int
poutcome	이전 마케팅 캠페인 결과	Category

emp.var.rate	고용 변화율(분기별 지표)	float
cons.price.idx	소비자 물가 지수(월별 지표)	float
cons.conf.idx	소비자 신뢰 지수(월별 지표)	float
euribor3m	3개월 유리존 은행 간 금리(일별 지표)	float
nr.employed	직원 수(분기별 지표)	float
y	상품가입여부	int

데이터 분석 결과, 종속 변수 y의 클래스 분포는 불균형하다. 'no(0)' 클래스가 'yes(1)' 클래스보다 훨씬 많은 비율을 차지하고 있어, 모델 학습 시 미가입 클래스로 편향될 가능성이 있다. 이러한 클래스 불균형은 모델의 재현율을 낮출 수 있어, 모델링 과정에서 고려해야 한다.

2.2 데이터 전처리

모델 학습에 앞서 데이터의 품질을 향상시키고 모델의 성능을 높이기 위해 다음과 같은 전처리 과정을 수행하였다.

2.2.1 결측치 처리

데이터 셋 내 범주형 변수에서 결측치는 'unknown'값으로 되어 있으며, 각 변수별 개수는 다음과 같다.

칼럼	**개수**	**칼럼**	**개수**
job	330	marital	80
education	1731	default	8597
housing	990	loan	990

'unknown'값은 별도의 유효한 카테고리로 간주하고, one-hot encoding을 적용해 모델이 이 정보를 학습에 활용할 수 있도록 진행하였다.

2.2.2 이상치 처리

age(나이)는 최소 17세, 중앙값 38세, 최대 98세로 확인되었다. 70대 이상 고객의 존재 가능성을 고려해 별도 이상치 처리는 진행하지 않았다.

duration(마지막 통화 시간)은 최대값이 4918초로 확인되었다.이상치 값은 전체 데이터의 약 7%를 차지했으며, 이상치 영향을 완화하기 위해 duration_log 변수를 추가하고, 이상치여부를 나타내는 플래그 변수를 추가했다.

pdays(이전 캠페인 후 지난 일수)는 최대값이 999이며, 중앙값과 3분위수 모두 999로 나타났다. 이는 대부분의 고객이 이전 캠페인 동안 연락을 받은 적이 없는 것으로, 값 999의 비율은 96.32%이다. 따라서, 이전에 연락한 적이 있는지 여부를 나타내는 이진 변수 previous_contact를 추가하였다.

previous(이전 캠페인 동안 연락 횟수)는 최대 7로, 대부분이 캠페인에 참여한 적이 없어 이상치로 간주하기 어렵다고 판단했다.

cons.conf.idx(소비자 신뢰 지수)는 약 1%의 이상치가 확인되어, 해당 데이터는 삭제하였다.

2.2.3 중복값 제거

데이터셋 내 중복된 행을 확인하고 제거 진행하였다.

2.2.4 최종 사용 변수

최종적으로 모델 학습을 위해 다음과 같은 변수들을 선택하고 전처리를 진행하였다.

age, job, marital, education, default, housing, loan, contact, month, campaign, pdays, previous, emp.var.rate, poutcome, cons.price.idx, cons.conf.idx, euribor3m, nr.employed, previous_contact, y를 사용하였으며,

job, marital, education, default, housing, loan, contact, month, poutcome, previous_contact 변수들에 대해 One-Hot Encoding을 적용하였고. 다중 공선성 방지를 위해 첫 번째 카테고리는 제거했다.

3 모델링

은행 마케팅 캠페인 데이터의 고객 상품 가입 여부를 예측하기 위해 다음과 같은 단계로 모델링을 진행하였다.

3.1 1차 모델링

클래스 불균형 문제를 고려하기 전에, 기본적인 성능을 확인하기 위해 여러 분류 모델을 학습하였다. 사용된 모델은 Decision Tree, Voting, Random Forest, AdaBoost, Stacking, Logistic Regression, XGBoost, Gradient Boost였으며, 각 모델의 정ㅇ확도, 정밀도, 재현율,

F1-score를 비교하였다.

모델	Accuracy	Precision	Recall	F1-score
Decision Tree	0.8373	0.3141	0. 3476	0. 3300
Voting	0.8884	0.5318	0.2640	0.3529
Random Forest	0.8924	0.5645	0.2921	0.3850
AdaBoost	0.8966	0.6870	0.1896	0.2972
Stacking	0.8995	0.6784	0.2340	0.3578
Logistic Regression	0.8981	0.7078	0.1973	0.3086
XGBoost	0.8968	0.6207	0.2690	0.3753
Gradient Boost	0.8986	0.6727	0.2338	0.3471

결과 분석에서 높은 정확도에도 불구하고, 재현율이 낮아 실제 '가입' 고객을 잘 예측하지 못하는 문제점을 확인하였다. 이는 데이터의 클래스 불균형 때문이라 판단하고, 다음 단계에서 불균형 처리 기법을 적용하기로 결정하였다.

3.2 2차 모델링(SMOTE 적용)

클래스 불균형 문제를 완화하고 '가입' 고객 예측 성능(재현율)을 향상시키기 위해 SMOTE(Synthetic Minority Over-sampling Technique)를 적용해 학습 데이터를 재 샘플링 했다. SMOTE 적용 후 각 모델의 성능은 다음과 같다.

모델	Accuracy	Precision	Recall	F1-score
Decision Tree	0.8264	0.2878	0.3771	0.3265
Voting	0.8678	0.4273	0.5429	0.4782
Random Forest	0.8747	0.4305	0.3815	0.4045
AdaBoost	0.8321	0,3453	0,5635	0.4282
Stacking	0.8748	0.4255	0.3500	0.3841
Logistic Regression	0.8318	0.3226	0.4615	0.3797
XGBoost	0.8822	0.4676	0.4079	0.4357
Gradient Boost	0.8676	0.4273	0.5429	0.4782

SMOTE 적용 결과, 전반적으로 재현율이 크게 향상되었으나, 정밀도는 감소하는 경향을 보였다. F1-score는 여러 모델에서 개선되었으며, 특히 Voting과 Gradient Boost 모델에서 상대적으로 높은 F1-score를 나타냈다.

3.3 3차 모델링(SMOTEENN 적용)

SMOTE의 오버샘플링과 ENN(Edited Nearest Neighbors)의 언더샘플링을 결합한 SMOTEENN 기법을 적용해 모델 성능을 추가적으로 개선하고자 하였다. SMOTEENN 적용 후 각 모델의 성능은 다음과 같다.

모델	Accuracy	Precision	Recall	F1-score
Decision Tree	0.7876	0.2842	0.5950	0.3846
Voting	0.8168	0.3341	0.6464	0.4405
Random Forest	0.8390	0.3660	0.6053	0.4562
AdaBoost	0.7479	0.2579	0.6713	0.3726
Stacking	0.8090	0.3152	0.6075	0.4150
Logistic Regression	0.7872	0.2907	0.6302	0.3979
XGBoost	0.8493	0.3859	0.5943	0.4679
Gradient Boost	0.8178	0.3333	0.6332	0.4367

SMOTEENN 적용 결과, 재현율은 더욱 향상되었지만 정밀도는 2차 모델링에 비해 전반적으로 감소했다. 이는 SMOTEENN의 특성상 오버샘플링과 함께 noisy한 데이터를 제거하는 과정에서 발생할 수 있다.

3.4 4차 모델링(XGBoost 하이퍼파라미터 튜닝)

F1-score를 기준으로 가장 우수한 성능을 보인 XGBoost 모델에 대해 하이퍼파라미터 튜닝을 진행하여 모델 성능을 최적화하고자 하였다. GridSearchCV를 사용해 주요 하이퍼파라미터를 탐색한 결과, 다음과 같은 최적의 파라미터와 성능을 얻었다.

최고 F1-score(교차검증)은 약 0.9581, 테스트 데이터의 정확도는 약 0.8501이었다. 클래스 0(미가입) 정보는 정밀도 0.95, 재현율 0.88, F1-score 0.91이었으며 클래스 1(가입)정보는 정밀도 0.39, 재현율 0.60, F1-score 0.47이었다.

하이퍼파라미터 튜닝을 통해 교차 검증 성능은 크게 향상되었으나, 테스트 데이터에서의 F1-score는 상대적으로 낮아 과적합 가능성을 시사했다.

3.5 5차 모델링(규제 포함 하이퍼파라미터 튜닝)

과적합 가능성을 줄이기 위해 XGBoost 모델에 규제 관련 하이퍼파라미터(gamma, reg_alpha, reg_lambda)를 포함해 추가적인 하이퍼파라미터 튜닝을 진행하였다.

GridSearchCV를 사용해 확장된 파라미터 그리드를 탐색한 결과, 다음과 같은 최적의 파라미터와 성능을 얻었다.

최고 F1-score(교차검증)은 약 0.9550, 테스트 데이터 정확도는 약 0.8483이었다. 클래스 0(미가입) 정보에서 정밀도 0.95, 재현율 0.88, F1-score 0.91이 나왔으며, 클래스 1(가입) 정보는 정밀도 0.38, 재현율 0.60 F1-score 0.47이 나왔다.

규제 파라미터를 추가하였으나, 테스트 데이터에서의 성능 향상은 미미했다. 교차 검증 F1-score는 이전 튜닝 결과와 비슷한 수준을 유지했으며, 테스트 데이터 성능 또한 큰 변화 없이 유사한 결과를 보였다.

4. 결과

4.1 모델링 결과

1차 모델링에서는 기본적인 분류 모델들을 학습해 데이터 클래스 불균형 문제를 확인하였다. 2차 및 3차 모델링에서는 SMOTE, SMOTEENN과 같은 오버샘플링 기법을 적용해 소수 클래스인 '가입' 고객의 예측 성능, 그 중에서도 재현율을 향상시키고자 하였다. 이 과정에서 재현율을 개선되었으나, 정밀도는 감소하는 경향을 보였다. 4차 및 5차 모델링에서는 F1-score를 최적화 하기 위해 XGBoost 모델에 대한 하이퍼파라미터 튜닝을 GridSearchCV를 통해 진행하였다. 4차 모델링에서는 기본적인 하이퍼파라미터 탐색을, 5차 모델링에서는 과적합 완화를 위해 규제 파라미터를 포함한 튜닝을 하였다. 하이퍼파라미터 튜닝을 통해 교차 검증 성능은 높게 나왔으나, 테스트 데이터에서의 성능 향상은 제한적이었다.

최종 모델 성능인 규제 포함 XGBoost에서의 정확도는 약 0.8483, 클래스 1(가입)의 정밀도는 약 0.38, 재현율은 약 0.60, F1-score는 약 0.47이었다.

클래스 불균형 처리 기법을 통해 '가입' 고객에 대한 재현율을 상당 수준으로 향상시키는 데 성공하였다. 특히 SMOTE와 SMOTEENN 적용이 재현율 향상에 기여했다. 최종적으로 하이퍼파라미터 튜닝을 거친 XGBoost 모델은 '가입' 고객을 탐지하는 능력이 비교적 우수하나, 예측의 정밀도는 아직 개선의 여지가 있다.

향후 모델 성능 향상을 위해서는 피처에 대한 정교한 엔지니어링이나 다른 불균형 데이터 처리 기법을 사용해볼 수 있고, 다른 모델을 이용해 볼 수 있을 것이다.

4.2 인사이트

모델 분석 결과, 학생, 은퇴자, 미혼고객, 주택 대출만 있고 개인대출이 없는 고객 또는 대출 정보가 없는 고객층에서 저기 예금 가입률이 높게 나타났다. 반면, 육체 노동직 종사자의 가입률은 상대적을 낮았다.

연락 방식으로는 모바일(cellular)이 유선(telephone)보다 효과적이었으며, 3월과 12월의 가입률이 높았다. 특히 4월 목요일의 가입률이 주목할 만하였다. 또한 이전 캠페인 성공 고객의 재가입률이 높은 점 또한 중요 인사이트다.

4.3 비즈니스 전략

이러한 분석을 바탕으로 다음과 같은 비즈니스 전략을 제안한다.

먼저, 학생, 은퇴자, 미혼 고객 등 가입 가능성이 높은 특정 고객 그룹을 대상으로 맞춤형 상품 안내 및 프로모션을 강화하고, 육체 노동직 종사자를 위한 제안을 재고한다.

다음으로, 모바일 연락 중심으로 3,12월 집중적인 마케팅 캠페인을 진행하며, 이전 캠페인 성공 고객을 리타겟팅하면 좋을 것이다.

마지막으로, 높은 가입률을 보이는 고객 그룹의 특성을 반영한 특화된 예금 상품 개발을 고려해, 가입률이 저조한 시기인 5월에 대한 마케팅 전략을 재검토해 볼 수 있겠다.

이 전략들을 통해 마케팅 자원의 효율성을 극대화하고, 정기 예금 가입 고객 유치를 증대시킬 수 있을 것으로 기대한다.