프로젝트 기획서

1. 주제

실제 고객 데이터 기반 은행 마케팅 비용 감축 최적 모델 선정

1. 개요(모티베이션)

최근 고객 데이터를 활용한 금융권의 마케팅이 활발해 지면서 우리도 해당 내용을 적용한 프로젝트를 진행하고자 했다. 하지만 국내 데이터를 확보하기 어려운 상황으로 kaggle의 은행 정기예금 홍보 데이터를 기반으로 실제 금융 상품 마케팅 이전에 고객 정보를 통해 미리 해당 마케팅 성공 여부를 확인하여 성공 가능성이 높은 고객들을 파악하고자 한다. 이를 통해서 은행의 마케팅 비용을 감소하고 성공 확률을 높이고자 한다.

1. 목적

고객 정보를 기반한 마케팅 성공 예측 분류 모델을 선정하여 이전에 성공 가능성이 높은 고객들을 파악하여 해당 고객들을 대상으로 한 홍보를 통해 마케팅 성공 확률을 올리고 비용을 줄인다.

1. 팀원 및 역할 분담

* 팀장: 박준배
* 경제 도메인: 조해원
* PPT제작: 박희연
* 기록: 김민성
* 코드 구현: 박준배

1. 조사 데이터

* Kaggle 데이터()

1. 프로젝트 구성

* 데이터 전처리 및 EDA
* 모델 설정 및 결과 도출
* 모델 평가

1. 데이터 구성
2. 은행 고객 데이터:
   1. age : 연령(숫자)
   2. Job: 직업의 종류
      1. 카테고리:

"관리","블루칼라","기업가","하녀","관리","은퇴","자영업","서비스","학생 ","기술자","실업자","알 수 없음"

* 1. Marital: 혼인 여부
     1. 범주:

"이혼", "기혼", "미혼", "모름", 참고: "이혼"은 이혼 또는 사별을 의미함

* 1. Education: 교육 수준
     1. 카테고리:

"basic.4y","basic.6y","basic.9y","high.school","문맹","professional.course","university.degree","알 수 없음" )

* 1. Default: 파산 경험
     1. 범주: "아니오","예","알 수 없음"
  2. Housing: 주택 융자가 있습니까?
     1. 범주: "아니오","예","알 수 없음"
  3. Loan: 개인 대출이 있습니까?
     1. 범주: "아니오","예","알 수 없음"

1. 최근 마케팅 연락 관련
   1. Contact: 연락
      1. 통신 유형: "핸드폰", "전화"
   2. Month: 해당 연도의 마지막 접촉 월
      1. 카테고리: "jan", "feb", "mar", ..., "nov", "dec"
   3. day\_of\_week : 마지막 연락 요일
      1. 카테고리: "월", "화", "수", "목", "금"
   4. duration: 마지막 접촉 지속 시간(초)(숫자).
      1. 중요 참고 사항: 이 속성은 출력 대상에 큰 영향을 미칩니다(예: duration=0이면 y="no"). 그러나 호출이 수행되기 전의 지속 시간은 알 수 없습니다. 또한 호출이 끝난 후 y는 분명히 알려져 있습니다. 따라서 이 입력은 벤치마크 목적으로만 포함되어야 하며 실제 예측 모델을 갖고자 하는 경우에는 버려야 합니다.
2. 타겟값
   1. y - 고객이 정기 예금에 가입했습니까? (이진법: "예1","아니요0")
3. 특이사항
   1. 속성 값 누락
      1. 일부 범주 속성에는 여러 개의 누락 값이 있으며 모두 "알 수 없음" 레이블로 코딩됩니다. 이러한 누락된 값은 가능한 클래스 레이블로 처리되거나 삭제 또는 대치 기술을 사용하여 처리될 수 있습니다
         * 해당 부분에 대한 전처리 필요

.

1. 분석 대상 및 모델
2. 분석 대상: 은행 고객 마케팅 데이터
3. 데이터 선정 기준
   1. 마케팅 시작 이전에 알 수 있는 정보
4. 분석 방법
5. 분류 모델

* 가능한 모든 분류 방법을 사용하여 아래 평가 방식을 적용하여 가장 합리적인 모델 선정

1. 모델 평가 방법
   1. 평가 방법에 대한 고찰
      1. 정밀도(Precision) 기반 평가
         1. 장점: 마케팅 비용 감소 효과
         2. 단점: 상대적으로 마케팅 성공 수의 대폭 감소 예상
      2. 재현율(recall) 기반 평가
         1. 장점: 더 많은 마케팅 성공 횟수
         2. 단점: 상대적으로 더 많은 비용 예상
      3. 정확도(accuracy) 기반 평가
         * 타겟 값 자체가 실패에 편향되어 있고 독립변수의 값도 편향되는 경향이 있기 때문에 부적절하다고 판단
      4. F1\_score
         * 정밀도와 재현율 사이의 관계 기반 분류 결정 임계값 조정
2. 프로세스

* 선형 연구 검토 및 데이터 전처리 논의
* 전처리 후 모델 설정
* 모델 성능 검토

1. 사용법???

* 고객 데이터 입력
* 해당 데이터 기반 마케팅 성공 예측
* 성공 예측 고객을 csv 파일로 만들어 저장

1. 한계점

* 기타 금융자산의 상호 연관성을 파악하기 어렵다(ex: 적금, 펀드 등)

예측상 가입한다는 사람들을 대상으로 마케팅을 진행한다.

정밀도(예측이 성공할 때 실제 성공할 확률) 우선!!!

아무리 모델을 돌려도 f1 값은 똥이 나오기 때문에 정밀도 기반의 모형 결정이 필수적으로 보인다. 그러기 위해서는 은행의 전화 마케팅 자체가 전체 마케팅 효용 중에서 차지하는 부분이 작아서 비용 감소의 효용이 잘못 분류해서 실제로 가입할 수 있지만 놓치는 것으로 발생하는 비용보다 커야 할 것이다.

* + - 1. 정밀도가 고객 경험 측면에서 부정적 고객 경험이 은행에 더 큰 손해라고 보기 때문에 정밀도를 중요하게 봤다.
      2. 전체 전화 영업 가입 모수 자체가 작기 때문에

고객 데이터 기반 은행 마케팅 성공 예측 최적 모델 선정