안녕하세요

최적 모델 선택을 통한 카드 채무 불이행자 데이터 결과 분석 및 채무 불이행자 예측 프로젝트를 맡은 7조 발표자 김윤성입니다

@발표순서는 다음과 같습니다

@@ 여러분 혹시 2000년대 초반에 있었던 카드대란을 들어보셨습니까?

2000년대 초반 수백만 명의 신용불량자가 카드빚을 갚지 못해 발생한 충격적인 사건입니다

카드사들은 경쟁 과잉으로 무분별한 신용카드 발급을 진행하였고 발급 자격이 되지 않지만 손쉽게 신용카드를 발행한 사람들은 결국 카드빚을 갚지 못하고 신용카드 현금서비스까지 이용하였습니다. 결국 이런 악순환이 반복되다가 파산까지 이르는 고객들이 급증하며 카드사들은 많은 고초를 겪은 사건입니다. 대표적으로 대기업이였던 LG카드도 파산하는 일까지 벌어지며 대한민국 경제에 큰 타격을 주었습니다

---------------------------------------------------------------------------------

@오래된 금융사들은 항상 리스크 관리를 해오지만 신용카드는 등장한지 오래 되지 않았기 때문에 사전 리스크 방지가 매우 중요합니다.

신용카드의 누적된 금액은 기간이 오래될 수록 체납자만 늘어나고 부채가 축적될 것이기 때문에 금융사들은 리스크 관리를 위해서 고객들의 체납율을 미리 예측하여 대비할 필요가 있습니다.

-----------------------------------------------------------------

@이런 리스크를 사전에 예측하고자 저희가 선택한 카드 체납 여부 데이터의 변수는

이름, 성별, 최종 학력, 신용 금액, 최근 6개월간 상환현황, 청구서 금액, 지급액 등 3만 명의 각종 특징들이 변수로 있습니다

@@먼저 데이터의 클래스 변수의 고르기를 보기 위해 디폴트 변수를 윈그래프로 나타낸 결과 채무자의 비율이 약 80%로 클래스의 불균형이 있는 것을 볼 수 있었습니다. 이를 통해 데이터 전처리 단계에서 오버 샘플링을 해야겠다는 계획을 세웠습니다.

------------------------------------------------------------------------------------------------------------------------

@데이터 내에서도 변수 중요도가 다를거라 예상하여 각각 변수들을 더 자세히 보기위해 다양한 시각화를 진행하였습니다

먼저 학력이 높을수록 채무불이행 가능성이 낮아진다는 결과를 얻을 수 있었습니다

다음으로 결혼 여부에 따라서 채무 불이행 가능성이 차이를 보이지만 미세한 차이임으로 결혼여부는 채무 여부의 값에 영향을 별로 주지 않는 변수라고 가정을 세웠습니다

----------------------------------------------------------------------------------------

@다음으로 나이에 채무 여부를 동시에 볼 수 있는 분포표를 본 결과 이 데이터셋 사람들의 나이는 25세부터 40세 사이의 사람들이 많았고 각 나이의 비율로 범주화 하고 분석하였더니 26대부터 30대가 가장 채무불이행율이 높았습니다

----------------------------------------------------------------------------------------------------

@다음으로 고객들의 신용 한도를 봤을 때 신용한도가 높아질수록 채무불이행 가능성이 낮아지는걸 확인할 수 있었습니다. 또한 신용한도의 범위가 상당히 커서 데이터 정규화를 해야겠다고 계획하였습니다.

------------------------------------------------------------------------

@마지막으로 변수들간의 상관관계가 높을 경우엔 공분선성이 생겨 모델 성능에 영향을 끼치므로 상관관계 분석을 하였다.

변수들간의 상관관계를 히트맵을 통해 분석해본 결과 월마다의 상환현황 변수끼리 상관관계가 높은 것이 확인 되었으며 상관관계를 없애기 위해 상관관계 높은 변수들을 합친 새로운 파생변수를 만들어 기본 데이터와 비교 가능한 대조군1 데이터를 생성했다.

------------------------------------------------------------------------------------------------------------------------------

@위에서 상관관계가 높은 변수들을 찾을 수 있었으며 상관관계가 높은 변수들이 실제로 모델에 얼마나 영향을 끼칠지 알아보기 위해 독립여부에 성능이 좌우되는 나이브 베이즈 분류기 모델을 선택하였다.

--------------------------------------------------------------------------------

@또 가장 대표적 모델로 다른 모델과 객관적으로 비교 할 수 있는 모델인 로지스틱 회귀 모형을 선택하였다.

----------------------------------------------------------------

@마지막으로 채무 불이행에 어떤 변수들이 영향을 많이 미치는지 순위를 매기기 위해 랜덤 포레스트 모델을 선택하였다.

---------------------------------------------

@먼저 랜덤포레스트 모델로 변수 중요도를 알아본 결과 최근 3개월 내 소비패턴 변수들만 중요하고 나머지 변수들은 영향을 주지 않는다는 것을 알 수 있었다

------------------------------------------------------------------------

@이를 통해 모델의 성능을 높이기 위해 영향이 높지 않은 변수들을 제거한 대조군2와 영향이 높지 않은 변수 제거한 값에서 대조군 1처럼 새로운 파생변수를 만든 대조군 3의 데이터를 새로 만들어 다시 데이터 전처리 단계로 돌아 갔다.

-----------------------------------------------------------------------------------------------

@변수들 마다 가지는 범위들이 크기 때문에 과대 적합을 낮추기 위해 스케일링을 하였고 했을때와 안했을때를 나눠 비교하기로 하였습니다.

@또한 클래스 불균형을 보완하기 위해 오버샘플링을 한 모델과 하지 않은 모델로 나눠 비교 하기로 하였습니다.

@먼저 모델 성능 평가를 위해 채무불이행자로 예측한 사람중 실제로 채무불 이행자에 대해 알 수 있는 정밀도와 실제 채무 불이행자 중 예측한 사람이 진짜 채무 불이행자일 확률인 재현율을 조화평균으로 측정해주어 종합적 성능평가가 가능한 F1 스코어를 1순위로 비교하였고 실제로 채무 이행자한테 불이행했다고 판단하는 경우가 회사에 타격이 더 크다고 판단하여 재현율을 2순위 지표로 잡고 각각 다른 총 8가지 데이터들을 분석하였습니다

@먼저 랜덤포레스트 모델에서는 당연히 오버샘플링 했을때가 더 f1 스코어가 높게 나왔으며 이중에서도 파생변수를 하나 더 만들어준 대조군 1이 가장 높게 나왔습니다

@다음으로 변수들의 상관성이 얼마나 모델에 영향을 끼치는가를 보기위해 나이브 베이즈 모델로 분석하였다. 오버 샘플링을 했을때 상관관계가 높은 변수들이 제일 많은 대조군1이 역시 제일 점수가 낮았으며 원본과 비교했을때 점수가 0.6정도로 아주 많은 차이를 보였다 이를통해 상관관계가 모델에 아주 많은 영향을 미치는 것을 알 수 있었다

@다음으로는 로지스틱 회귀로 전반적으로 몇몇개의 변수들이 달라졌다고 대조군끼리의 차이는 별로 없었지만 오버 샘플링을 할때와 하지 않을때 엄청난 점수 차이를 나는 것을 확인할 수 있었다

@이 3가지 모델들을 비교해본 결과 랜덤포레스트 모델이 전반적으로 가장 높은 점수를 기록했으며 랜덤포레스트와 같은 앙상블 모델이지만 예측모델에 대한 에러를 줄여줄 수 있는 부스팅 기법을 기반으로하는 xgb 모델을 사용한다면 더 좋은 점수를 얻을 것이란 가설을 세우고 모델링 해보기로 하였다

@xgb로 모델링을 해본 결과 예상과는 다르게 랜덤 포레스트와 크게 다르지 않다는 것을 알 수 있었는데 이는 결국 트리를 만들어 비슷한 베이스로 분석을 하기 때문이라고 판단하였다

@@그렇다면 이전에 데이터 전처리를 통해 모델의 예측력을 실제로 높히는데 도움이 되었는가를 보기위해 예측력이 가장 좋았던 랜덤포레스트 모델에서 나온 F1 스코어들을 실험계획법을 활용하여 분산분석을 진행하였습니다

요인으로는 스케일링을 하였을때와 하지 않았을때, 오버샘플링을 하였을때와 하지 않았을때, 상관성이 높은 변수들을 묶어 새로운 파생변수를 만들었을때와 만들지 않았을때, 상관관계 높은 변수를 없앨때와 그대로 두었을때 이렇게 총 4가지로 분석을하였습니다

분석을 한 결과 스케일링과 파생변수를 같이 만들었을때와 스케일링과 변수삭제를 동시에 했을때는 p-value가 유의수준보다 높게 나와 제외를 하고 보겠습니다

@김동익 준거 넣기

@파레토 차트로 각 요인들이 실제 예측에 얼마나 영향을 주나 봤더니 오버 샘플링이 가장 크게 나왔고 두번째로 오버샘플링과 스케일링을 같이한것 세번째로 스케일링이 나왔습니다

@좀 더 자세히 보기 위해 주효과도를 활용하여 봤을때 스케일링과 오버 샘플링이 아주 큰 기울기로 영향을 많이 준다는 것을 볼 수 있었고 변수 삭제와 파생변수 만드는 전처리과정은 그렇게 많은 영향을 미치진 못한 것을 볼 수 있었다

@@이 모델링들과 실험계획법을 통하여 예측력을 높이려면 꼭 스케일링과 오버 샘플링은 해야하며 모델은 트리 기반 모델이 본 프로젝트에 가장 적합하다고 판단하였습니다