#빅콘테스트 일요일아침 발표 스크립트#

**1. 페르소나 (시작)**

: 안녕하세요 저희는 발표전에 간단한 2가지 사례로 먼저 시작하겠습니다. 청년 창업가 A씨는 대출받기위해 창구를 찾음, 신용등급 없고 -> 대출못받음

B보험사 팀장 K씨는 현재 기존 신용평가 모형으로 고객평가 x , 새로운 신용평가 모델이 필요하다 생각함.

**2. 표지**

: 대출 연체 예측 모델 발표를 맡은 일요일 아침 팀 전현우입니다.

목차입니다. ~~ ~~ ~~ 하게 진행됩니다.

**3. 데이터소개**

1) 배경

- 저금리 우량고객 은행권, 고금리 불량고객 비은행권 사이 중금리 시장 진입을 위해 평가모델 필요

2) 데이터소개

- 고객정보, sci 한화생명, skt 가 있다

- 전체 고객중 96%가 미연체자, 4% 연체자가 있습니다.

**4. EDA 및 전처리**

1) 전처리

- 이상치제거

- 범주화(신용카드 발급 수 범주화)

: 연속형 변수 중 필요한 것에 대해 범주화를 했습니다. 그 중 하나인 신용카드 발급 수는 연체자 비율이 비슷한 7이상의 값을 하나로 묶어주었고 blah blah

+ 현우형 (여러 사례가 있는데 그 중에 신용카드가 있다)

- 정규화

: 연속형 변수 중 0인 비율이 높았습니다. 그래서 Log변환읋 해줬습니다. 연속형 데이터는 구간같은 것들이 단위가 큰데 로그를 사용하면 단위가 줄어들기 때문에 정규화를 했습니다.

- 표준화

: 직업별로 표준화를 진행하여 소득에 직업정보가 미치는 영향을 줄였습니다.

- correlation

:

2) EDA

- 최근 신용등급

: 95%가 신용등급이 나타나지 않았습니다. 연체자와 비연체자 간 특징을 보이는 변수에 대해 EDA를 해봤습니다.

- 2산업분류 대출건수

: 연체자 비율도 증가하는 경향이 있습니다.

- 대출총금액

: 비연체자에 비해서 연체자가 3천만원 이하의 소액규모 대출을 많이 받았습니다.

- 최근 1년 보험료 연체율

: 1년 보험료 연체율 역시 높을수록 연체자 비율 증가함

- 직업과 나이의 연체자 비율

: 20대 기업/단체임원, 70대 단순 노무직

- 통신요금 연간 최대 연체금액

: 대부분의 비연체자가 연간 최대 연체금액이 10만원 넘지못함, 반면에 연체자는 10만원을 넘는 비율이 더 높았다.

- 납부방법

: 4가지 납부방법이 있었는데 G납부방법에서 가장 높은 연체자 비율을 보였습니다.

- 신용카드 수

: 신용카드 건수가 증가할수록 연체자 비율이 급격하게 감소하는 모습을 보입니다.

- 남은할부금

: 연체자의 50%가 남은 할부금이 없었고 비연체자보다 높았습니다.

3) 파생변수 생성

- 유지기간 그룹화

: 다음으로 파생변수에 대해 설명하겠습니다. 히트맵으로 신용대출 유지기간을 탐색했습니다. 두기간 모두 3년 ~ 5년일 때 연체자 비율이 높은 특징을 보였고 교호작용을 반영하기 위해 앞의 두 변수를 반영하여 다음과 같은 파생변수를 생성했습니다.

- 건당 대출금액

: 또 신용에 따라서 한번에 빌릴 수 있는 금액이 다르기 때문에 총금액을 총 건수로 나누어 파생변수를 만들었습니다. 플랏을 보시면 연체자가 더 낮은 것을 알 수 있습니다.

- 단말기 대금 상환율

: 휴대폰 같은 경우는 금전적으로 여유가 있을 때 바꾸는 경우가 많은데 어느 정도 상환을 했냐를 척도를 나타내는 단말대금 상환율을 변수로 만들었습니다.

- 파생변수 생성 총 18개

: 연체자와 미연체자를 잘 구분할 수 있다고 판단되는 18개의 파생변수를 만들었습니다.

**5. 분석 방향 결정**

1) 주성분분석

- 다음으로 분석방향결정에 대해 소개하겠습니다. 대부분 기법에서 사용하는 차원 축소를 사용하지 않았습니다. 그 이유는 주성분 분석을 사용했을 때 Cumulative

데이터가 sparse 해서 0인 비율이 너무 많아서 PcA를 활용하는게 적절한 선택이 아닌 것 같다. 그리고 트리기반 모형을 사용하기 때문에 사용할 필요가 없었다.

2) 불균형 문제

- 다음으로 96 대 4의 불균형 데이터의 경우에 일반적인 분류 알고리즘에서는

불균형 데이터의 경우에는 비율이 낮은 class에 대해 민감하게 반응하지 않기 때문에 일반적인 경우에 좋은 성능이 나타나지 않습니다. 머신러닝 알고리즘에 추가적인 방법을 사용해 잘 분류하도록 해야한다.

- 불균형 데이터를 고려하는 평가 메트릭이 필요하다.

: 기존의 머신러닝 알고리즘은 accuracy를 사용하는데 전체 데이터 중 맞춘 데이터 비율을 이용하는데, 저희 데이터는 미연체자 보다는 연체자를 예측하는게 훨씬 중요합니다. 그래서 전체 중 맞춘 이를 기준으로 하는 accuracy 식 표현은 적절하지 않습니다.

F-measure 의 경우는

이런 불균형 데이터는 accuracy로 평가하기 안 좋다, 따라서 F-measure로

(현우형 추가)

기존에 사용하는 accuracy 같은 경우는

예상한 값 중에 맞은 비율이 accuracy, 이런 평가 척도는 지금 데이터에 좋지 않다. 그래서 f-measure를 했다.

실제 연체자 중에 얼마나 잘 예측했는지 찾는 지표와 우리가 예측한

recall : 실제 1 중에 예측해서 맞은 1

precision : 예측한 1중에 실제 1

- 불균형데이터 처리 : 데이터 전처리 과정에서 샘플링

: 다음은 샘플링을 사용한 방법입니다. Under, Over, Smote

언더 샘플링의 경우는 데이터의 크기를 줄이기 때문에 정보손실이 발생합니다.

오버 샘플링의 경우는 똑같은 데이터를 여러개 만드므로 과적합 위험

스모트는 KNN알고리즘 기반이기 때문에 High dimension에서 성능이 떨어집니다.

따라서 샘플링처리 하지 않았습니다.

- 불균형데이터 처리 : 모델 적합과정에서의 접근

: 부스팅 기법은 잘 분류해내지 못한 데이터에 집중하기 때문에 majority 보다는 minority에 더 집중할 수 있습니다. 그 이유는 대부분 틀리는 경우가 minority에서 생기고 여기에 가중치를 주기 때문이다. 그래서 저희는 GBM이나 XGB 의 부스팅 알고리즘을 사용햇습니다.

- 모델 평가과정에서의 접근(threshold)

: 불균형 데이터이기 때문에 majority class와 minority class의 weight 조절이 필요하다. 그래서 우리는 Threshold를 낮춰서 minority class가 발생했을 때 더 민감하게 반응하도록 했습니다.

**6. 모델 분석 프로세스**

- 데이터의 특성을 고려해서 알맞은 모델 선택

: 우리 데이터는 TARGET 0과 1을 구분하는 분류문제, 지도학습 Train과 Test 셋으로 지도 학습(Supervised Learning)으로 예측모델 구축했다.

우리 데이터는 범주형 변수도 많았고 연속형인 경우 0의 비율이 높아서 PCA를 쓰지 않음, 그래서 분석방향에서 말했듯 부스팅 모델을 사용했다.

- 모델링 프로세스

: 전체적인 모델링 프로세스는 다음과 같습니다. 데이터 전처리 & 파생변수를 만들고 모델로 학습시키고 성능을 비교하고 초적값을 찾아 최종 모형을 구축할 예정입니다.

- GBM 모델

: GBM은 부스팅 기법을 기반으로 약한 학습기를 결합해 오차를 줄여나가는 하나의 강화학습기를 만듭니다. GBM 사용이유 불균형데이터에서 boosting을 사용하기로 했기 때문에

#Hyper parameter#

n\_tree, Max\_depth : tree가 깊어지면 과적합이 우려되므로 조절하는 parameter,

learn\_rate : 부스팅 모델이 오류에 대해 예측을 하는데 다음모델에 어느정도 반영을 할지 정함,

stopping rounds, tolerance, metric : 조기종료 조건으로 학습을 할 때 우리가 제시한 metric에 어느 수준 향상되지 않으면 종료 -> 과적합을 막아준다.

- XGB

: GBM을 개선한 모델로 GBM에 비해서 정규화식이 추가되었다.

#Hyper parameter#

alpha : L1 정규화를 조절해줌 가중치의 크기에 대한 penalty(lasso)

lambda : L2 정규화 모수로 weight가 너무 커지는 것을 방지 (ridge)

- 데이터에 맞는 Hyper-Parameter 조절

: Overfitting과 Underfitting을 고려하며 직접 돌려보면서 최적의 Hyper-Parameter를 찾았습니다.

- 휴리스틱 방법으로 최적의 Threshold 찾기

: 다음과 같은 예측치를 얻었습니다. 예측치는 Threshold에 따라 다음과 같은 F-measure graph를 얻었습니다. 그 중 Flat하고 F-measure를 최대로 하는 걸 고려해서 찾았고 그게 0.2다. 0.5에 비해 0.36에서 0.46으로 크게 증가합니다.

- F-score

: 이를 고려한 Threshold 성능을 비교했는데 XGB가 높게 나왔습니다. 저희는 더 높일 수 있는 방법이 없을까 생각해봤습니다.

- Stacking을 통한 base model들의 앙상블

: 스태킹을 사용하기 위해서 RF, GBM, XGB, DNN을 만들어 사용했지만 이전의 XGB보다 개선된 결과를 보이지 않았다. xgb와 비슷한 결과, 하지만 계산속도가 느리고 XGB에 비해서 변수에 대한 효과를 설명할 수 없다. + 확장성

**7. 결과 해석 및 활용방안**

- 변수 중요도(Relative Importance)

: 트리를 만들 때 어ᄄᅠᆫ 변수가 어느정도의 효과를 보였는지 정량적으로 계산한 것입니다. 다음과 같은 결과가 나왔는데 가장 높은 것은 평균 통화시간, 빈도 등 SCI신용평가 한화생명 SKT 모두 고르게 높은 값을 얻었습니다.

- 변수가 연체 상환에 각각 어떤 영향을 미쳤을까?

: 앞과 같은 중요한 변수가 단독으로 모델에 어떤 효과를 주는지 PDP를 그려 확인했습니다. + PDP란? 가장 먼저 신용평가 데이터를 확인했는데 높은 변수를 가진 Ratio / CRDT / Per\_CARD / CRDT\_OCCR\_MDIF & ~ 를 확인했습니다.

#SCI 평가정보

Ratio\_LNIF

: 전체대출금액 / 전체대출건수의 파생변수로 한번에 얼마나 빌리는지 보는 파생변수입니다. 그 금액의 한도가 200만원 이상을 빌리는 경우 연체위험도가 급격하게 감소하고 있습니다.

CRDT\_CARD\_CNT

: 개설수가 증가할수록 연체위험도가 큰 폭으로 떨어진다.

Per\_CARD

: 질문들어오면 현우형

CRDT\_OCCR\_MDIF & SPTCT\_OCCR\_MDIF

: 교호작용으로 넣은 변수가 importance가 높게 나와서 효과적으로 시각화하기 위해서 3차원으로 표현했다. 두 유지기간에서 3~5년일 때 연체 위험도이 높다

#한화생명

PREM\_OVDU\_RATE

: 보험료 연체율이 40%이상일 때 대출 연체위험도가 확 증가한다

GDINS\_MON\_PREM / MAX\_MON\_PREM / FMLY\_TOT\_PREM

#SKT

AVG\_CALL\_FREQ / AVG\_CALL\_TIME / MOBL\_PRIN

LT1Y\_MXCO\_AMT (최근 1년 납부요금 연체금액 중 최대 연체금액)

: 연간 최대연체금액이 2만원 이상일 때 급격하게 연체 위험도이 올라간다.

**8. 활용방안 (마지막)**

- 신용평가사

: 다른 데이터를 활용했을 때 결과가 더 좋은 것을 확인했습니다. 이는 신용평가의 정확도나 신뢰도를 향상시킬 것입니다.

- 보험사

: 기존 한화생명의 중금리대출서비스는 제한된 대상이였습니다. 모델 도입 후 더 다양한 고객층을 확장할 수 있습니다.

- 통신사

: 신용평가에 통신데이터의 중요도와 역할을 확인했다. 이종 산업에서 활용될 수 있도록 Quality가 보장된다.