제 1회 산학연계공모전

(주) 플랫포머스__클린베테랑_고객-매니저 매칭 성공여부 예측

Ya ho~!

빅데이터경영통계전공

#이상우 #윤경서 #신예주

목차

#1 데이터 분석

기업에서 제공한 데이터에 대한 분석을 바탕으로 진행 방향 설정

#2 데이터 전처리

데이터의 결측치와 이상치 처리

#3 Feature Engineering

데이터를 활용하여 유의미한 feature 생성

#4 Modeling / 평가지표

CatBoost와 ExtraTrees를 이용한 모델링 / 내부성능 확인을 위한 추가적인 평가지표 이용

#5 가중평균

모델링 후 submission의 예측률을 높이기 위한 여러 시도

#6 아쉬운 점

공모전을 진행하며 느꼈던 아쉬웠던점

#01

데이터분석

매니저의 근무환경 — **매칭성공률**

매니저의 근무환경이 좋을수록 고객에게 제공되는 서비스의 질이 높아질 것이라고 가정

-> 매니저와 관련된 데이터를 이용하는 것에 초점을 맞춤

#02
데이터 전처리

삭제

- 접수시각
  - #결측치가 90% 이상

변경

TEST 만 해당

- 반려동물
  - #test에만 있는 값을 train과 같게 변경
- 매니저성별
  - #test의 매니저 성별 데이터 1값이 하나라, 0으로 변경

결측치

- 평수
- 부재중여부
- 우선청소
- 결제형태
- 반려동물
- CS교육이수여부
- 청소교육이수여부
- 매니저사용휴대폰
- 매니저이동방법

CatBoost : 각 열값의 최빈값으로 채움

ExtraTrees : 결측치를 살려서 범주형 - 미응답
수치형 - 0.5

#03
Feature Engineering
- Cat Boost

날짜형식

서비스 주소
근무가능지역

mean
encoding

고객 및 매니저
개인정보

날짜, 시간
고객, 매니저

범주형 feature
56개

수치형 feature
43개

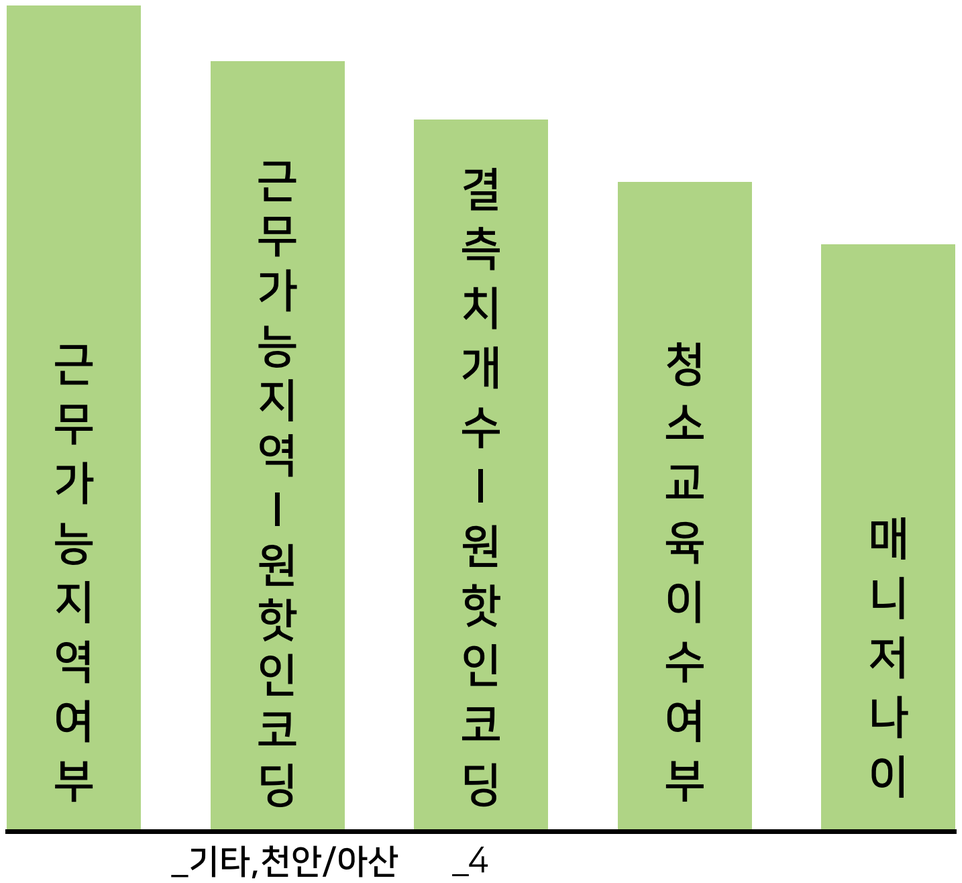
총 '99개'
피쳐 생성

범주형 변수 원핫인코딩 후
-> 3076개의 피처

#03

Feature Engineering
- Cat Boost

<feature 중요도>



'3076개의 피처'

shap 이용

importance
0보다 큰 값

'총 915개의 피처'

-> 처음 생각했던 분석 방향성에 맞게 매니저와 관련된 열을 사용하여 만든 feature가 높은 importance를 보임

#03

Feature Engineering
- Extra Trees

날짜형식

고객 및 매니저
개인정보

서비스 주소
근무가능지역

mean
encoding

범주형 feature
44개

수치형 feature
43개

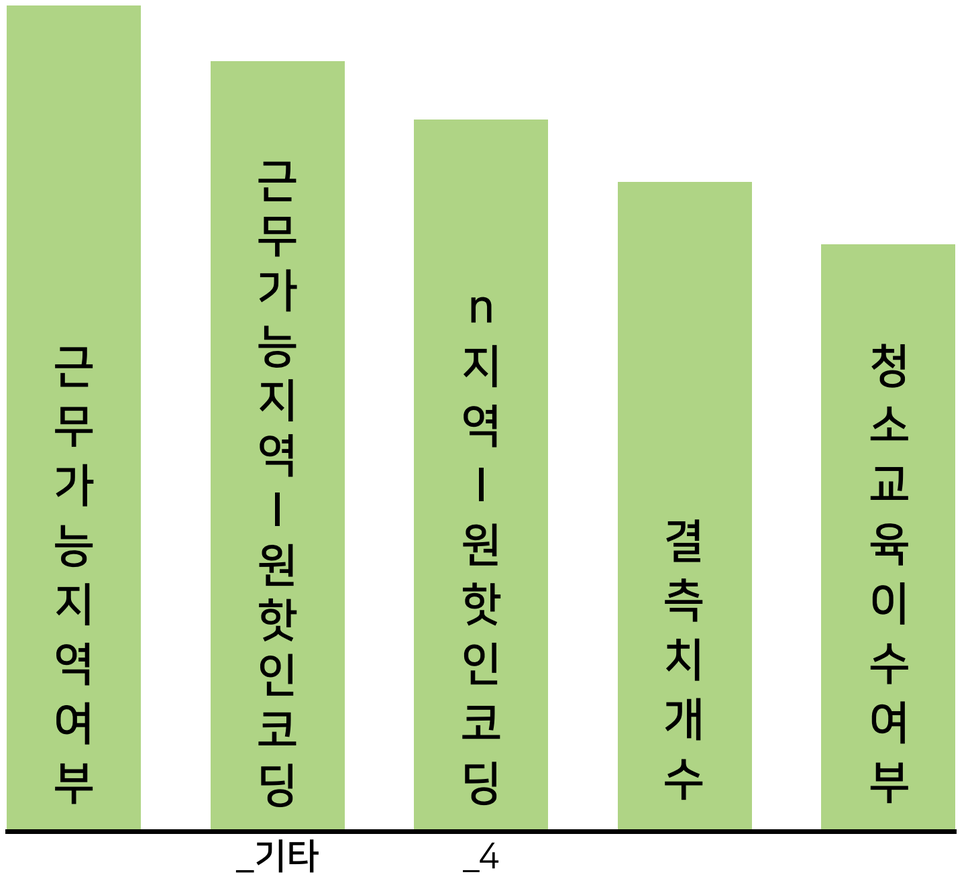
총 '87개'
피쳐 생성

범주형 변수 원핫인코딩 후
-> 821개의 피처

#03

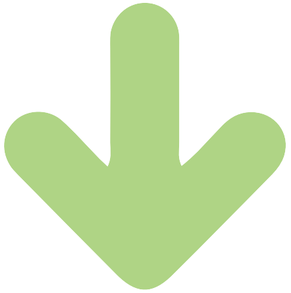
Feature Engineering
- Extra Trees

<feature 중요도>



'821개의 피처'

shap 이용



importance
0보다 큰 값

'총 535개의 피처'

-> 처음 생각했던 분석 방향성에 맞게 매니저와 관련된 열을 사용하여 만든 feature가 높은 importance를 보임

#04
Modeling

#CatBoost

- 범주형 변수**多**

- 별도의 모델 튜닝 X (randomstate=0만 고정)
  -> 과적합의 위험성 ↓

※ 모델 튜닝을 하지 않아도 자체 성능이 매우 높아 튜닝을 할 경우,
train에 매우 과적합될 것이라 생각되어 모델은 튜닝하지 않고 사용

CatBoost

#04
Modeling

#Extra Trees

- 무작위성 ↑

- 별도의 모델 튜닝 X (randomstate=0만 고정)

  -> 과적합의 위험성 ↓

※ 모델 튜닝을 하지 않아도 자체 성능이 매우 높아 튜닝을 할 경우,
train에 매우 과적합될 것이라 생각되어 모델은 튜닝하지 않고 사용

Extra Trees

#04 평가지표

roc_auc의 내부 성능이 너무 높게 나와서
자체적으로 검증할 수 있는 평가지표를 찾음

ROC_AUC

x축은 매칭성공여부가 1인 사람을 올바르게 분류하는 비율
y축은 매칭성공여부가 0인 사람을 1로 분류하는 비율

Recall

실제로 1인 것 중 1이라고 분류한 비율로, x축과 비슷한 평가지표라고 생각되어 추가적으로 내부 평가지표로 사용

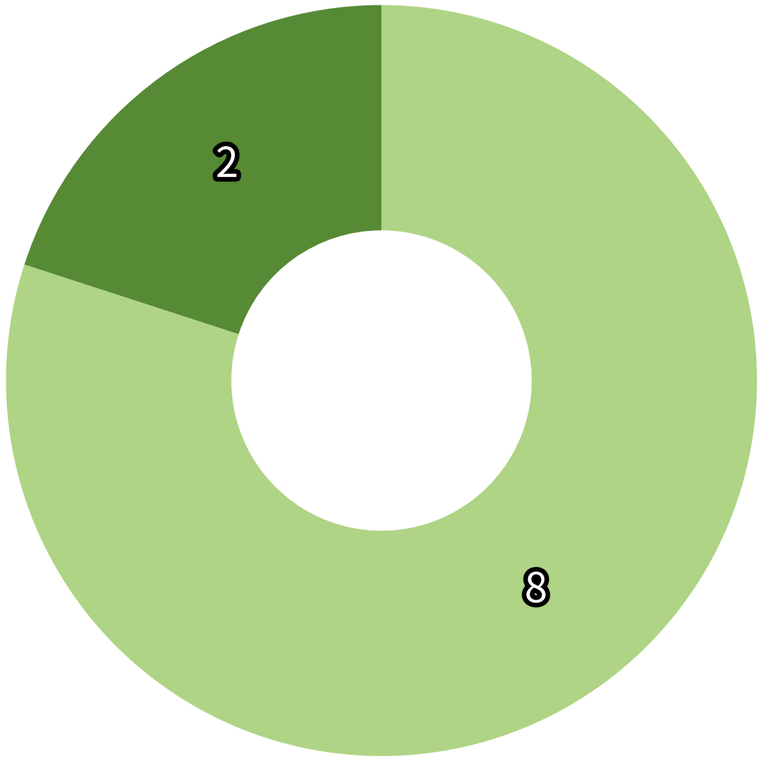
Precision

1이라고 분류한 것 중 실제 1인 것의 비율로, 실제 0인데 1이라고 예측된 것의 위험성을 최소화하는 것이 y축과 비슷하다고 생각되어 내부 평가지표로 사용

Average Precision

recall과 precision을 x축, y축으로 하는 precision-recall 그래프에서 선 아래 면적을 나타내는 것으로, 값이 높으면 성능이 좋을것이라고 생각하여 내부 평가지표로 사용

#05
가중평균



#가중평균

모델링을 통해 뽑아낸 각각의 파일들 중 성능이 더 높은 것에 가중치를 주는 방식으로 submission의 평균을 냄

ExtraTrees submission(8)

+

CatBoost submission(2)

- 서로 다른 계열의 모델을 섞음
  -> 서로 부족한 부분 보완

- Extra 성능 > CatBoost
  -> 8:2의 비율로 섞음
  가장 성능이 높았던 비율

#06
아쉬운 점

#Features

- 수치형 변수
- 접수시각 데이터 활용 X
- 고객정보, 매니저정보에 W2V (Word 2 Vec) 활용 X

#modeling

과적합을 방지하기 위해 파라미터 튜닝 X

-> 파라미터 튜닝을 통해 성능을 더 높일 수 있지 않았을까...

#DNN

DNN 모델을 만들기 위해 계속 시도했으나, 시간 부족 & 학습과정에서의 오류로 사용X

-> DNN모델을 활용을 했다면 성능을 더 높일 수 있지 않았을까...

THANK YOU

Ya ho~!