Catboost를 활용한

매니저 매칭 성공 여부 예측

제 1회 산학연계 공모전

TEAM - SOTA

18학번 권유진
18학번 박승주
18학번 최민석
18학번 한승수

목차

문제 정의

1. 매칭성공여부의 의미

서비스 사후에 남기는 평가로 기업 내부적으로 성공/ 실패로 분류

문제의 의도: 고객이 선호하는 조건에 맞는 매니저를 보내 성공확률을 높이자!

Key Idea1

고객의 수요에 맞는 매니저가 매칭됐을 때 성공적일 것이다.

-> **고객의 요청사항에 부합하는 매니저**인지 확인할 수 있는 피쳐를 생성한다.

Key Idea2

고객과 매니저들의 특성을 추출하고 조합한다.

-> PCA, Kmeans clustering과 같은 **특성 추출 알고리즘**을 활용한다.

Key Idea3

데이터에 수치형으로써 의미 있는 데이터가 많지 않았다.

-> **범주형 변수**들을 이용한 모델링이 중요할 것이다.

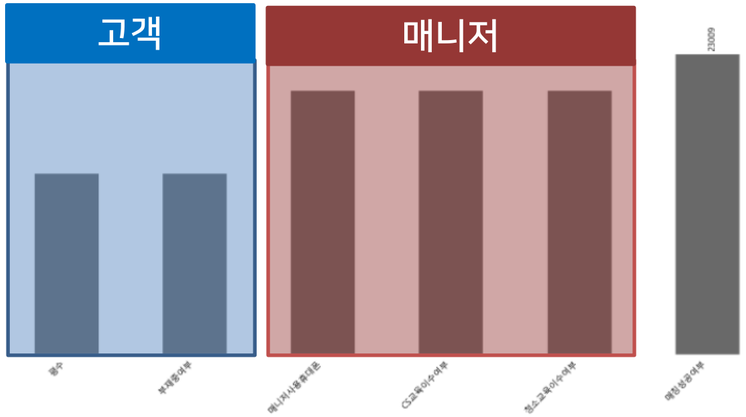
Preprocessing
Feature Engineering
K-means
Feature Selection
Modeling

1. Train, Test Dataset 변수 통일

고객ID
매니저ID
매니저주소
매니저최초가입일
매니저 최초서비스일

→ **변수 제거**

2. 결측값 처리



결측값들의 크기가 비슷한 변수들이 존재한다.

↓

“데이터 생성 과정에서 특정 개체의 답변이 존재하지 않는다”

결측값 여부도 고객과 매니저의 하나의 특징이라고 판단 → **결측치를 측정값과 다른 값으로 대체**

02 본론 – Feature Enginnering

1. 고객 관련 변수 생성

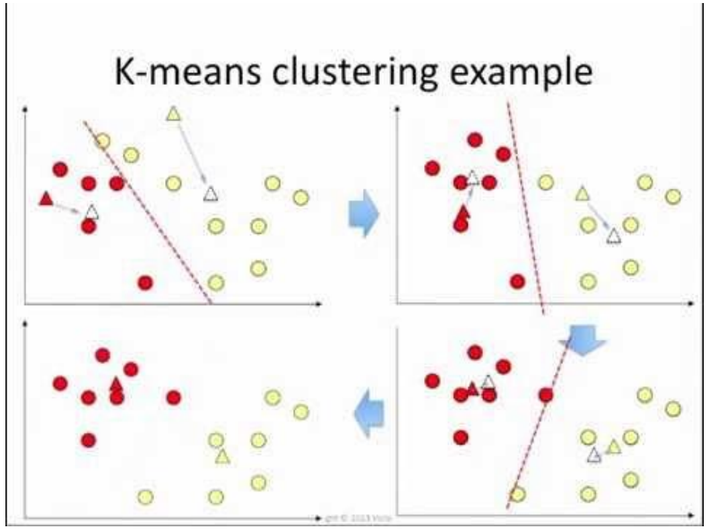
- 고객가입일 → 고객가입월, 고객가입일일, 고객가입일(숫자형)
- 최초서비스일 → 고객서비스월, 고객서비스일, 최초서비스일(숫자형)
- 접수일 → 접수월, 접수일일, 접수일(숫자형)
- 잔여회차, 현재회차 → 잔여회차, 종료여부
- 서비스시작시간, 서비스종료시간 → Endtime, starttime, betweentime

2. 매니저 관련 변수 생성

- 매니저생년월일 → 매니저생년월일(범주형), 나이대
- 서비스주소, 근무가능지역 → ser_add_dae, ser_add_so, so_가능, dae_가능, 지역_가능‘
- 부재중여부, 부재중서비스가능여부 → 부재중 청소

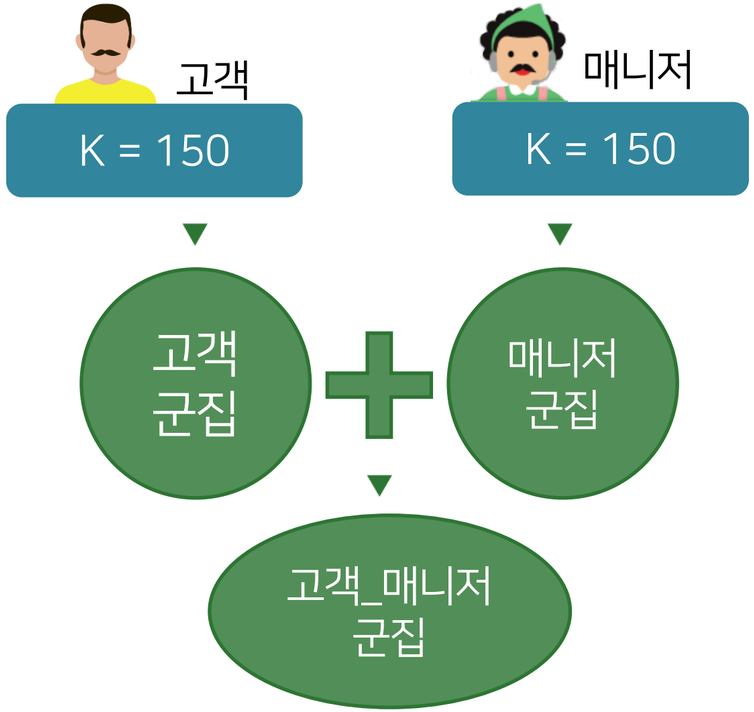
K-means Clustering

각 군집의 평균(mean)을 활용하여 K개의 군집을 형성하는 알고리즘.
여기서 평균은 각 클러스터의 중심과 데이터들의 평균 거리를 의미



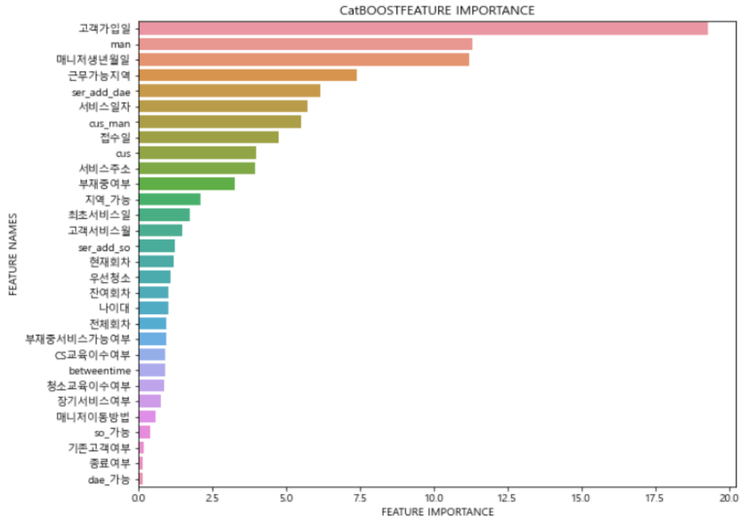
매칭성공여부 = 고객과 매니저의 상호작용

특정 고객 군집과 특정 매니저 군집의 관계 파악 중요!



1. Feature importance 활용

Model 학습 과정에서 중요도가 낮은 피쳐들을 제거하여
가장 높은 성능을 보이는 피쳐셋을 찾는 방법



2. Catboost Select_features

Catboost 자체 내장 함수로 후진소거법을 이용하여
가장 높은 성능을 보이는 피쳐셋을 찾는 방법.

3. 직관에 따른 Feature Selection

많은 피쳐들 중, 유의미할 것으로 생각되는 피쳐들을 자체적으로 필터링한다.

3가지의 방법들 중, 역설적으로 **직관적인 피쳐셀렉션**이 가장 높은 성능을 보임

CatBoost

범주형 변수에 특화된 부스팅 기법

- Information Gain이 동일한 feature combination
- 낮은 Cardinality 가질 시, One-hot Encoding 사용 (default-2)
- 높은 Cardinality 가질 시, Ordered Target Encoding 사용

43개의 Feature 모두 범주형 변수 처리
➡ CatBoost에서 압도적으로 높은 성능

- Ordered target Encoding

Mean Encoding의 Data Leakage 문제를 해결한 방법
- 현재 데이터의 인코딩하기 위해 이전 데이터들의 인코딩 된 값을 사용한다.

- One-hot Encoding

범주형 변수에 고유 index를 부여하여 출연여부에 따라 1과 0으로 채우는 방법

Trainset 과 testset의 차이

제공된 testset과 trainset에 질적 차이가 존재함

- Validationset, hyper-parameter 조정 등의 중요성 하락
- 과적합의 위험성이 높아짐

➡
1. 학습 시 모든 traindata를 이용하여 모델에 학습
2. Randomsearch, gridsearch 등 하이퍼파라미터 조정 X

결론 및 결언

1. K-means 활용

고객과 매니저의 군집 특성에 따라 매칭성공여부가 달라진다.

2. Catboost

범주형 데이터가 많고, 피쳐 생성에 한계가 있을 경우에 뛰어난 성능을 보이는 모델링 방식이다.

3. 직관에 따른 Feature Selection

Feature Selection 과정에서 모델의 결과 뿐만 아니라 도메인 지식을 활용하면 좋은 피쳐셋이 된다.

발표를 들어주셔서 감사합니다

제 1회 산학연계 공모전

TEAM - SOTA