SCDC 공모전
분석 계획서

DATAVITA

고지형 박재우 이형선

삼성카드

프로세스

전체적인 분석 과정은 크게 아래의 세 파트로 나누어 진행

① 데이터 전처리 ② 모델링 ③ 결과 해석

데이터 전처리

① 데이터 불균형 처리
- SMOTE 기법 활용, 오버샘플링

② 분포 맞춤형 변수 생성
- K-Bins & GMM 군집화

모델링

① 변수 선택
- 변수 중요도 및 유의도 기반

② 모델 최적화
- 스태킹 앙상블 기법 활용
- 여과 모델로 LightGBM 활용

결과 해석

① 모델 해석
- Shapley Value
- Global Surrogate Model

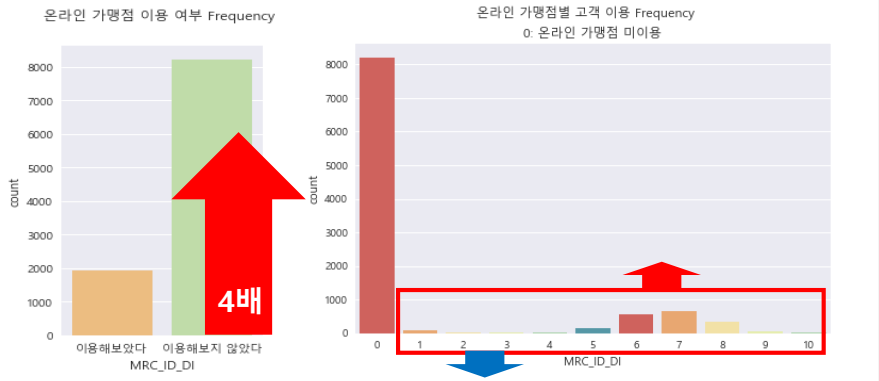
② BI 도출
- (B2B) 비즈니스 솔루션
- (B2C) 고객 맞춤형 카드 추천

데이터 전처리

데이터의 분포를 탐색하여 ① 타겟 변수의 불균형 ② 가맹점 이용에 따른 분포 차이 발견

① SMOTE 기법을 활용한 오버샘플링 ② K-Bins & GMM 군집화를 해결책으로 고안

타겟 변수의 불균형 분포

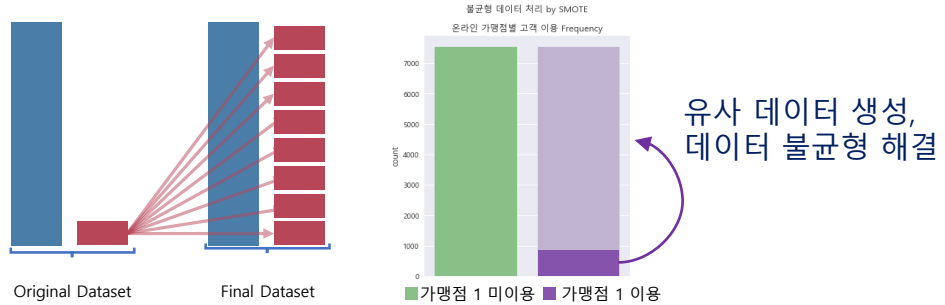


가맹점 이용에 따른 분포 차이

구간에 따라 가맹점 이용에 따른 분포 차이가 명확하게 나타나는 변수 발견

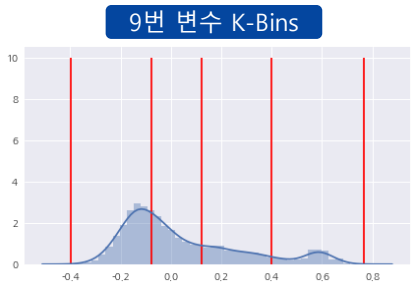
SMOTE 기법을 활용한 오버샘플링

SMOTE(Synthetic Minority Over-Sampling Technique)

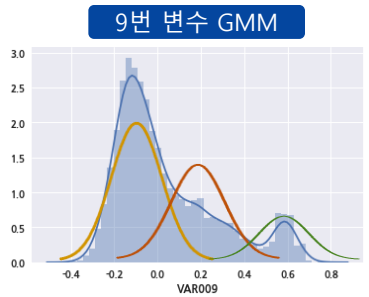


유사 데이터 생성,
데이터 불균형 해결

K-Bins & GMM 군집화를 이용한 변수 생성



K-Means 군집을 통해
K개의 구간으로 변수 카테고리화

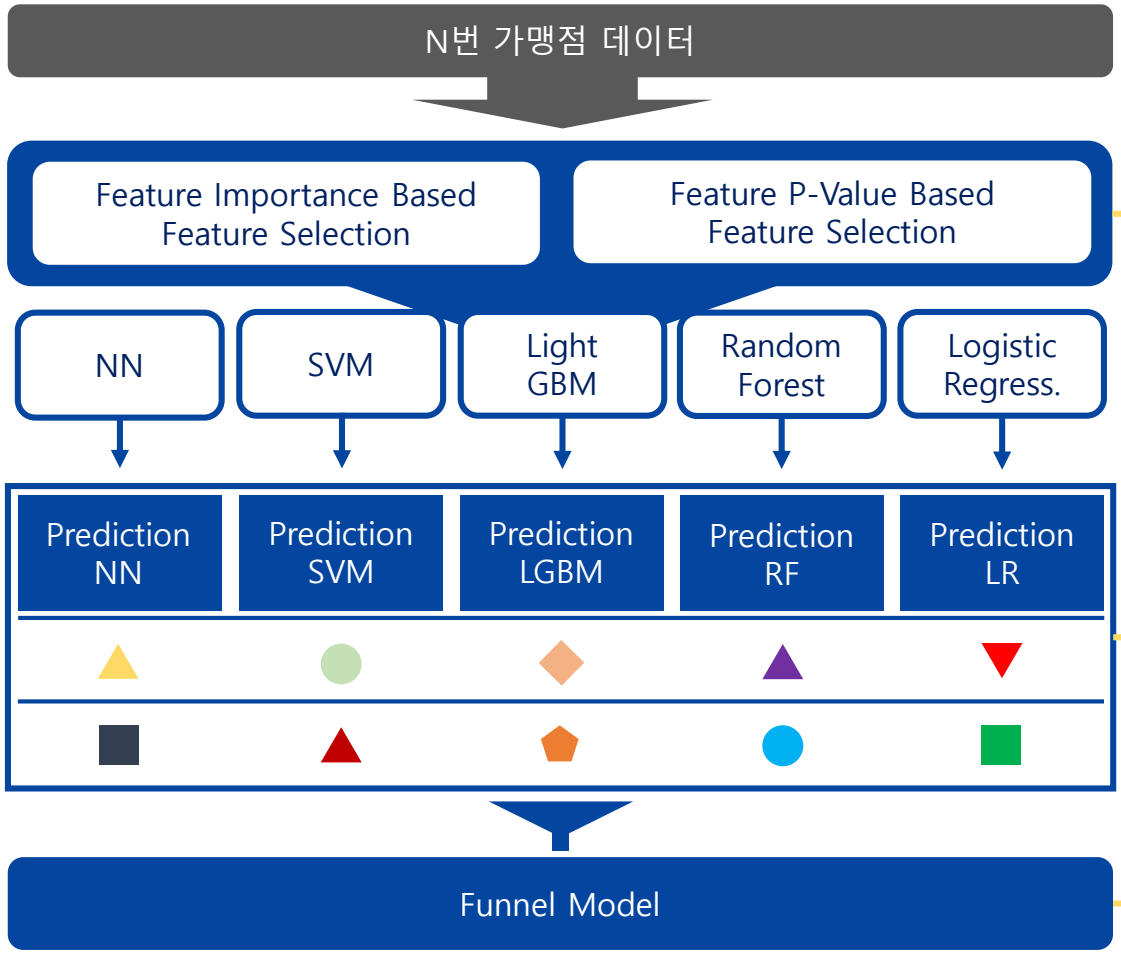


BIC를 기준으로 정규분포 적합,
K개의 구간으로 변수 카테고리화

모델링

가맹점 맞춤형 모델을 학습하기 위해 다음과 같은 과정으로 모델링 진행

① 유의미한 변수 선택 ② 성능 향상을 위한 모델 최적화



① 변수 선택
- 가맹점 별로 전처리가 완료된 데이터 입력
- 다음과 같은 기준으로 유의미한 변수 선별
  - 트리 기반 모델에서 추출 가능한 변수 중요도
  - Lasso 회귀 모델에서 추출 가능한 변수 유의도

② 모델 최적화
- 스태킹 앙상블(Stacking Ensemble) 활용
  - 여러 모델의 예측 결과를 변수로 재학습하는 기법
  - 선형, 비선형 학습법을 반영하여 예측 일반성 확보
- 스태킹 앙상블 적용을 위한 데이터셋 생성

②-2) 여과 모델
- 여러 모델의 결과를 종합하여 여과(Funnel)하는 모델
- LightGBM 알고리즘 사용
  - 각 모델의 중요도 판별 가능
  - 대용량 데이터에서 빠른 학습 속도
  - 정확도가 입증된 부스팅 계열 모델

결과 해석

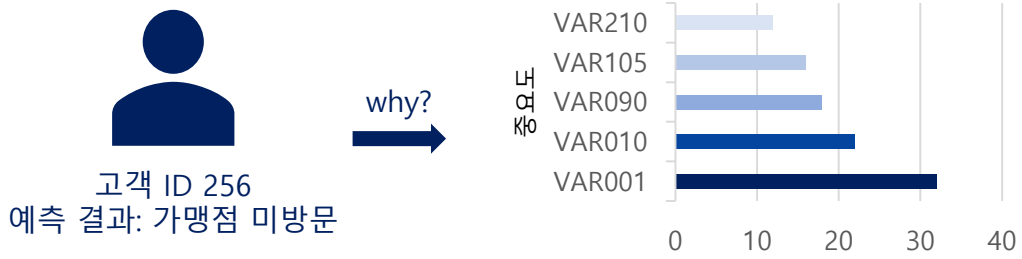
해석 가능한 인공지능 방법론을 활용하여 복잡한 모델 해석

① Shapley Value로 변수의 중요도 분석 ② Global Surrogate Model로 변수의 영향력 추론

해석 가능한 인공지능(XAI, eXplainable AI)

- 예측 정확도가 우수한 인공지능 모델은 대게 복잡성이 그에 비례하므로 해석이 어려움 ⇒ 다양한 XAI 방법론이 고안됨
- 학습된 모델과 데이터 샘플을 활용, Shapely Value와 Global Surrogate Model을 통해 복잡한 모델을 해석하고자 함

Shapley Value



- 모든 가능한 변수 조합에서 어떤 변수가 얼마나 기여했는지 측정
- 모델 전반과 데이터 각각의 해석 가능
- 계산 비용이 변수의 수에 비례하기 때문에, 적절한 샘플링을 통해 해결

⇒ 각 변수의 중요도 측정 가능

Global Surrogate Model

- 유사한 모델로 원래 모델을 대신(Surrogate) 설명
- 선형 모델을 사용한다면 각 변수의 회귀계수 도출 가능
- 설명력이 강한 대리 모델을 탐색하기 위해 다수의 샘플링을 통해 선택

⇒ 각 변수의 변화에 따른 방문 확률 증가율 측정 가능

결과 해석

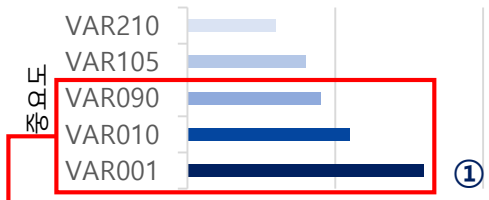
B2B와 B2C 측면에서 각각 BI 도출 및 아이디어 제시

① 변수 해석을 통한 비즈니스 솔루션 제안 ② 설득력 있는 고객 맞춤형 카드 추천

Case 1. 가맹점 A

비즈니스 솔루션 제안

? 가맹점 A 이용 고객은 어떤 특징을 보일까요?



① 변수 중요도 기준 상위 N개 변수 추출

VAR010이 증가할 때 이용 확률은 10배 증가
VAR090이 증가할 때 이용 확률은 1.2배 증가
VAR001이 증가할 때 이용 확률은 3배 감소

② 영향력이 큰 변수를 타겟팅하는 마케팅 기획 방향 수립

Case 2. 고객 B

설득력 있는 고객 맞춤형 카드 추천

(기존) 고객이 직접 선택한 교통, 주유, 음식 등의 혜택을 조합하여 추천

(개선) 고객의 구매 가능성이 높은 가맹점 위주로 맞춤 카드 추천

(차별점) 고객의 주요 행동 패턴을 함께 제시하여 설득