#1. Project - H백화점 고객 성별 예측 (장동원)

**〮 Background**

H백화점 관련 데이터를 토대로 H백화점의 고객 성별 예측

**〮 Summary**

1. Data preprocessing

* EDA(탐색적 자료 분석)

계절별, 시간대별 구매건수에서 Null 값 발견. 구매 건수와의 비교통해 Null값을 0으로 치환.( # **train\_data.fillna(0, inplace = True)**)

* Features analysis(변수 분석)

1. 데이터 값이 비대칭임을 발견(남:여=3:7 정도) [원형 그래프/막대형 그래프 통해 시각화] →데이터의 남녀 비율 조절 필요성 발견
2. 상관관계 분석[**matplotlib** 활용] → 성별과 관련성이 없는 id값과 음의 상관관계를 나타내는 변수 제거.
3. Gender 값에 따른 각 변수들의 변화를 시각화하여, 새로운 파생변수에 관한 인사이트 도출 → 대표적으로,

“평균구매가격을 기준으로 남녀의 데이터를 시각화 하였을 때, 최빈값은 여자의 비율이 더 높게 나타나지만, 구매가격이 증가할수록 남자의 비율이 조금 더 높아진다” 라는 해석 도출

* 위의 결과를 통해 평균가격을 기준으로 한 파생변수 추가(‘고객등급’)
* Features engineering(변수 조작: 파생변수 추가)

1. 평균구매가격을 기준으로 한 ‘고객등급’을 새로운 변수로 추가함(평균구매가격 기준으로, 20%마다 나누어 5등급 분류)
2. 주구매코너를 기준으로 각 코너의 여성비율을 만들어 변수화

* 비율에 따라 5등급 부여(1~0.8: 1등급, …, 0.2~0: 5등급)
* 예측 위해 주구매코너의 value값(str형식)을 category화 통해 숫자로 변경

1. Model & Algorithms(모델 및 알고리즘)
2. TomeKLinks, SMOTETomek : 데이터의 남녀 비율 비대칭 해소위해 적용

* SMOTETomek 사용 시 더 균형있게 샘플링 됨.

1. Voting ensemble, Averaging predictions, stacking, Bagging

* LogisticRegression, DecisionTreeClassifier, KNeighborsClassifier, SVM

모델 등을 조합하여 예측함

1. AdaBoost, Gradient Boosting, XGboost

* 분류 이후 가중치 조절을 통한 정확한 예측 수행. XGboost를 사용했을 때가 가장 높은 accuracy(0.788)을 나타냄.

1. Cross Validation(K-fold cross validation) 통한 신뢰도 검증

* 가장 예측률이 좋았던 XGboost를 모델로 Cross Validation한 결과,

0.006의 분산값이 나와 높은 신뢰도를 검증함.

1. Review

유의미한 파생변수를 만들어내는데 어려움이 있었음. 더 유의미한 파생변수가 있었다면 예측 정확도를 더 높일 수 있지 않았을까 하는 아쉬움이 있음. H백화점의 마케팅 전략, 매장별 특성과 같은 도메인적 지식이 더 주어졌더라면 더 유의미한 파생변수를 만들어낼 수 있었을 것임.

1. Report

Jupyter notebook with Python을 토대로 H백화점 고객 성별 예측을 수행함.