**Reselling platform**

Cosumer to consumer 온라인 쇼핑몰에서 구매하는 소비자의 특징을 알아보고, 구매를 활성화하기 위한 방안을 도출해보고자 합니다.

**Contents**

먼저, 데이터 셋을 살펴보겠습니다. 두 번째로, 예측모델을 만들고 성능을 확인합니다. 마지막으로, 모델을 해석해보겠습니다.

1. **Exploration**
   1. **Problem**
      * 분석 대상은 베스티에를 콜렉티브입니다. 이 온라인 쇼핑 사이트에서는 중고명품을 사고팔 수 있습니다.
      * C2C 플랫폼이 성장하기 위해서는 거래액이 늘어야하는데, 이번에는 구매자의 특성을 살펴보면서 구매가 늘어나기 위한 요소를 확인합니다.
   2. **Data set**
      * Data set은 캐글에서 수집되었고, Vestiaire Collective의 사용자 정보가 담겨있습니다.
      * 전처리 이전에 98,913 사용자별 24개 특성이 있었습니다.
      * 결측치와 중복치는 없었고,
      * 의미가 중복되는 것과 특성 값이 너무 다양한 특성을 삭제하여 15개 특성만 남겼습니다.
      * 분석에 활용된 data set은 98,913개 사용자 계정별 15개 특성입니다.
      * **(특성 기술)**
      * 특정에는 “사용자의 선호 언어, 팔로워, 팔로잉 수, 좋아요를 표시한 제품, 판매하지는 않았지만 올려 둔 제품, 판매한 제품, 검수 통과율, 관심 상품 수, 구매 상품, 성별 칭호, 어플리케이션 종류, 프로필 사진 업로드 여부, 마지막 로그인 후 지난 일 수, 가입 경과일 수”가 있습니다.
      * 우리의 타겟은 사용자가 구매한 제품 수를 활용하여, 한 개라도 구매한 이력이 있으면 구매자로 분류하고, 구매 이력이 없으면 비구매자로 분류합니다.
   3. **Exploratory data analysis**
      * 전체 사용자 중 구매한 사용자는 5.5%뿐으로 아주 작다는 걸 알 수 있습니다.
      * 구매자 5,419계정 중에서도, 대부분인 5,000여명이 5개 이내로 구입하고 있습니다.
      * 반면에 소수의 계정에서 수백개의 중고물품을 구매하여 구매간의 극명한 비대칭을 보이고 있습니다.
2. **Model**
   * + 데이터의 특성을 살펴봤으니, 이제 사용자가 구매를 할 지에 대한 예측 모델을 만들어 보겠습니다.
   1. **Baseline**
      * 우선, 기본적인 모델을 만들어서 예측 모델이 이 기준모델보다 예측 정확률이 높아질 수 있도록 설계하고자 합니다.
      * 기본 모델은 Logistic regression을 이용했습니다. Logistic regresiion은 특성과 타겟의 선형 관계에서 확률로 전환하고, threshold를 정하여 그 threshold보다 높은 경우만 구매자로 분류하고, 낮으면 비구매자로 분류하는 방법입니다.
      * 이 모델로 구매자를 예측해서, 평가 데이터로 정확도를 확인합니다.
      * 정확도를 확인하는 기준으로 recall을 이용하겠습니다.
      * Recall은 전체 구매자 중 모델이 구매자로 분류한 비율을 의미합니다.
      * 이 모델의 recall은 75%로 나오는데, 전체 구매자 중 75%만 구매자로 분류되었고, 25%는 비구매자로 잘못 분류한 것 입니다.
      * 75%를 우리의 기준 성능으로 정하겠습니다.
   2. **Improved model**
      * 두 가지 모델을 설계했습니다.
      * **(Random Forest, recall is 0.3)**
      * 먼저, Random Forest 모델을 설계했습니다. Random Forest는 Tree를 여러 개 만들어 보고, Tree들의 다수결 결과를 통해 사용자를 구매자와 비구매자로 분류합니다.
      * R.F는 복원 추출로 다양한 data set을 이용하고, 특성도 무작위로 추출하고 Tree들을 만들기 때문에, 특정 data set에만 맞는 과적합 문제를 완화할 수 있습니다.
      * 이 모델로 예측한 결과, 평가데이터의 구매자중 30%만을 맞추었고, 구매자 중 70%는 비구매자로 잘못 분류했습니다.
      * 이 것은 기준 모델의 성능인 75%보다 낮기 때문에 이용할 수 없습니다.
      * **(Gradient boosting decision tree, recall is 0.76)**
      * 다음으로, 살펴볼 모델은 GBDT입니다. 역시 여러 tree를 생성하여 타겟을 예측하는 모델입니다.
      * Tree의 마지막 노드인 leaf의 수를 제한하여, 학습 데이터 셋에만 적합하게 되는 과적합 문제를 완화하고,
      * 먼저 생성한 Tree의 잔차를 줄여나가는 Tree를 계속 만들어서 합치칩니다. 이렇게 하면 학습 데이터 셋에서 예측도가 떨어지는 과소적합 문제도 완화할 수 있습니다.
      * 일반적으로 강한 성능을 자랑하여, 많이 쓰이는 모델입니다.
      * GBDT모델의 recall을 확인 해 보니, 평가 데이터 셋의 구매자 중 76%를 구매자로 올바르게 맞췄고, 24%는 비구매자로 잘못 분류했습니다.
      * **(Improved model)**
      * GBDT가 baseline 모델보다 근소하지만 recall 성능이 향상되었기 때문에, GBDT를 이용해서 구매 확률을 예측하도록 하겠습니다.
3. Interpretation
   * + 이제는 모델에서 중요하게 보는 특성이 무엇인지 살펴보며 모델을 해석해보겠습니다.
   1. Interpretation of features
      * 특성중요도는 특정한 특성이 없어지거나, 원래 데이터와 달라질 때, 모델의 성능이 얼마나 변하는지를 확인하여 특성이 얼마나 중요한 관련이 있는지 확인하는 기준
      * 특성중요도를 보면, socialProductsLiked부터 상위 4개가 급격하게 높아지는 것을 알 수 있습니다.
      * ‘hasIosApp’ 특성까지의 중요도 증가율에 비해서, 갑자기 2배 정도로 증가합니다.
      * 따라서 상위 4가지 특성이 구매확률과 가장 연관이 깊다고 정하겠습니다.
      * 4가지 특성은; 사용자를 팔로우 하는 수가 가장 영향이 크고, 다음으로는 로그인을 한 이후 경과일수, 관심 상품에 등록한 제품 수, 사용자가 좋아요를 표시한 제품 수 입니다.
      * 주의할 점은, 특성중요도가 양의 상관관계, 음의 상관관계 여부, 인과관계를 의미하는 것은 아니라는 점입니다.
      * 따라서 특징이 달라지면 구매확률에 어떻게 영향을 미치는지를 다음에 살펴보겠습니다.
      * **(PDP, ICE 100)**
      * 개별조건부 기대 (ICE)는 다른 특성은 모두 동일하고, 하나의 특성만 변화시켰을 때, 구매확률이 어떻게 변하는지 확인할 수 있습니다.
      * 무작위로 100개 사용자를 뽑아서, 4가지 특성을 각각 변화시켜 본 ICE입니다.
      * PDP는 전체 구매자의 ICE를 평균낸 것으로, 전체 데이터셋에서의 평균적인 관련성을 나타냅니다.
      * **(팔로워 수와 구매확률의 관계)**
      * 팔로워 수는 특이하게도 1~2명 정도만 들어나면 구매확률이 증가하지만, 그 이후에는 감소합니다.
      * 구매를 해서 follower가 늘어난 것인지, follower가 늘어나서 구매가 늘어난 것인지 불확실하여 추가 연구가 필요합니다.
      * **(로그인 경과일수, 관심상품 수와 구매확률의 관계)**
      * 상식에 부합하는 결과입니다.
      * 로그인을 최근에 한 활동적인 사용자일수록 구매 확률이 높고, 관심 상품이 많은 사용자일수록 구매로 이어질 확률이 높을 것 입니다.
      * **(좋아요를 한 제품 수와 구매확률의 관계)**
      * 좋아요를 많이 누를수록 구매 확률이 낮아진다는 것은, 상식에 맞지 않습니다.
      * 아마도 좋아요가 무리하게 많은 경우가 불가능하여, ICE가 잘못 예측한 것 같습니다.
      * 하지만, 정확한 판단을 위해서는 역시 추가 연구가 필요합니다.
   2. Action proposal
      * **(두 가지 특성의 PDP)**
      * 두 가지 특성을 조합하여 변화시키고, 구매확률이 어떻게 변화하는지를 보는 히트맵입니다. 히트맵 안의 수치는 구매확률을 의미합니다.
      * 구매확률이 80% 이상인 것만 확인 해 보겠습니다.
      * **(관심상품)**
      * 관심 상품으로 많이 넣을수록 구매 확률이 증가하기 때문에, 고객의 참여도를 높여 관심상품을 많이 넣을 수 있도록 유도해야 합니다.
      * **(최근 로그인)**
      * 최근 로그인을 했을수록 구매확률이 높아지기 때문에, 계정의 활동성을 높여야 합니다.
   3. 3가지 제안
      * 관심상품을 추가하는 단계까지 오면, 구매 확률이 매우 높아지기 때문에, 소비자의 구매 퍼넬에서 intent 단계로 진입시키기 위한 노력에 집중합니다.
      * 흥미로운 컨텐츠를 지속적으로 제공하는 등의 방법으로 로그인을 계속 유도합니다.
      * Social product like가 많다고 구매확률이 계속 증가하지는 않기 때문에, 구매 의도를 추측할 수 있는 활용성이 높은 지표로 재설계 할 필요가 있습니다.

(발전시킬 부분)

3가지 제안, 해석 부분의 인과관계가 설득력이 떨어짐