|  |  |
| --- | --- |
| **분석 주제명** | **앱 사용성 데이터를 사용한 대출 신청 예측 및 고객 군집 분석을 통한 서비스 세분화** |
| **분석 내용**  **요약** | 대출 신청 여부 예측을 위해 주어진 데이터셋을 EDA하여 특성을 파악한 후 파생 변수 생성, 결측치 보간 등의 과정을 거쳐 데이터셋을 전처리하고 통합한다. 이후 Catboost, XGBoost, LightGBM, DNN 모델로 예측 모델의 성능을 비교한 후 최종 모델을 결정한다. 다음으로 BRICH, DBSCAN 알고리즘을 사용하여 대출 신청을 한 고객과 하지 않은 고객을 대상으로 군집을 분석한 후 유의미한 정보를 통해 핀다의 발전을 위한 서비스를 제안하고자 한다. |
| **분석방법** | **[user\_spec 데이터셋을 통한 유저 특성 분석]**   1. **수치형 변수의 기초통계량과 시각화를 통한 데이터셋의 분포도 확인**   먼저 주어진 user\_spec 데이터셋의 기초통계량 확인을 통해 해당 데이터셋에 특별한 이상치가 존재하는지 확인했다. 기초통계량 상에서는 특별히 눈에 띄는 이상치를 발견하지는 못하였다. 보다 명확한 데이터의 분포를 확인하기 위해 user\_spec 데이터 내 각 피처들의 히스토그램을 시각화하고 각 피처들 간 상관관계를 히트맵으로 나타내었다.  히트맵으로 데이터셋 내 각 컬럼들의 상관계수를 나타내는 히트맵을 확인해 보았을 때, 피처들 간의 상관관계가 높지 않음을 확인할 수 있었다. 이에 따라 상관계수를 높일 수 있는 파생변수 생성 및 변수선택법의 활용이 필요하다 판단하였다.   1. **파생변수의 생성**   분포도 확인을 통하여 user\_spec 데이터셋 내에서 범주화 및 세분화 할 수 있는 피처들에 대하여 파생변수 컬럼을 생성하였다.  • **[‘birth\_year’] -> [‘age\_cat’]**   * 기존의 birth\_year 컬럼에서 나이에 대한 컬럼을 생성 한 후, 이를 다시 나잇대로 변환한 컬럼이다.   **• user\_id.duplicated() -> [‘freq’]**   * 데이터셋 확인 결과, 동일한 user\_id 를 가진 고객이 여러 번 유저 스펙을 기입한 정황을 확인했다. 이에 따라, 동일한 고객이 유저 스펙을 중복하여 기입한 빈도를 나타내는 컬럼을 생성했다.   • **[‘company\_enter\_month’] ->[‘ work\_year’]**   * 기존의 company\_enter\_month 컬럼에 기반하여 유저가 2022년을 기점으로 직장 내에서 일한 연차를 나타내는 컬럼을 생성했다. * **[‘credit\_score’] -> [‘score\_type’]** * 기존의 credit\_score 컬럼에서 연속적인 숫자의 분포로 나타나져 있는 값들을 KCB 방식에 따라 등급화 시켰다.  1. **결측치 처리** 2. **파생변수 생성 전** 3. 파생변수를 생성하기에 앞서 기존의 user\_spec 데이터셋 내에서 결측치 비율을 확인했다. 결측치 비율에 따라, 결측치가 10% 이내의 ‘birth\_year’(1%), ‘gender’ (1%), ‘credit\_score’(7%) 변수의 경우 결측치 행을 모두 drop 시켰다. 4. 상대적으로 결측치 비율이 높은 ‘personal\_rehabilitaion\_yn’, ‘personal\_rehabilitaion\_complete\_yn’ 컬럼의 경우, 고객의 특성 기재란에 개인회생자 여부가 필수 기재 사항이 아닌 핀다 어플의 특성을 반영하여 개인회생자 여부를 기재하지 않은 고객의 경우, 개인회생자가 아닐 확률이 높을 것이라 판단, 개인회생자 여부가 결측치일 경우 개인회생자가 아닌 고객이라 판단하고 이에 따라 결측치를 처리하였다. (fillna(0)) 5. ‘personal\_rehabilitaion\_complete\_yn’ 컬럼의 경우에도, 만약 개인회생자가 아니라면 이를 변제할 의무 또한 없으므로 ‘personal\_rehabilitaion\_yn’ 컬럼이 0인 경우, 변제 유무와 관계가 없는 유저들이므로 ‘personal\_rehabilitaion\_complete\_yn’ 컬럼을 해당 컬럼과 전혀 관계가 없는 숫자인 -99로 채워주었다. 6. **앞서 설명한 파생변수 ‘ age\_cat’, ‘freq’ ,’score\_type’ 의 생성** 7. **파생변수 생성 이후 결측치 처리** 8. 파생변수의 생성 이후, 결측치를 재확인하고, 여전히 결측치가 남아있는 ‘work\_year’, ‘existing\_loan\_cnt’, ‘existing\_loan\_amt’ 를 보간하기 위하여 iterativeImputer 모델을 활용했다. 9. 모델을 활용하기 이전에 데이터셋 내의 범주형 변수들을 원핫 인코딩을 통해 더미화시키고, 각 피처 간의 상관계수 기반의 회귀 예측 모델인 iterativeImputer 을 통해 훈련, 결측치들을 보간해주었다. 10. **‘credit\_score’ 컬럼을 기반으로 파생변수 ‘score\_type’ 을 생성**   **[log\_data 데이터셋 활용 방안]**  여윳돈 계산기 서비스 이용 고객은 대출 상환을 목적으로 두는 경향을, DSR 계산기 서비스 이용 고객은 추가 대출을 목적으로 두는 경향이 있음을 파악하였다. 따라서 고객(user\_id) 별로 여윳돈 계산기 서비스 혹은 DSR 계산기 서비스의 이용 빈도를 추출한 후 고객 정보(user\_spec) 데이터셋에 추가하였다.  **[loan\_result 데이터셋 가공]**  loan\_result는 예측해야 하는 is\_applied 변수를 포함한 데이터셋이다. 추후에 user\_spec 데이터셋과 함께 is\_applied 변수값을 예측하기 위한 모델을 실행할 때 사용할 예정이다. 이를 위해 loan\_result 데이터셋을 전처리와 변수를 추가한 user\_spec 데이터셋과 application\_id를 기준으로 합친 이후 loan\_result에 없지만 user\_spec에는 있는 데이터를 먼저 drop을 시키고 loan\_limit과 loan\_rate가 결측치인 행을 drop을 시켜주었다. *또한, 유저별 낮은 대출 금리와 기대출금액과 대출한도 사이의 연관성이 있다고 판단하여 이를 활용한 유저별 가장 낮은 대출 금리 / 현재 application\_id의 금리,* *현재 application\_id의 대출한도 / 유저의 기대출금액을 파생변수를 생성하였다.* 이후 user\_spec과 loan\_result의 데이터를 합쳐 is\_applied가 결측값인 행은 target 셋으로 분리하였다.   1. **변수 선택법의 활용**    * 1. 기준으로 log\_data 데이터셋 내의 여윳돈계산기 서비스, DSR계산기 서비스의 빈도수를 추가하여 user\_spec 데이터를 재생성한다.      2. 해당 user\_spec 데이터셋 내의 다양한 유저 스펙 컬럼들 중 변수들의 중요도를 기준으로 모델링에 필요한 컬럼들을 추출하고자 한다.      3. 이에 따라 최종 종속 변수인 loan\_result 데이터셋의 ‘is\_applied’ 컬럼을 user\_spec 데이터셋에 ‘application\_id’를 기준으로 추가시키고, ‘is\_applied’ 컬럼을 target 변수로 설정, 사이킷런의 SelectKBest 모듈을 활용해 변수 선택 모델링을 시행한다. 여기에서 변수 선택의 기준은  chi2 방식을 활용해 카이제곱 검정통계량을 기준으로 20개의 컬럼을 추출하였다.         + 1. 여기서 변수 선택의 기준을 카이제곱 통계량으로 둔 것은, 해당 데이터셋 내에서는 타 기준과 대비해 카이제곱검정이 변수 간 관련성 여부를 판단하는 기준점으로 적합하다 판단했기 때문이다.      4. 변수 선택법 모델을 돌린 후 대출 신청 여부라는 해당 대회의 주제를 고려하여 변수를 다시 한 번 선택하는 과정을 거쳤다.      5. 이후 ‘PrepayCal\_freq’, ‘existing\_loan\_cnt’, ‘existing\_loan\_amt’, ‘gender’, ‘personal\_rehabilitation\_yn’, ‘personal\_rehabilitaion\_complete\_yn’ 컬럼은 제거하였다.  * *변수 선택법으로 선택된 변수들을 1차로 걸러준 후, 대출 신청 예측이라는 주제를 고려하여 만든 파생 변수와 선택된 변수들을 추가하였으며 범주가 많은 purpose 변수는 빈도수를 고려하여 대환대출, 생활비, 기타 3가지로 줄여 주었다.*   **[예측 모델]**  예측 모델의 경우 여러 개 모델을 선정해 성능을 비교하고자 한다. 머신러닝의 Catboost, XGBoost, LightGBM 모델과 딥러닝의 DNN 모델을 선정하였다.  *예측하려는 is\_applied의 분포를 살펴본 결과 0과 1의 비율이 17:1으로 엄청난 클래스 불균형을 띄고 있어 모델 예측에 어려움이 있을 것이라 판단하였다. 따라서 SMOTE + ENN 방식을 활용한 복합 샘플링 방식으로 1과 0의 비율에 큰 차이가 나지 않도록 맞추어주는 과정을 거쳤다.*  *이후 데이터셋을 확인한 결과 중복되는 샘플의 개수가 약 300만 개 정도 있음을 확인하였고 똑 같은 데이터의 가중치를 부여하는 것과 동일한 역할을 하여 과대적합의 우려가 있으므로 삭제해주었다.*  머신러닝 모델의 세 알고리즘은 모두 Gradient Boosting 기반으로, 대출 신청 여부의 0과 1의 비율이 불균형하다는 데이터 불균형의 문제를 고려하여 scale\_pos\_weight라는 하이퍼 파라미터의 조정을 통해 예측의 성능을 높이고자 한다. *이외의 하이퍼 파라미터 튜닝의 경우 Optuna 라이브러리를 사용해 자동탐색을 통해 최적의 하이퍼 파라미터를 찾고 하이퍼 파라미터 튜닝 후에도 성능을 저해하는 변수를 찾기 위해 Permutation Importance를 계산하여 예측 모델에 반영할 계획이었으나 데이터셋에 과적합되는 양상을 띄어 머신러닝 모델을 최종 모델로 선정하지 않았다.*  DNN 모델의 경우 입력 데이터의 개수에 맞게 차원을 구성하여 층에 맞게 줄여 마지막에 sigmoid 활성화 함수를 통해 출력값을 이진으로 출력할 수 있게 구성한다. Epoch의 값을 충분히 설정해준 후 early stopping을 통해 validation F1-score가 가장 높게 훈련된 후에 일정 에포크 수 이상 모델의 성능이 좋아지지 않는다면 훈련을 중지할 예정이며 ModelCheckpoint 기능을 활용해 가장 높은 validation F1-score를 기록한 모델을 저장해 예측을 진행하였고 4개의 모델 중 높은 validation F1-score를 기록하였고 예측 분류 역시 본래의 데이터셋과 비슷한 비율을 유지하여 이를 최종 예측 모델로 선정하였다.  **[군집 분석]**   1. **군집 분석 알고리즘**   *대용량 데이터에 적합한 birch와 DBSCAN 모델을 사용하려 하였으나 컴퓨팅 자원의 한계로 모델링이 돌아가지 않아 분석 방향을 변경하였다. 본 연구진은 범주형 변수들이 user에 대한 정보를 많이 담고 있다고 판단하여 범주형 변수들로 군집 분석을 할 수 있는 알고리즘을 찾아 사용하였다.**이렇게 고객의 특성을 파악하여 서비스를 제공하고자 변수를 선택한 결과, K-modes clustering방법을 사용하여 군집화 모델을 구축하였으며 클러스터 개수는 elbow method로 결정하였다.* |
| **분석결과 활용 및 시사점** | 군집 분석을 한 결과, 클러스터 별로 특징이 뚜렷하게 구분되었다. 고객의 특성을 반영한 변수들은 성별과 대출 신청의 목적에 따라 대출 승인 여부가 확실하게 구분되었다. 또한 소득 수준에 대한 변수들도 직업 형태와 소득 종류, 신용 점수에 따라 유의미한 차이를 보였다. 두 경우 모두 클러스터가 2개로 구분되었는데, 대출이 승인된 집단과 아닌 집단으로 나뉘었다. 대출이 승인된 집단은 주로 고객이 남자였고, 20-30대의 젊은 층이 비중이 높았다. 또한 정규직 종사자가 많아 안정적인 소득을 얻는 사람이 많았고, 신용 점수도 높았다. 반대로 대출 승인이 나지 않은 집단은 성별이 여자고, 40대가 주를 이루며 직업의 형태가 안정적이지 않고 그에 따라 신용 점수도 낮은 사람들이 대부분이었다.  개인 소득에 따른 대출 신청 결과는 당연하다고 볼 수 있으나, 개인 특성에 따른 변수들의 군집 분석 결과에서 대출 목적이 생활비 마련이나 대환대출의 비중이 크고, 젊은 층의 이용률이 높다는 점은 주목할 만하다. 따라서 이 유형의 고객 맞춤 서비스를 적극적으로 제안하는 것이 필요하다. 또한 서비스 이용률이 낮은 중장년층과 소득 수준이 낮은 집단에도 대출 신청 서비스의 장벽을 낮추기 위한 대출 관리 서비스나 어플리케이션 사용 가이드 등의 서비스를 제안하는 방식도 적절한 서비스 제안으로 볼 수 있다. |