Title

신용카드 사용자 데이터를 활용한 신용카드 연체 예측

Student ID, Name
2019251126, 황성아

Introduction

최근 신용카드 시장은 급속하게 성장하고 있습니다. 특히 디지털화의 발전으로 인해 모바일 신용카드나 인터넷 신용카드 등 새로운 유형의 신용카드도 출시되고 있고, 이를 이용하는 소비자들이 늘고 있습니다. 하지만 금융감독원이 발표한 바에 따르면 148 개의 국내 여신전문금융회사의 연체율은 2019 년말 1.68%, 2020 년말 1.26%, 2021 년 말 0.86%으로 점차 줄어드는 추세였지만 2022 년 말 증가세로 반전해 1.25%가 되었다고 합니다. 신용카드 연체는 고객 뿐만 아니라 카드 회사에서도 다양한 문제점을 초래합니다. 이자 및 연체로 지불 문제, 채권 회수 문제, 운영 비용 증가, 평판 손상 등 다양한 문제가 발생할 수 있습니다. 따라서, 신용카드 회사들은 고객의 연체를 최소화하려고 노력하고 있으며, 이를 방지하기 위해서는 신용카드 발급 시 적정한 기준을 적용할 필요가 있습니다.

따라서 본 프로젝트의 목표는 신용카드 연체에 어떠한 요인과 밀접한 연관이 있는지 알아보아 신용카드 발급 시 기준을 더욱 효율적으로 설정하며, 정확한 신용카드 연체 예측 모델 구축을 통해 신용카드 연체 가능성이 큰 고객을 사전에 예측할 수 있도록 하는 것 입니다. 이를 통해 신용카드 연체를 사전에 방지하여 연체율이 감소하고 신용카드 회사들이 겪고 있는 문제점들이 최소화되기를 기대합니다.

Data

본 프로젝트에서 사용하게 될 데이터는 '신용카드 사용자들의 개인 정보 데이터'로, 데이콘(Dacon)이라는 사이트의 '월간 데이콘 신용카드 사용자 연체 예측 AI 경진대회'에서 사용된 데이터 중 'train'데이터 셋을 다운로드하여 사용할 예정입니다.

사용할 데이터는 20 개의 변수와 26457 개의 행으로 이루어져 있으며, 이 데이터셋의 변수는 gender(성별), car(차량 소유 여부), reality(부동산 소유 여부), child_num(자녀 수), income_total(연간 소득), income_type(소득 분류), edu_type(교육 수준), family_type(결혼 여부), house_type(생활 방식), DAYS_BIRTH(출생일), DAYS_EMPLOYED(업무 시작일), FLAG_MOBIL(핸드폰 소유 여부), work_phone(업무용 전화 소유 여부), phone(전화 소유 여부), email(이메일 소유 여부), occyp_type(직업 유형), family_size(가족 규모), begin_month(신용카드 발급 월)과 같은 신용카드 사용자들의 개인 정보 독립 변수들과 예측하고자 하는 사용자의 카드 대금 연체를 기준으로 한 credit(신용도) 종속 변수로 구성되어

있습니다. 종속변수 credit(신용도)은 0,1,2 값으로 구성되어 있으며, 값이 낮을 수록 높은 신용의 사용자들 의미합니다.
[데이터 링크 : https://www.dacon.io/competitions/official/235713/data]

Methods

가장 먼저 시각화를 통해 데이터를 탐색하여 보고 시각화를 바탕으로 전처리 진행, 그리고 전처리를 진행 한 데이터를 바탕으로 예측 모델링을 하였습니다. 모델링 부분에서 범주형 변수의 경우 Encoding, 수치형 변수의 경우 표준화를 진행하여 Train/Valid/Test Set 으로 나눈 후, XGBoost, LightGBM, Random Forest 3 가지 알고리즘을 활용해 예측 모델을 구축하고 정확도를 측정하였습니다. 또한, Grid Search 와 Random Search 를 활용하여 파라미터를 최적화 해 보았습니다.

[분석 기법 설명]

1. XGBoost
   : 트리 기반의 앙상블 학습에서 가장 각광받고 있는 알고리즘 중 하나로, 일반적으로 머신러닝 알고리즘 중 뛰어난 예측 성능을 발휘한다. 병렬 CPU 환경에서 병렬학습이 가능하다는 점에서 기존 GBM 에 비해 빠른 수해 성능을 보장한다. 또한 자체에 과적합 규제 기능이 있으며 tree pruning 기능으로 더 이상 긍정 이득이 없는 분할을 가지치기를 통해 분할 수를 더 줄이는 장점도 있다. 반복 수행 시마다 내부적으로 학습 데이터 세트와 평가 데이터 세트에 대한 교차 검증을 수행해 과적합을 방지할 수 있다.

2. LightGBM
   : XGBoost 와 함께 부스팅 계열 알고리즘 중에서 가장 각광받고 있는 알고리즘 중 하나로, XGBoost 보다 예측 수행 시간이 빠르고 메모리 사용량이 적다는 점 등에서 XGBoost 의 단점을 보완한 알고리즘이다. Root Node 와 가까운 Node 를 우선적으로 대칭 분할하는 다른 트리기반 알고리즘과는 다르게 Loss 변화가 가장 큰 Node 를 지속해서 비대칭 분할한다. 이를 통해 깊은 비대칭 트리를 생성하여 예측 오류의 손실을 감소할 수 있다. 또한, 카테고리형 변수의 자동 변환과 최적 분할이 가능하다.

3. Random Forest
   : 앙상블 알고리즘 중 비교적 빠른 수행 속도를 자랑하며, 기본 결정 트리보다 일반화 된 알고리즘이다. 룰이 명확하고 트리 별 시각화가 가능하다는 장점이 있다.

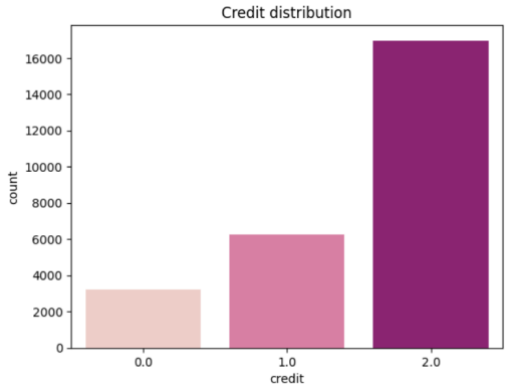
[Hyper parameter 최적화 방법]

1. Grid Search
   : 하이퍼 파라미터의 후보를 정하여 각 후보 값으로 모델을 학습시켰을 때 가장 성능이 좋은 하이퍼 파라미터의 조합을 선택한다.

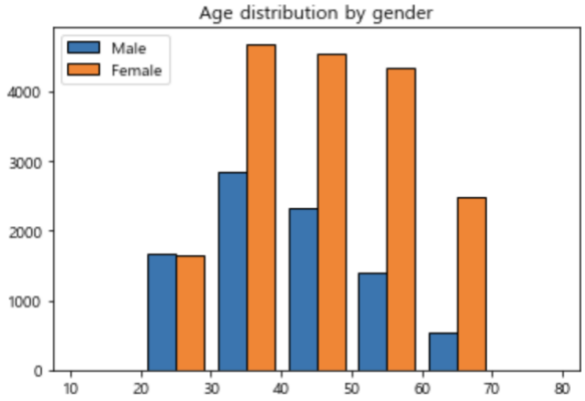
2. Random Search
   : Grid Search 에 비해 불필요한 반복 수행 횟수를 대폭 줄이면서, 동시에 정해진 간격(grid) 사이에 위치한 값들에 대해서도 확률적으로 탐색이 가능하므로, 최적 하이퍼 파라미터 값을 더 빨리 찾을 수 있다.

Experimental Results

1) 시각화

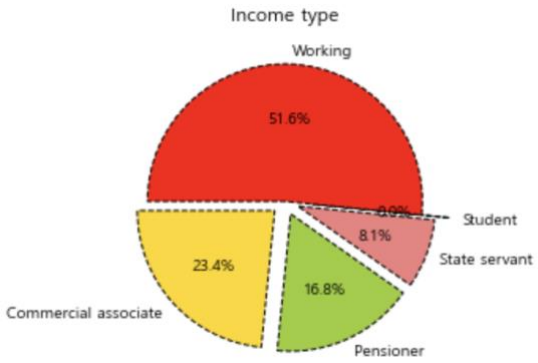


<그림 1>

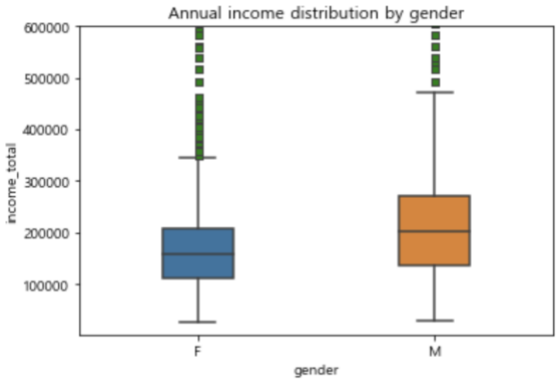


<그림 2>

<그림 1> 의 결과에 따라, 0 일 수록 신용이 높은 사람을 나타내므로 고 신용자의 수가 가장 낮고 저 신용자의 수가 가장 많은 편임을 알 수 있습니다. <그림 2>에 따르면 남성이 여성보다 약 2 배 많고, 30~40 대에 가장 많이 분포해 있는 것을 알 수 있습니다.



<그림 3>



<그림 4>

<그림 3> 에 따르면 고객 대부분은 수입이 있는 상태이며 학생의 수는 거의 없는 것을 확인할 수 있습니다. <그림 4>를 보면 남성이 여성보다 연간 소득이 높은 편이고, 여성은 주로 100000~200000, 남성은 주로 150000~300000 구간에 속해 있는 것을 확인할 수 있습니다.

2) 전처리

- 의미 없는 변수 제거: index 는 분석에 의미가 없고, flag_mobil 값은 모두 1 로 동일하므로 필요가 없다고 판단하여 제거하였습니다.
- 이상치 처리 및 파생변수 1 생성: child_num 의 변수의 이상치는 3 으로, family_size 의 이상치는 5 로 변경 후, 두 변수의 상관관계가 높아 하나의 변수 famchild_sum 으로 합쳐 파생변수를 생성하였습니다.
- DAYS_EMPLOYED 가 양수이면 무직으로 판단하여 0 처리: DAYS_EMPLOYED 변수는 일을 한 날짜만큼의 음수로 표현되어 있는 변수이므로 음수이면 무직이라 판단하여 0 으로 대체하였습니다.
- 음수 값을 양수 값으로 변경 및 파생변수 2 생성: 데이터 수집 당시를 0 이라 두고 날짜를 역으로 세어 값이 음수로 표현되어 있는 변수들을 양수 값으로 변경 후, 상관관계가 높은 DAY_BIRTH 와 DAYS_EMPLOYED 변수의 차이 값으로 파생변수(before_EMPLOYED)를 생성하였습니다.
- 결측 값 처리: occyp_type 변수에 결측값이 존재하기 때문에 DAYS_EMPLOYED 변수에서 무직이라고 판단한 데이터를 무직이라는 카테고리를 만들어 할당하여 결측 값을 채우고 나머지 결측 값은 데이터 오류라 판단하여 삭제하였습니다.

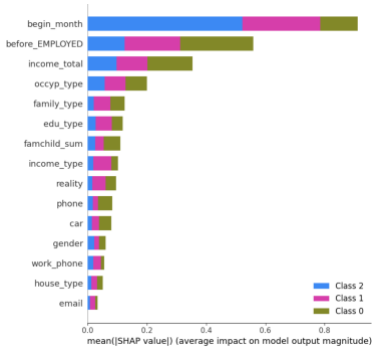
3) 모델링

1. XGBoost 결과

```
xgb = XGBClassifier(random_state = 42)
xgb.fit(X_train, y_train)
xgb_preds = xgb.predict(X_test)
```

```
#정확도
print('accuracy:{:0.4f}'.format(accuracy_score(y_test, xgb_preds)))
```

```
accuracy:0.7031
```



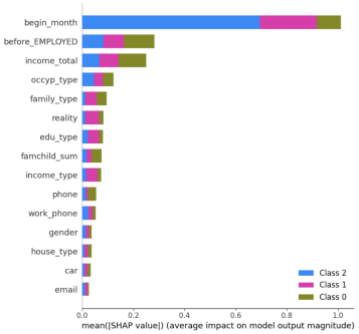
XGBoost 알고리즘을 활용해 모델링 한 결과, 정확도는 0.7031 이며, begin_month, before_EMPLOYED, income_total 변수의 중요도가 top3 로 가장 높은 것으로 나타났습니다.

2. LightGBM 결과

```
lgb = LGBMClassifier(random_state=42)
lgb.fit(X_train, y_train)
lgb_preds = lgb.predict(X_test)
```

```
#정확도
print('accuracy:{:0.4f}'.format(accuracy_score(y_test, lgb_preds)))
```

```
accuracy:0.7004
```



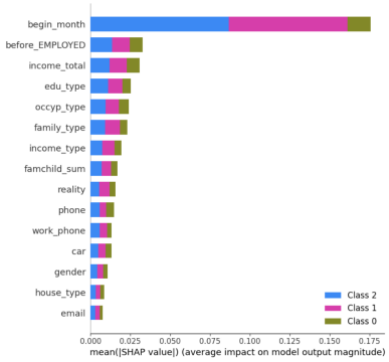
LightGBM 알고리즘을 활용해 모델링 한 결과, 정확도는 0.7004 이며, XGBoost 에서와 동일하게 begin_month, before_EMPLOYED, income_total 변수의 중요도가 top3 로 가장 높은 것으로 나타났습니다.

3. Random Forest 결과

```
rf = RandomForestClassifier(n_estimators=3000,
                            max_depth= 16,
                            random_state = 42)
rf.fit(X_train, y_train)
rf_preds = rf.predict(X_test)
```

```
#정확도
print('accuracy:{:0.4f}'.format(accuracy_score(y_test, rf_preds)))
```

```
accuracy:0.7058
```



Random Forest 알고리즘을 활용해 모델링 한 결과, 정확도는 0.7058 이며, 마찬가지로 동일하게 begin_month, before_EMPLOYED, income_total 변수의 중요도가 top3 로 가장 높은 것으로 나타났습니다.

Discussion and Conclusions

위의 세 가지의 알고리즘을 통해 예측 모델을 구축하여 본 결과, Random Forest 알고리즘을 활용하여 최적화한 모델의 정확도가 0.7058 로 가장 높았습니다. 또한, 위에서 진행한 세 가지 모델의 변수 중요도 top3 는 모두 begin_month, before_EMPLOYED, income_total 변수로 나타났습니다. 따라서, 신용카드 연체와 가장 연관이 깊어 신용카드 발급 시 주의 깊게 살펴보아야 할 정보는 신용카드 발급 월, 태어나서 일을 하기 전까지의 기간, 연간 총 소득이라고 판단했습니다.

이번 분석을 바탕으로 연체 가능성이 높은 고객을 식별하여 리스크 관리 및 대응이 편해지고, 연체로 인하여 발생하는 비용이 절감되고, 연체 관리 서비스의 질적인 발전을 통하여 사용자의 고객 서비스가 개선되며, 고객 신용 위험을 더 잘 관리하여 시장 경쟁력을 강화하는 등 예산안의 효율화가 가능해 질 것이라고 생각합니다. 신용카드 연체를 사전에 방지하여 연체율이 감소하고 신용카드 회사들이 겪고 있는 다양한 문제점들이 최소화될 것 입니다.

References

- *데이터 변수 설명*. DACON. (n.d.). Retrieved May 5, 2023, from https://www.dacon.io/competitions/official/235713/talkboard/402821/
- Newsis. (2023, May 2). *높아지는 연체율에... 금융사들'역대급' 충당금 쌓는다*. newsis. Retrieved from https://newsis.com/view/?id=NISX20230502_0002288266&cID=15001&pID=15000
- 이혜승. (2004). 데이터 마이닝 기법을 이용한 신용카드 연체고객 예측 및 감소 방안 제시 (국내석사학위논문). 고려대학교 대학원, 서울.
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*, 3146–3154.