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

본격적인 분석에 앞서 데이터 셋을 살펴본 결과 occyp_type 변수에는 결측 값이 존재하였고 DAYS_BIRTH, DAYS_EMPLOYED, begin_month 변수의 값들은 데이터 수집 당시를 0 이라 두고 날짜를 역으로 세어 값이 음수로 표현되어 있기에 분석에 불편함이 예상되는 등 전처리가 필요해 보여 가장 먼저 간단한 전처리를 진행할 것 입니다.

간단한 전처리를 진행한 후에는 변수 간 다양한 시각화를 통해 수치형 변수의 경우 값들이 어떠한 분포를 띄고 있는지, 범주형 변수의 경우 어떤 범주에 많은 데이터들이 해당되어 있는지, 또한 독립 변수들 간 어떠한 관계를 띠고 있는지, 어떠한 요인들이 종속변수에 가장 영향을 크게 미치고 있는지 등을 나타내 볼 것입니다. 이를 바탕으로 중요한 인사이트를 뽑아내고 파생변수를 생성하거나 범위를 정해 데이터를 묶는 등 추가적인 전처리도 진행할 것 입니다. 또한, K-means 와 같은 군집분석 기법을 활용하여 신용카드 사용자들을 군집으로 묶어 나타내고 각 고객 군집이 어떤 특징을 띄는지 확인해볼 것 입니다.

최종적으로는 신용카드 사용자들의 개인 정보 독립변수들을 활용하여 카드 대금 연체를 기준으로 한 신용도를 예측하는 예측 모델을 구축하여 대금 연체 가능성을 예측해 볼 계획입니다. 독립변수는 수치형 변수와 범주형 변수가 모두 존재하고, 종속변수는 3 개의 범주로 이루어져 있으므로, 다중분류 모델을 채택하여 진행할 것 입니다. 다중분류에 사용할 알고리즘은 다양한 머신러닝 알고리즘들 중 뛰어난 성능을 자랑하고 있는 트리 기반의 Random Forest, lightGBM, XGBoost 이며, Grid Search 나 Random Search 와 같은 최적화 방법으로 하이퍼 파라미터를 최적화 후 성능 비교를 통해 가장 성능이 좋은 최종 모델을 채택할 것 입니다.

본 프로젝트에서 예상되는 어려움 첫 번째로는 전처리 부분에 있을 것 같습니다. 전처리를 어떻게 하는지에 따라 분석 전반적으로 영향을 많이 미치기 때문에 전처리에 가장 많은 시간을 쏟아 고민해볼 예정입니다. 두 번째로는 군집분석 부분에 있을 것 같습니다. 사용해야할 변수들도 많은 편이고, 데이터 형태도 수치형과 범주형이 혼합되어 있는 상태이기 때문에 군집화가 잘 돼서 의미 있는 해석을 내야하는 부분에서 어려움이 예상됩니다.

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.