

분류 기획서

은행 이탈 데이터를 사용한 이진 분류 분석

분석 목적

- 은행 고객 이탈에 영향을 주는 변수를 파악하여 고객 이탈을 예방하는 방안을 마련한다.

대회 개요

- Binary Classification with a Bank Churn Dataset**

Binary Classification with a Bank Churn Dataset

Playground Series - Season 4, Episode 1

[k https://www.kaggle.com/competitions/playground-series-s4e1](https://www.kaggle.com/competitions/playground-series-s4e1)

- 시작일
 - January 2, 2024
- 마감일 (대회 마감)
 - January 31, 2024
- 평가
 - the ROC curve
- 최종 스코어
 - 1st score: 0.90585
- 목표 스코어
 - Top 20% - Silver
 - 전체 참여 수 3633
 - 상위 20% 726등
 - ROC Curve ≥ 0.89286

선정된 주제 이슈 조사

- ▼ 최근 인공지능(AI)과 금융을 접목한 서비스를 출시하는 금융사들이 늘고 있다.

출처: 동아일보(23.08)

카드사 "AI로 고객이탈 예방"... 은행들은 "자체 생성형AI 개발"

최근 인공지능(AI)과 금융을 접목한 서비스를 출시하는 금융사들이 늘고 있다. 소비자의 편의성을 높이는 장점이 있지만 AI 투명성·공정성 확보와 고객 정보 보호가 과제로 꼽힌다. ...

 <https://www.donga.com/news/Economy/article/all/20230807/20605914/1>

와 버티컬 거대 언어모델(LLM) 개발 중

내 초거대 AI팀 조직

출시 목표로 초거대 AI 기반 'AI 뱅커' 개발 중

생성형 AI 적용을 위한 전담 태스크포스(TF) 설치

가능성 분석하는 '카드이용 활성고객 예측 모형'


AI 마케팅 시스템 '에임즈' 구축

▼ 디지털 전환으로 업종 간 경계가 모호해지고 경쟁이 본격화된 가운데 전통적인 은행의 마케팅 비용은 증가했지만, 고객이 이탈하는 상황에 직면했다.

출처: 뉴데일리경제(23.11)

은행권 '조용한 이탈' 증가... "초개인화 서비스 강화해야"

2024년 은행의 최우선 과제로 예금 확보가 꼽혔다. 지난 2020년만 해도 주요 과제로 부각됐던 대출 증대, 브랜드 인지도 강화 등의 항목보다 예금 확보가 최우선 순위로 부상한 것이다. 28일 하나금융경영연구소에 따르면 2024

 <https://biz.newdaily.co.kr/site/data/html/2023/11/28/2023112800069.html>



▼ 디지털 채널 경쟁의 본격화로 고객 이탈이 우려된다는 전망이 나왔다. 국내 은행들은 디지털 경쟁에서 지면 금융 상품의 단순 제조자로 전락할 수 있으므로 경쟁력을 키워야 한다.

출처: 서울경제(21.01)

<https://www.sedaily.com/NewsView/22H436FENS>

▼ 하나 은행에서 거래 은행 이탈 요인은 모바일 채널 편리성, 금융 소비자 디지털 자산 관리 기대 상승 등이 국내 금융 소비자의 금융거래 특징이라는 분석을 내놨다.

출처: 비즈트리뷴(24.01)

"은행 이탈 요인은 모바일 편리성, 디지털 자산관리 기대감 상승"-하나금융연구

하나은행 산하 하나금융경영연구소가 베이비부머 세대 모바일금융 유입 가속화, 거래은행 이탈 요인은 모바일채널 편리성, 금융소비자디지털 자산관리 기대 상승 등이 국내 금융소비자의 금융거래 특징이라는 분석을 내놨다. 4일 하나금융경영연구소는 '대한민국

 <https://www.biztribune.co.kr/news/articleView.html?idxno=301042>

하나금융경영연구
Hana Institute of Financial Research

▼ MZ세대(1980년대 초~2000대 초 출생) 은행 고객의 이탈이 가속화되면서 은행들이 고령화 문제에 맞닥뜨렸다.

출처: nate뉴스(23.12)

늘어나는 시중은행들...MZ고객 모시기 진담 : 네이트 뉴스

한눈에 보는 오늘 : 경제 - 뉴스 : [서울신문]MZ세대(1980년대 초~2000대 초 출생) 은행 고객의 이탈이 가속화되면서 은행들이 고령화 문제에 맞닥뜨렸다. 미래에도 안정적인 수익을 창출하려면 지금부터 잠재 고객을 최대한 확보

<https://news.nate.com/view/20231212n01431>

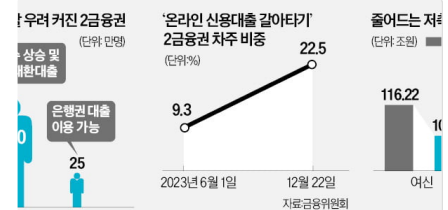


▼ 신용 점수가 올라가면 은행권으로 우량 차주가 대거 이탈할 수 있다.

신용 사면에 고객 이탈...저축은행 '이중고' 시름

신용 사면에 고객 이탈...저축은행 '이중고' 시름, 신용사면에 고객 이탈 약 25만명, 신용점수 회복 받아 1금융권으로 갈아타기 나설 듯 온라인 대환대출도 이탈 부추겨 저축銀 여수신 1년새 8% 감소

<https://www.hankyung.com/article/2024011713351>



출처: 한국경제(24.01)

▼ 고소득 고객은 5대 은행에서 이탈할 가능성이 높은 '고위험 고객'이다.

"돈 되는 고객이 먼저 이탈한다"... 오픈뱅킹 속 위기의 대형은행

대형 은행이 오픈 뱅킹(Open Banking) 때문에 핵심 고객을 잃게 될 위기에 처했다는 분석이 나왔다. 베인앤드컴퍼니(Bain & Company)

<https://www.ciokorea.com/news/37389>

출처: CIO Korea 뉴스레터(18.02)

분석 사례 조사

▼ 신용 카드 소지 여부에 따라 은행 고객 이탈률이 다르게 나타난다.

<http://jiisonline.evehost.co.kr/files/DLA/2-8-2.pdf>

출처: 한국지능정보시스템 학회 논문지(02.12)

▼ Logistic Regression을 이용한 이탈고객예측모형

원문보기 - ScienceON

🔗 <https://scienceon.kisti.re.kr/commons/util/originalView.do?cn=CFKO200811850424975&oCn=NPAP08305944&dbt=CFKO&journal=NPRO00293790>

▼ 금융 CRM전략을 위한 이탈고객 모형분석

dcollection.yonsei.ac.kr

https://dcollection.yonsei.ac.kr/public_resource/pdf/000000306644_20250306210948.pdf

문제 확인

- 고객 이탈에 가장 큰 영향을 주는 변수는 무엇일까?
- 고객 이탈 여부를 가장 잘 예측할 수 있는 모델을 무엇일까?
- 해당 예측으로 생각할 수 있는 인사이트는 어떤 것이 있을까?

데이터

• Train

[train.csv](#)

- 행(row) : 165034
- 열(column) : 14
- 데이터 상위값

| | id | CustomerId | Surname | CreditScore | Geography | Gender | Age | Tenure | Balance | NumOfProducts | HasCrCard | IsActiveMember | EstimatedSalary | Exited |
|---|----|------------|----------------|-------------|-----------|--------|------|--------|-----------|---------------|-----------|----------------|-----------------|--------|
| 0 | 0 | 15674932 | Okwudilichukwu | 668 | France | Male | 33.0 | 3 | 0.00 | 2 | 1.0 | 0.0 | 181449.97 | 0 |
| 1 | 1 | 15749177 | Okwudiliolisa | 627 | France | Male | 33.0 | 1 | 0.00 | 2 | 1.0 | 1.0 | 49503.50 | 0 |
| 2 | 2 | 15694510 | Hsueh | 678 | France | Male | 40.0 | 10 | 0.00 | 2 | 1.0 | 0.0 | 184866.69 | 0 |
| 3 | 3 | 15741417 | Kao | 581 | France | Male | 34.0 | 2 | 148882.54 | 1 | 1.0 | 1.0 | 84560.88 | 0 |
| 4 | 4 | 15766172 | Chiemenam | 716 | Spain | Male | 33.0 | 5 | 0.00 | 2 | 1.0 | 1.0 | 15068.83 | 0 |

• Test

[test.csv](#)

- 행(row) : 110023
- 열(column) : 2
- 데이터 상위값

| | id | CustomerId | Surname | CreditScore | Geography | Gender | Age | Tenure | Balance | NumOfProducts | HasCrCard | IsActiveMember | EstimatedSalary |
|---|--------|------------|-----------|-------------|-----------|--------|------|--------|-----------|---------------|-----------|----------------|-----------------|
| 0 | 165034 | 15773898 | Lucchese | 586 | France | Female | 23.0 | 2 | 0.00 | 2 | 0.0 | 1.0 | 160976.75 |
| 1 | 165035 | 15782418 | Nott | 683 | France | Female | 46.0 | 2 | 0.00 | 1 | 1.0 | 0.0 | 72549.27 |
| 2 | 165036 | 15807120 | K? | 656 | France | Female | 34.0 | 7 | 0.00 | 2 | 1.0 | 0.0 | 138882.09 |
| 3 | 165037 | 15808905 | O'Donnell | 681 | France | Male | 36.0 | 8 | 0.00 | 1 | 1.0 | 0.0 | 113931.57 |
| 4 | 165038 | 15607314 | Higgins | 752 | Germany | Male | 38.0 | 10 | 121263.62 | 1 | 1.0 | 0.0 | 139431.00 |

• sample_submission

sample_submission.csv

- 행(row) : 110023
- 열(column) : 2
- 데이터 상위값

| | id | Exited |
|---|--------|--------|
| 0 | 165034 | 0.5 |
| 1 | 165035 | 0.5 |
| 2 | 165036 | 0.5 |
| 3 | 165037 | 0.5 |
| 4 | 165038 | 0.5 |

변수 설명

- Customer ID: 각 고객의 고유 식별번호
- Surname: 고객의 성
- Credit Score: 고객의 신용점수
- Geography: 고객이 거주하는 국가
- Gender: 고객의 성별
- Age: 고객의 나이
- Tenure: 고객이 은행을 이용한 기간

- Balance: 고객의 계좌 잔액
- NumOfProducts: 고객이 이용하는 은행 상품의 수(ex. 예금, 적금)
- HasCrCard: 신용카드 보유 여부
- IsActiveMember: 활성 회원 여부
- EstimatedSalary: 고객의 예상 연봉
- Exited: 고객 이탈 여부

문제 해결 전략

- 데이터 전처리
 - 상관관계 분석을 통한 변수 선택
 - 불필요한 피처 제거
 - Id, CustomerId, Surname 제거
 - 변수 변환
 - Int 변환
 - HasCrCard, Age, IsActiveMember
 - 인코딩
 - Geography, Gender
 - OneHot Encoder
 - Label Encoder
 - ...
 - 스케일링
 - Standard Scaler
 - Robust Scaler
 - MinMax Scaler
 - ...
 -
- 모델링
 - Logistic Classifier : 기본적인 선형 분류 모델로 해석력 뛰어나다.
 - LightGBM Classifier : 대용량 데이터에서 빠른 학습 속도와 높은 성능 보여준다.

- CatBoost Classifier : 범주형 변수를 효과적으로 처리하는 부스팅 모델이다.
- GradientBoosting Classifier : 앙상블 기법을 활용해 뛰어난 성능을 제공한다.
- ...
- 하이퍼 파라미터 최적화
 - 베이지언 최적화
 - 랜덤서치

일정표

타임라인 (2)

| Aa Name | 📅 Date | ⋮ Tags |
|-------------------------|------------------------------|--------|
| ✓ <u>주제 선정</u> | @2024년 4월 11일 | |
| 📢 <u>발표</u> | @2024년 4월 26일 | |
| 📅 <u>기획서 제출</u> | @2024년 4월 16일 | |
| 📅 <u>결과물 제출</u> | @2024년 4월 26일 | |
| ✓ <u>발표자료 최종 점검</u> | @2024년 4월 25일 | |
| 📅 <u>분류/회귀 분석 동시 진행</u> | @2024년 4월 11일 → 2024년 4월 12일 | |
| 👤 <u>멘토링</u> | @2024년 4월 20일 오후 7:00 | |
| 📅 <u>추가 대회 진행</u> | @2024년 4월 22일 → 2024년 4월 24일 | |
| 👤 <u>멘토링</u> | @2024년 4월 22일 오후 8:00 | |
| 📅 <u>보고서 작성(피드백 반영)</u> | @2024년 4월 22일 | |
| 📅 <u>분류/회귀 분석 동시 진행</u> | @2024년 4월 15일 → 2024년 4월 19일 | |
| 📅 <u>보고서 작성</u> | @2024년 4월 19일 | |

팀원 소개 및 역할

- 김수지 [PM]
 - 기획서 작성
 - 보고서 작성
 - 발표 자료 작성
 - 전복 데이터 세트를 사용한 회귀 분석

- 데이터 전처리
 - 분석 및 EDA
- 박지은 [PL]
 - 기획서 작성
 - 보고서 작성
 - 발표
 - 전복 데이터 세트를 사용한 회귀 분석
 - 모델링
 - 모델 평가
- 변진영 [PM]
 - 기획서 작성
 - 보고서 작성
 - 발표 자료 작성
 - 은행 이탈 데이터를 사용한 이진 분류 분석
 - 모델링
 - 모델 평가
 - 전복 데이터 세트를 사용한 회귀 분석
 - 모델 평가
- 이소희 [PM]
 - 기획서 작성
 - 보고서 작성
 - 발표 자료 작성
 - 은행 이탈 데이터를 사용한 이진 분류 분석
 - 분석 및 EDA
- 이정수 [PL]
 - 기획서 작성
 - 보고서 작성
 - 발표 자료 작성
 - 발표
 - 은행 이탈 데이터를 사용한 이진 분류 분석
 - 데이터 전처리

- 모델 평가