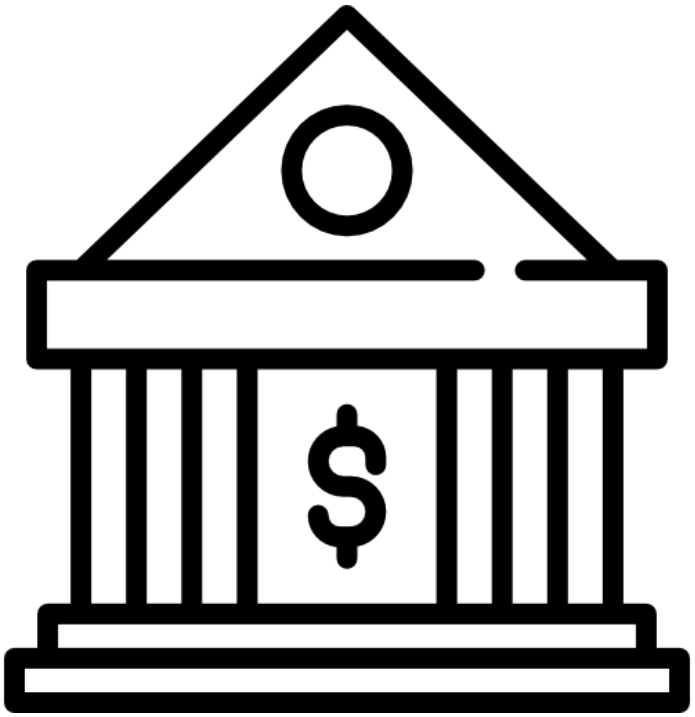
은행 이탈 고객 예측 분류 분석

프로젝트 타입	팀 프로젝트
Tool	Python



목차

Binary Classification with a Bank Churn Dataset

: 은행 이탈 고객 예측

Binary Classification with a Bank Churn Dataset
Playground Series - Season 4, Episode 1

https://www.kaggle.com/competitions/playground-series-s4e1

데이터 설명

1. **데이터 크기**

	행(row)	열(column)
Train	165034	14
Test	110023	13
Submission	110023	2

2. **변수 설명**

Feature	mean	Type	Train NA	Test NA	Origin NA	Submission NA
Id	순번	int64	0	0	0	0
CustomerId	고유 식별번호	int64	0	0	0	-
Surname	성	object	0	0	0	-
CreditScore	신용점수	int64	0	0	0	-
Geography	거주 국가	object	0	0	0	-
Gender	성별	object	0	0	0	-
Age	연령	float64	0	0	0	-
Tenure	은행 이용 기간	int64	0	0	0	-
Balance	계좌 잔액	float64	0	0	0	-
NumOfProducts	은행 이용 상품 수	int64	0	0	0	-
HasCrCard	신용카드 보유 여부	float64	0	0	0	-
IsActiveMember	활성 회원 여부	float64	0	0	0	-
EstimatedSalary	예상 연봉	float64	0	0	0	-

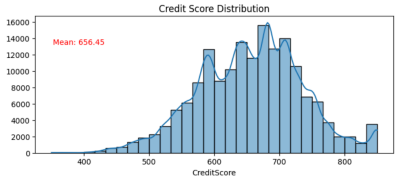
Target	mean	Type	Train NA	Test NA	Origin NA	Submission NA
Exited	이탈 여부	int64	0	-	0	0

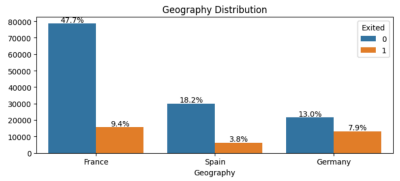
원본 데이터 EDA 시각화

원본 EDA 시각화

: train, test 분포 동일

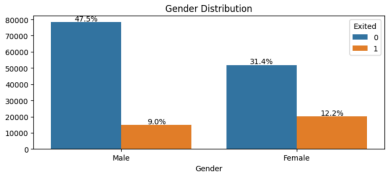
[Train : Histogram,BarPlot]



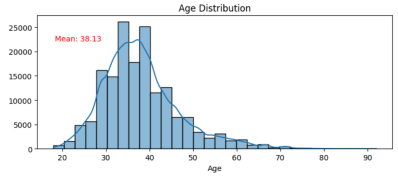


- **CreditScore(신용점수)**
  - 평균
    - 656.45
  - 분포
    - 대체로 정규분포 형태
    - 신용 점수가 600-700 사이에 가장 많은 고객이 분포
    - 점수 분포가 넓어 다양한 고객층이 존재

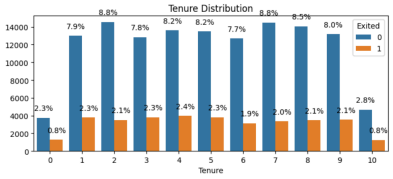
- **Geography(거주 국가)**
  - 분포
    - France가 전체의 57.1%로 가장 많고, Spain 22%, Germany 20.9%로 구성
  - 이탈
    - 세 국가 모두 이탈하지 않은 고객 비율이 높음
      - France는 이탈율이 낮은 편, Germany는 상대적으로 높은 이탈율



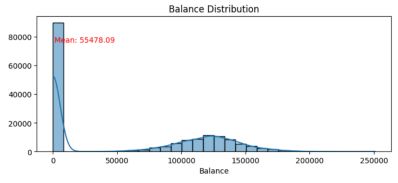
- **Gender(성별)**
  - 분포
    - 남성이 56.5%, 여성이 43.6%로 약 6:4 비율
  - 이탈
    - 남성과 여성 모두 이탈하지 않은 비율이 높음
    - 여성이 남성보다 이탈율이 높음



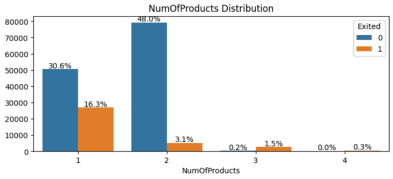
- **Age(연령)**
  - 평균
    - 38.13
  - 분포
    - 대체로 정규분포 형태
    - 30,40대 고객이 가장 많음
    - 연령이 높아질수록 고객 수가 급격히 감소



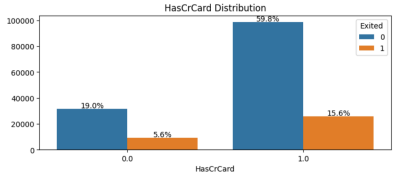
- **Tenure(은행 이용 기간)**
  - 분포
    - 모든 기간에 고르게 분포
    - 0년과 10년에서 고객 수가 소폭 감소
  - 이탈
    - 이용 기간과 관계없이 이탈하지 않은 비율이 높음



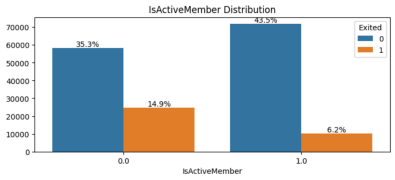
- **Balance(계좌 잔액)**
  - 평균
    - 55478.09
  - 분포
    - 0인 경우가 가장 많으며, 0을 제외한 나머지 값들은 정규분포에 가깝게 분포
    - 50,000 ~ 200,000 구간에 고객이 집중



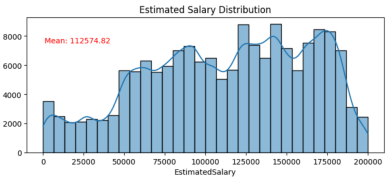
- **NumOfProducts(은행 이용 상품 수)**
  - 분포
    - 1개 또는 2개 상품을 보유한 고객이 대부분
  - 이탈
    - 상품 개수가 1개인 고객의 이탈율이 상대적으로 높음



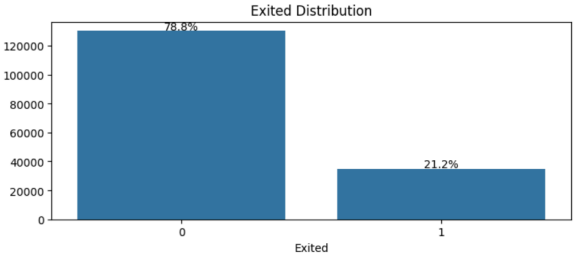
- **HasCrCard(신용카드 보유 여부)**
  - 분포
    - 신용카드 보유 고객이 75.4%로, 비보유 고객 24.6%보다 많음
  - 이탈
    - 신용카드 보유 여부와 관계없이 이탈하지 않은 고객이 많음



- **IsActiveMember(활성 회원 여부) -**
  - 분포
    - 비활성회원이 50.2%, 활성회원이 49.7%로 근소한 차이
  - 이탈
    - 비활성 회원의 이탈율이 더 높음

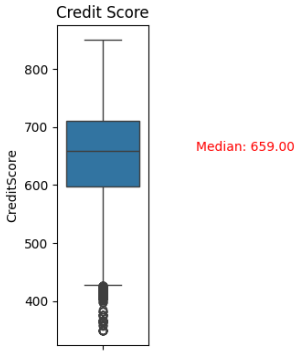


- **EstimatedSalary(예상 연봉)**
  - 평균
    - 112574.82
  - 분포
    - 0에서 200,000까지 고르게 분포
    - 주로 50,000 ~ 175,000 구간에 집중

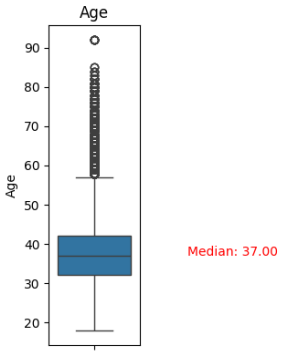


- **Exited(이탈 여부)_Target**
  - 분포
    - 이탈하지 않은 고객이 78.8%로, 이탈 고객 21.2%보다 많음

[단일 변수 BoxPlot으로 이상치 확인]

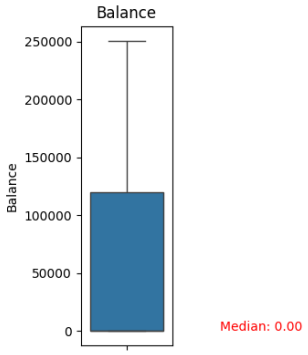


- **CreditScore(신용 점수)**
  - 중앙값
    - 659
  - 이상치
    - 존재



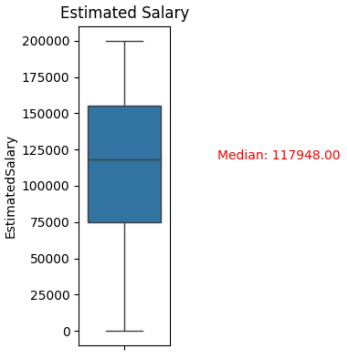
- **Age(연령)**
  - 중앙값
    - 37
  - 이상치
    - 존재

        - 특히 400 이하의 낮은 점수 고객 포함, 일부 고객층에서 낮은 점수 확인
    - 분포
        - 약간 오른쪽으로 치우친 분포로, 왼쪽 꼬리가 길게 나타남



- **Balance(계좌 잔액) -**
    - 중앙값
        - 0
    - 이상치
        - 없음
    - 분포
        - 대부분의 고객이 0의 잔액을 보유하며, 나머지 값들은 균일하지 않은 분포를 보임

        - 60세 이상의 연령에서 이상치 다수 확인, 특히 80세 이상 고객이 드물게 포함되어 고령층 소수 포함
    - 분포
        - 오른쪽으로 긴 꼬리를 가진 분포



- **EstimatedSalary(예상 연봉)**
    - 중앙값
        - 117,948
    - 이상치
        - 없음
    - 분포
        - 0에서 200,000까지 고르게 분포하며, 특정 구간에 고객이 집중되어 있지 않음

전처리 계획 요약

1. **원본 진행**

    원본 진행

- Id, Surname
    - 의미없는 변수 제거
- 이상치 처리 X
- 중복치 처리 X
- 파생변수 X
- 인코딩 처리 O

2. **변수 전처리 진행**

    변수 전처리 진행

- Id, Surname
    - 의미없는 변수 제거
- 이상치 처리 O

- 중복치 처리 O
- 인코딩 처리 O
- 파생변수 X

3. **변수 전처리 + 파생변수 진행(의미 중심)**

   변수 전처리 + 파생변수(의미 중심)

- Id, Surname
  - 의미없는 변수 제거
- 이상치 처리 O
- 중복치 처리 O
- 인코딩 처리 O
- 파생변수 O

4. **변수 전처리 + 파생변수 진행(성능중심)**

   변수 전처리 + 파생변수(성능 중심)

- 이상치 처리 O
- 중복치 처리 O
- 인코딩 처리 O
- TF-IDF + SVD + 파생변수 O

최종 모델 선정

지표	원본	변수 처리	변수 처리 + 파생변수 (의미)	변수 처리 + 파생변수(성능) ✔
Colab	0.8902	0.8911	0.8907	**0.8936**
Kaggle Private	0.88890	0.88891	0.88871	**0.89251**
Kaggle Public	0.88641	0.88635	0.88589	**0.88890**

인사이트 도출

: 도메인 지식 기반의 중요도 파악 후 우선순위를 통한 인사이트

: 변수 전처리 + 파생변수 진행(의미 중심)을 통한 인사이트

1. **인사이트 도출 변수**
   - **Age, Age_group**
   - **Balance**
   - **CreditScore**
   - **EstimatedSalary**

- **HasCrCard**
- **IsActiveMember**
- **NumOfProducts_1,2,3,4**
- **Tenure**
- **Balance_Tenure**
- **CreditScore_Age**
- **CreditScore_Tenure**
- **EstimatedSalary_Age**
- **EstimatedSalary_NumOfProducts**
- **NumOfProducts_Age**
- **NumOfProducts_Balance**
- **NumOfProducts_Tenure**
- **Tenure_Age**

2. **우선순위 결정 후 인사이트 도출**
   - **1순위: 상관계수와 Feature Importance, Permutation Importance 모두 높은 변수**

     📎
     - **이탈 방지**
       1. **고령층을 위한 맞춤형 상품 개설**
          a. 연령대 별 상품 선호도 분석
          b. 정기예금 상품 출시 및 금리 혜택 제공
       2. **고령층을 위한 디지털 은행 이용 교육 or 오프라인 이용 접근성 확대**
          a. 디지털 접근성 강화
       3. **신용점수 회복 방안 구축, 신용 점수 관리 프로그램 개설**
          a. 개인 맞춤형 신용 점수 관리 프로그램
     - **고객 유지**
       1. **젊은층을 위한 맞춤 전략**
          a. 모바일 앱을 통한 서비스
       2. **신용점수가 높은 VIP 고객들을 위한 맞춤 컨설팅 서비스**
          a. 투자 설계 혹은 자산 관리 매니저 제공

     - **Age : 고령층일수록 이탈 가능성 상승**
       - 상관계수: 0.34 (높음)
       - Feature Importance: 398.74 (상위권)
       - Permutation Importance: 0.0176 (상위권)
     - **CreditScore_Age : 고령층이면서 신용점수가 낮은 고객 이탈 가능성 높음**
       - 상관계수: -0.29 (상위권)
       - Feature Importance: 539.36 (최상위권)
       - Permutation Importance: 0.0011 (중간)

   - **2순위: Feature Importance와 Permutation Importance 높은 변수**

📎

- **이탈 방지**
  1. **초기 고령층 고객 대상 프로모션**
     a. 고령층을 위한 초기 가입 리워드 혹은 서비스 길잡이 제공
  2. **저수익 고객을 위한 안정적인 상품 개발**
     a. 부담이 적은 저위험/고안정 상품
  3. **단일 상품 이용 고객들 추가 상품 구매를 위한 전략 구비**
     a. 교차 판매를 통한 추가 패키지 상품 추천
     b. 기존 상품과 연계하는 부담이 적은 상품 추천
- **고객 유지**
  1. **활성 고객 대상 추가 행사 진행 및 지속적인 커뮤니케이션 강화**
     a. 정기적인 고객 만족도 조사 및 피드백 반영

- **CreditScore : 신용점수가 낮을수록 이탈 가능성 높음**
  - 상관계수: -0.03 (낮음)
  - Feature Importance: 455 (최상위권)
  - Permutation Importance: 0.0002 (중간)
- **Tenure_Age : 초기 고령층 고객 이탈 가능성 높음**
  - 상관계수: -0.13 (중간)
  - Feature Importance: 425.00 (최상위권)
  - Permutation Importance: 0.0001 (낮음)
- **EstimatedSalary_Age : 예상연봉이 낮은 고령층 고객 이탈 가능성 높음**
  - 상관계수: -0.13 (중간)
  - Feature Importance: 437.00 (최상위권)
  - Permutation Importance: 0.0000 (낮음)
- **NumOfProducts_Age : 이용 상품 수가 적은 고령층 고객 이탈 가능성 높음**
  - 상관계수: -0.32 (높음)
  - Feature Importance: 353.73 (상위권)
  - Permutation Importance: 0.0031 (상위권)
- **EstimatedSalary_NumOfProducts : 이용 상품수가 적으면서 예상 연봉이 높을수록 이탈 가능성 높음**
  - 상관계수: 0.18 (중간)
  - Feature Importance: 406.01 (최상위권)
  - Permutation Importance: 0.0002 (중간)

- **3순위: 상관계수와 Feature Importance가 중간 정도인 변수**

📎

- **이탈 방지**
  1. **잔고가 많은데 이용 상품수가 적은 고객**
     a. 상품을 충분히 경험하지 못한 고객
     b. 고객 데이터를 기반으로 다양한 상품 추천
        i. 저위험 단기 투자 상품과 같은 짧은 체험 프로그램 제공
  2. **고자본 고객을 만족시킬 수 있는 상품 추천**
     a. 다양한 상품(위험선호도에 따른 다양한 상품을 접할 기회 제공)
  3. **초기 고자본 고객들을 위한 지속적인 커뮤니케이션 및 서비스 혜택 강화**
     a. 개인 컨설팅 서비스
     b. 프리미엄 전용 금융 세미나와 같은 고객 관심도를 이끌어 낼 수 있는 은행 행사 유치
- **고객 유지**
  1. **저자본 고객들을 위한 부담이 적은 안정적인 상품 추천**
     a. 투자보단 적금과 같은 안정적인 금융 상품으로 고객 신뢰도 강화
     b. 기초 금융 교육 프로그램 제공
        i. 투자 프로그램
  2. **장기 고객 유지를 위한 충분한 혜택 및 지속적인 소통**
     a. 장기 고객 전용 상품, 혜택 및 우대 금리 제공

- **NumOfProducts_Balance : 잔고가 많으면서 이용 상품수가 적은 고객 이탈 가능성 높음**
  - 상관계수: -0.21 (중간)
  - Feature Importance: 243.39 (중간)
  - Permutation Importance: 0.0024 (중간)
- **Balance : 잔고가 많을수록 이탈 가능성 높음**
  - 상관계수: 0.13 (중간)
  - Feature Importance: 183.01 (중간)
  - Permutation Importance: 0.0005 (낮음)
- **Balance_Tenure : 잔고가 많으면 초기 고객 이탈 가능성 높음**
  - 상관계수: 0.03 (낮음)
  - Feature Importance: 261.35 (중간)
  - Permutation Importance: 0.0005 (낮음)

- **4순위: 상관계수는 낮으나 Feature Importance 또는 Permutation Importance가 높은 변수**

📎

- **이탈 방지**
  1. **비활성 고객의 니즈를 파악하고 활성화를 위한 조사 실시**
     a. 비활성 고객의 주요 불만사항 분석하여 해결책 제안
  2. **고객에게 적합한 수의 상품 추천**
     a. 고객과 적극적인 커뮤니케이션
     b. 단일 상품 이용 고객에게 추가 상품 제안
     c. 3개 이상의 상품 이용 고객에게 최적화를 제안함으로써 충성도 강화
- **고객 유지**
  1. **활성화 멤버들의 만족도 조사 실시 및 혜택 제공**
     a. 정기적으로 만족도 평가 실시하고 피드백 반영 후 조정
  2. **적당한 상품 수를 이용하며 유지하고 있는 고객을 적극적으로 관리**
     a. 정기적 소통을 통해 고객 니즈 변화에 민감하게 대응

- **IsActiveMember : 비활성 멤버일수록 이탈 가능성 높음**
  - 상관계수: -0.21 (중간)
  - Feature Importance: 121.36 (중간)
  - Permutation Importance: 0.0123 (상위권)
- **EstimatedSalary : 예상 잔고가 높을수록 이탈 가능성 높음**
  - 상관계수: 0.02 (낮음)
  - Feature Importance: 361.67 (상위권)
  - Permutation Importance: 0.0001 (낮음)
- **NumOfProducts : 이용 상품이 2개일 때 이탈 가능성 낮아진다 / 이용 상품이 1,3,4개 일때 이탈 가능성 높음**
  - 상관계수: 0.31, -0.38, 0.22, 0.09 (중간~높음)
  - Feature Importance:
    - NumOfProducts_2: 23.08 (중간)
    - NumOfProducts_3: 7.02 (낮음)
    - NumOfProducts_1: 0.70 (낮음)
    - NumOfProducts_4: 0.00 (낮음)
  - Permutation Importance:
    - NumOfProducts_2: 0.0456 (최상위권)
    - NumOfProducts_3: 0.0004 (낮음)
    - NumOfProducts_1: 0.0006 (낮음)
    - NumOfProducts_4: 0.0000 (낮음)

- **5순위: 상관계수, Feature Importance, Permutation Importance 모두 낮은 변수**

📎

- **이탈 방지**
  1. **신용 카드 가입 혜택의 적극적인 홍보를 통해 사용 유도**
     a. 현금 리워드, 할인, 포인트 적립 등 다양한 혜택 제공
     b. 광고를 통한 정기적 혜택 알림 발송
- **고객 유지**
  1. **신용 카드 사용 장점을 부각하여 기존 고객들의 적극적 사용 유도**
     a. 고객 소비 패턴 데이터 분석을 통해 맞춤형 혜택 혹은 신규 신용카드 추천

- **Tenure : 초기 고객 이탈 가능성 높음**
  - 상관계수: -0.02 (낮음)
  - Feature Importance: 37.33 (낮음)
  - Permutation Importance: 0.0000 (낮음)
- **HasCrCard : 신용카드 보유하지 않을수록 이탈 가능성 높음**
  - 상관계수: -0.02 (낮음)
  - Feature Importance: 57.00 (낮음)
  - Permutation Importance: 0.0005 (낮음)
- **NumOfProducts_Tenure : X(상관계수가 0)**
  - 상관계수: 0.00 (낮음)
  - Feature Importance: 87.00 (낮음)
  - Permutation Importance: 0.0000 (낮음)