PROJECT

데이터가 말합니다
대출, 가능합니다

Home Credit 데이터를 활용한 신규·기존 고객군 신용 리스크 분석

7조
김성철, 김수연, 이하은, 조민지

CONTENT

1 프로젝트 개요

1 프로젝트 개요

프로젝트 배경

Home Credit: 1997년 체코 설립, 글로벌 비은행 금융기관
전통 금융권에서 소외된 저신용·신용이력 부족층 대상 무담보·소액 대출 제공

프로젝트 주제

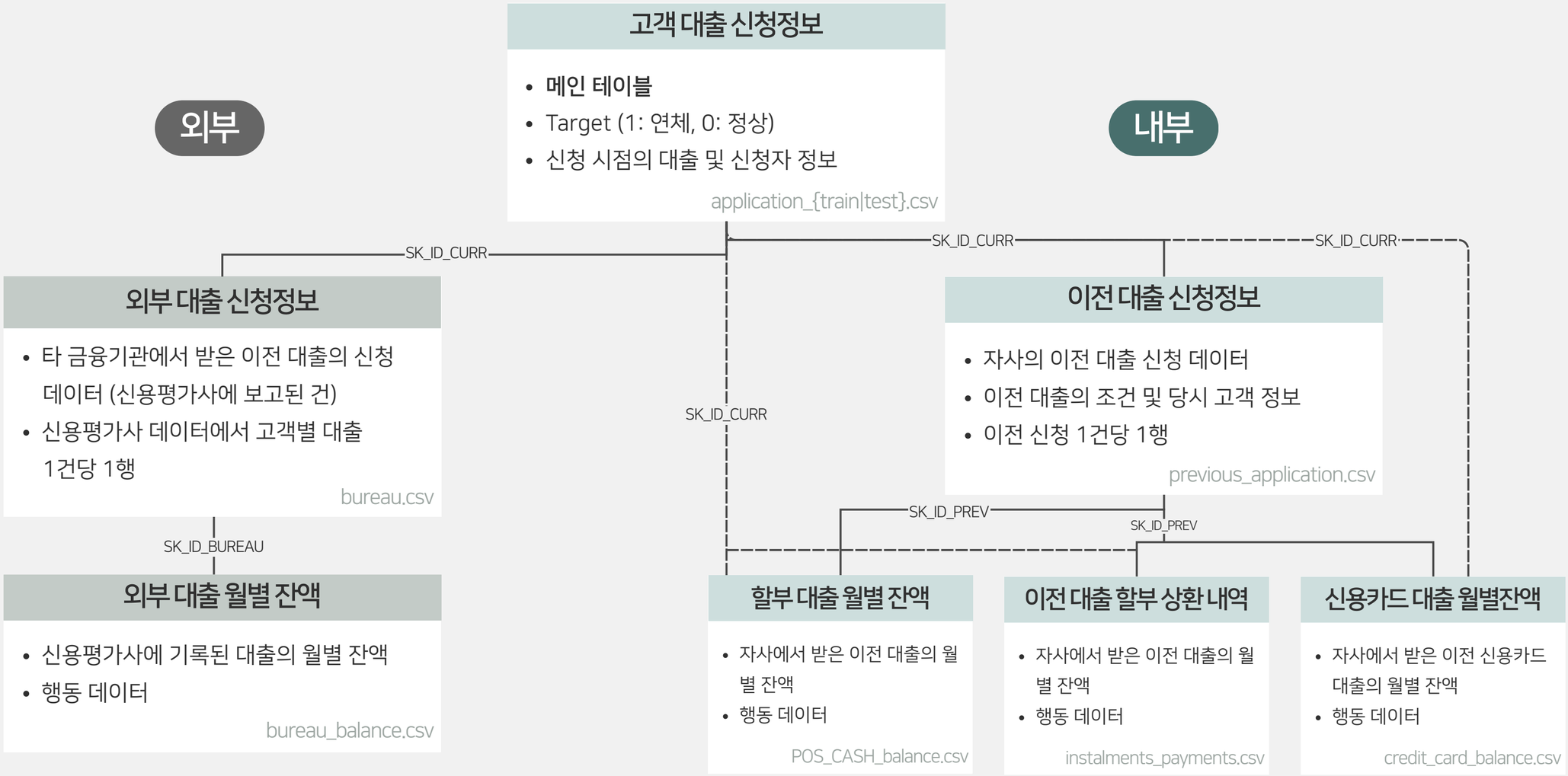
대출 이력 부족·소액 대출 중심 고객군에 적합한 자사 전용 신용점수 시스템 구축

프로젝트 목표

신규·기존 고객군 특성을 반영한 맞춤형 리스크 관리 방안 마련

2 데이터 소개

2 데이터 소개 DATA MAP



2 데이터 소개

Home Credit 데이터를 기반으로 한 타겟·고객군 기준 및 출처 요약

데이터 출처

Kaggle 공개 데이터셋
"Home Credit Default Risk"

TRAGET 데이터의 의미

대출 신청 고객의 연체 여부
(1: 연체, 0: 정상)

신규 고객군 분류 기준

과거 자사 대출 이력이 **없거나**
한 번만 이용했던 고객

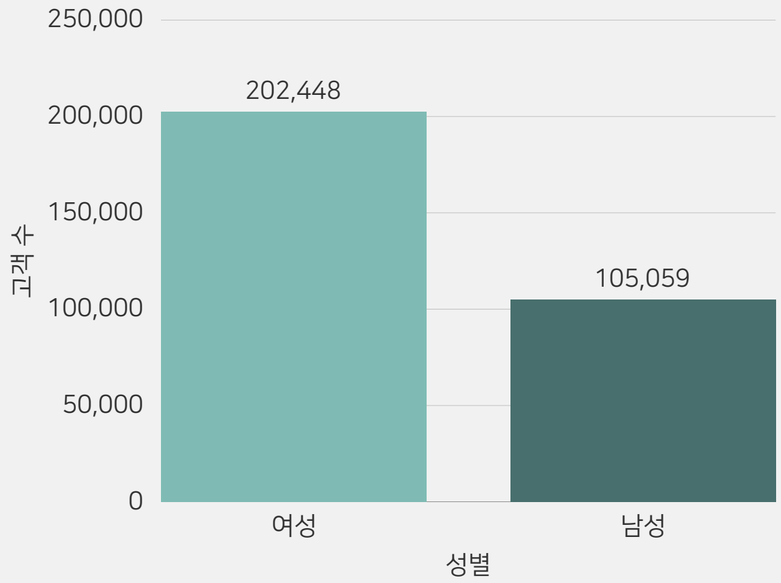
기존 고객군 분류 기준

과거 자사 대출 이력이 **2회 이상**인 고객

3 EDA

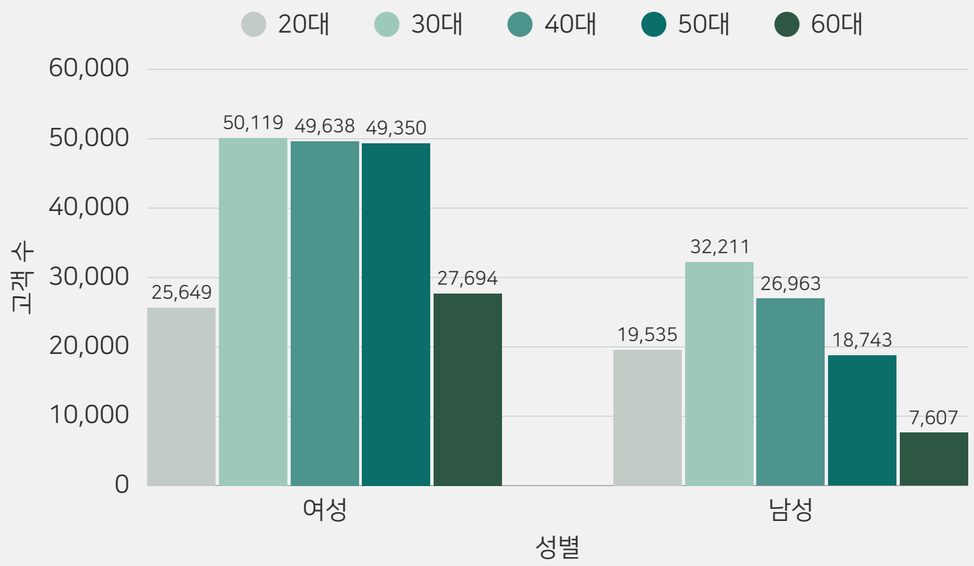
3 EDA · 고객 기본 특성

성별에 따른 대출 신청자 분포



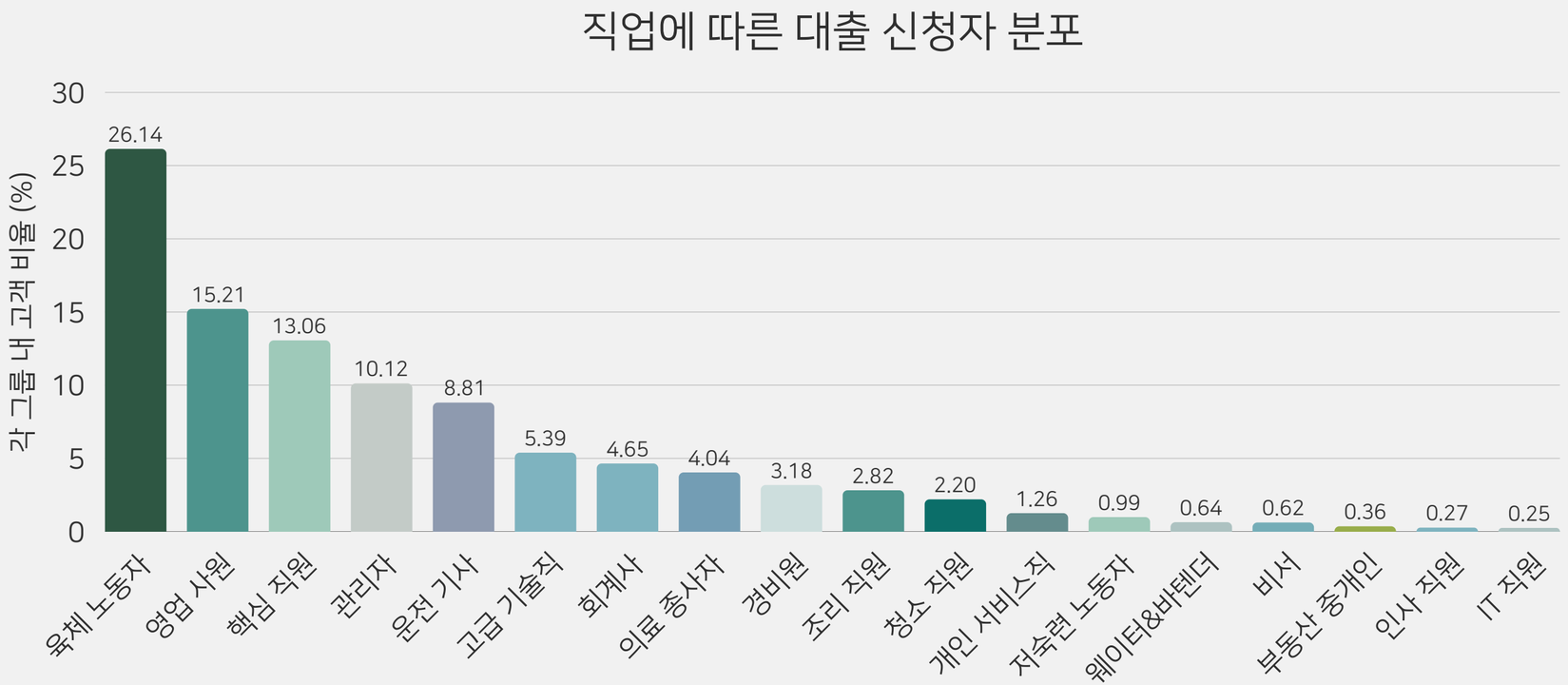
대출 신청 고객의 성비는
여성이 남성보다 약 2배가량 많음

성별 및 연령대에 따른 대출 신청자 분포



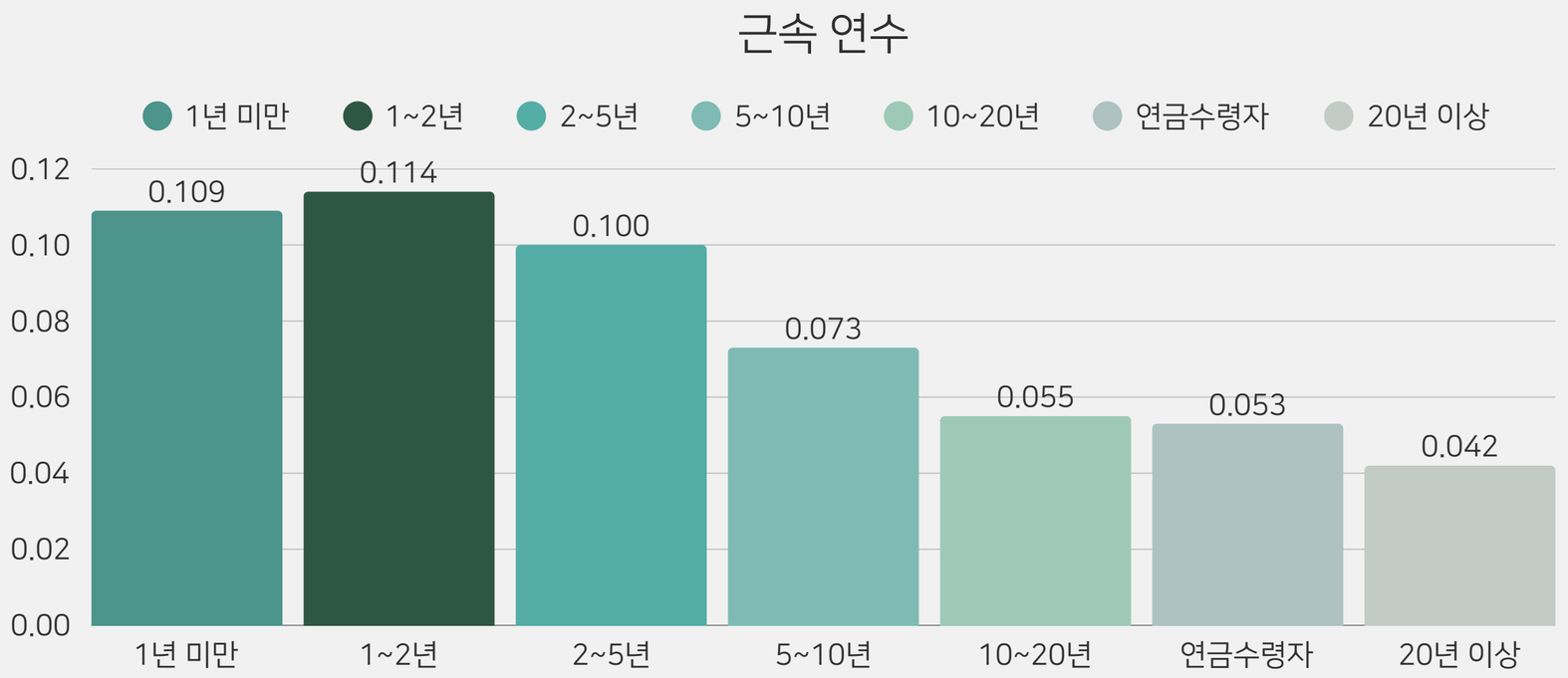
▲ 대출 신청의 **핵심 연령층** : **30대 ~ 50대**

▼ 대출 신청이 **낮은 연령층** : 20대는 소득·신용 한계로,
60대는 대출 규제로 수요가 상대적으로 적을 수 있음



대출 신청자는 **육체노동자·영업사원·핵심 직원** 순으로 비중이 높게 나타남

3 EDA · 고객 기본 특성

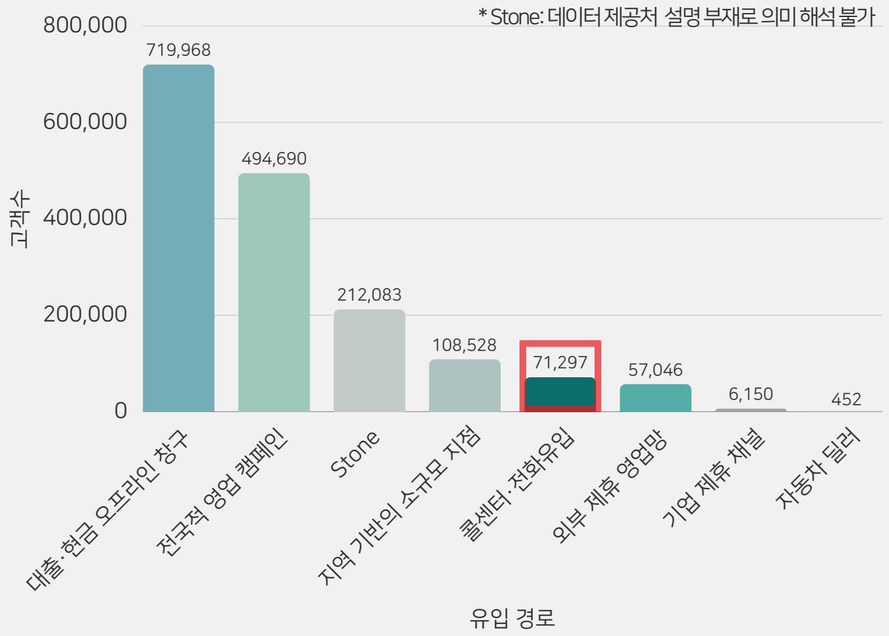


▲ 근속연수가 **2년 이하**로 짧은 구간에서 **연체율이 높음**

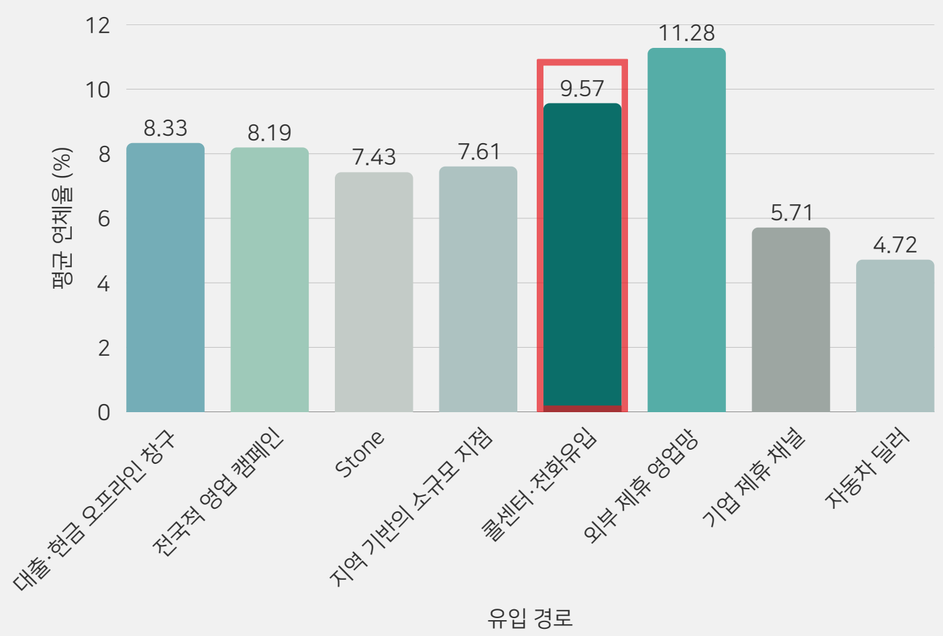
▼ 근속연수가 **길수록 연체율은 낮아짐**

3 EDA · 과거 대출 신청

대출 신청 유입 경로 분포



대출 신청 유입 경로별 연체율



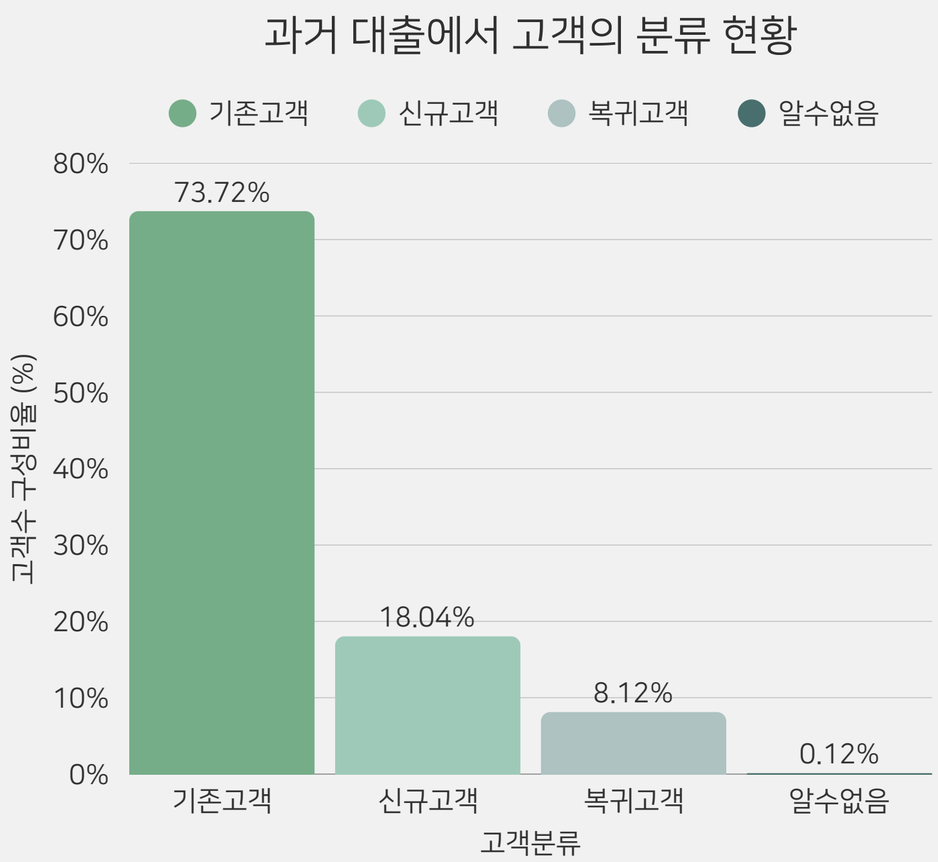
콜센터 · 전화유입 건수는 상대적으로 적으나 **연체율은 9.57%로 높은 편**

비대면 채널 특성상 고객 신뢰성·정보 정확도가 낮을 가능성

3 EDA · 과거 대출 신청

과거 대출 고객 중 **기존 고객이 73.72%**로 대부분을 차지

신규 고객은 18.04%, 복귀 고객은 8.12%로
상대적으로 적은 비중을 보임



4 KPI

4 KPI · 고객 분석을 위한 핵심 지표 구조

1 KPI 설계 배경

분석 결과와 신용평가 지표 FICO·NICE를 기반으로 **신용점수에 활용할 핵심 변수 도출**

리스크 관리에 효과적인 핵심 지표 선별

* KPI: Key Performance Indicator, 성과를 측정하는 핵심 지표
* NICE : 한국의 대표적인 신용평가사
* FICO : 미국의 대표 신용평가 모델

2 금융 요소 기반 평가 요소 및 세부 지표 구성

평가 요소	평가 요소의 상세내용
상환이력	장·단기 연체 이력, 연체 진행 일수, 연체 해제 여부, 조기 상환 여부
부채수준	고위험 대출 여부, 장기 대출 여부, 대출 잔액 변화, 대출 상환 여부
신용거래기간	신용카드·외부 대출·내부 대출 등 신용 거래 기간 경과
신용형태	신용·체크카드 사용 기간 및 금액, 할부·리볼빙·현금서비스 이용 여부
외부 신용 평가	외부 신용평가 기관에 등록된 신용 점수

4 KPI · 도입

KPI 파생변수생성

01
장·단기
연체 여부

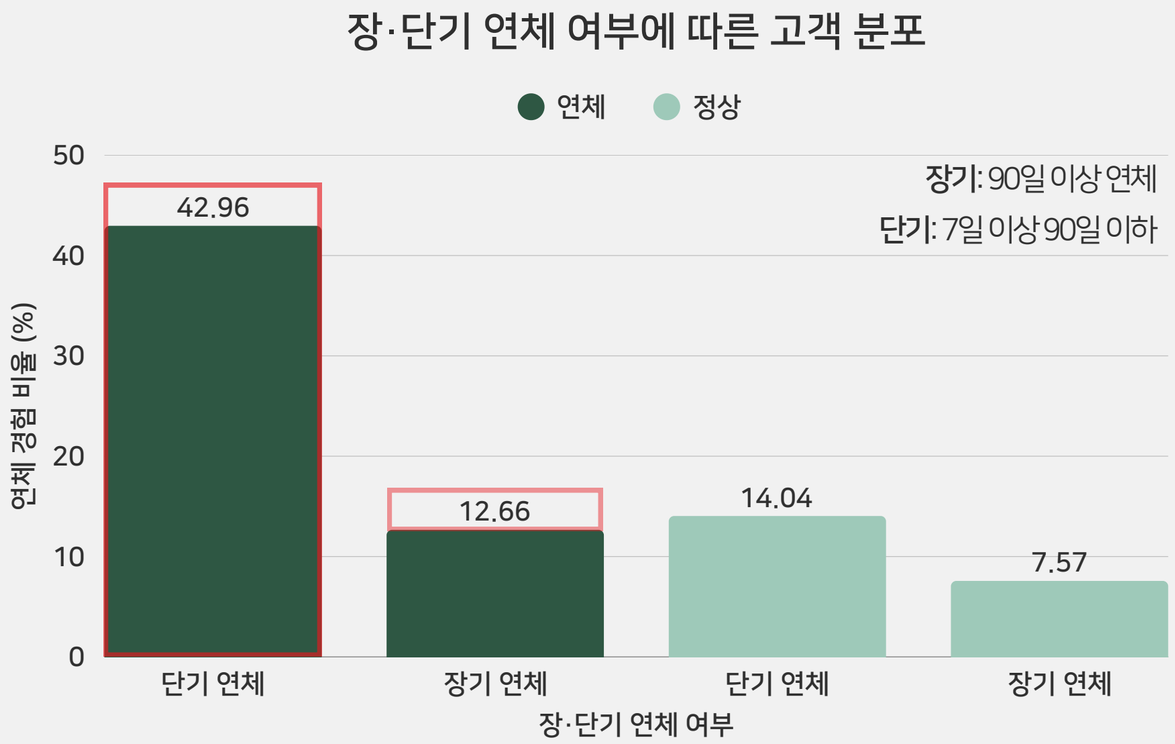
02
대출 전액
상환 비중

03
신용카드
사용 기간

04
신용카드
사용 적정성

05
대출 상품
유형별 이용률

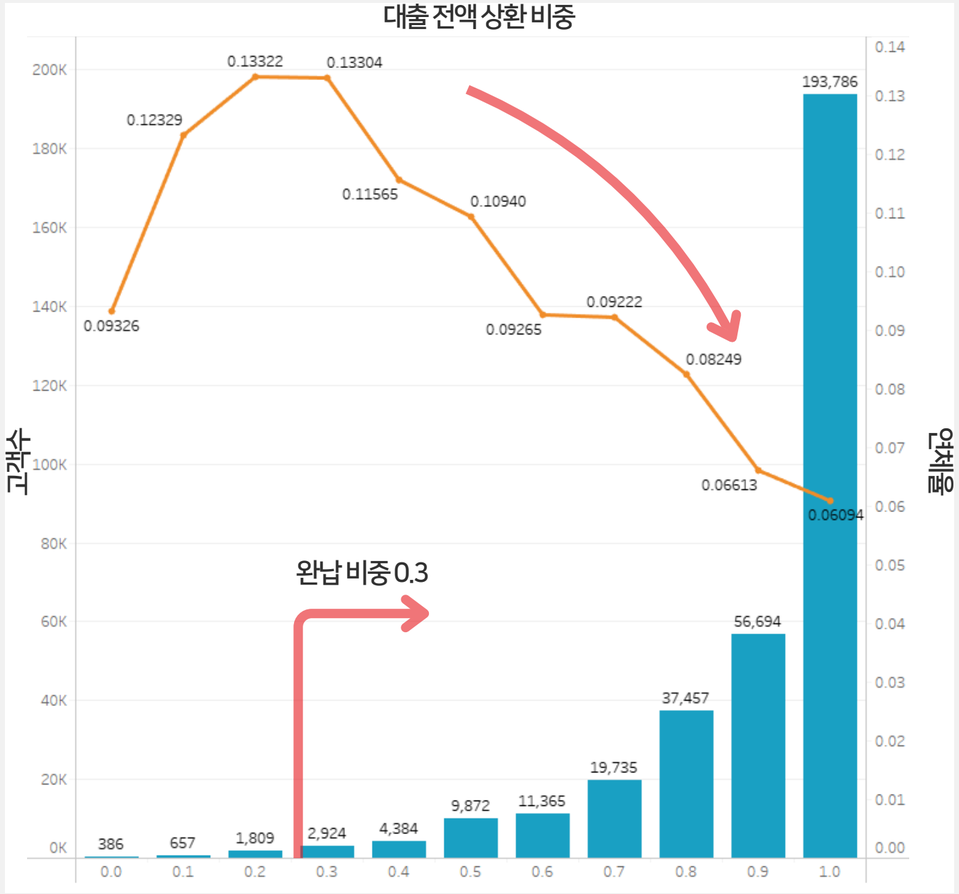
4 KPI 변수 생성 · 장기·단기 연체 여부에 따른 고객 분포



연체 고객은 정상 고객보다 **장·단기 연체 경험 비율이 모두 높으며**

특히 **단기 연체 경험**에서 차이가 두드러짐

4 KPI 변수 생성 · 대출 전액 상환

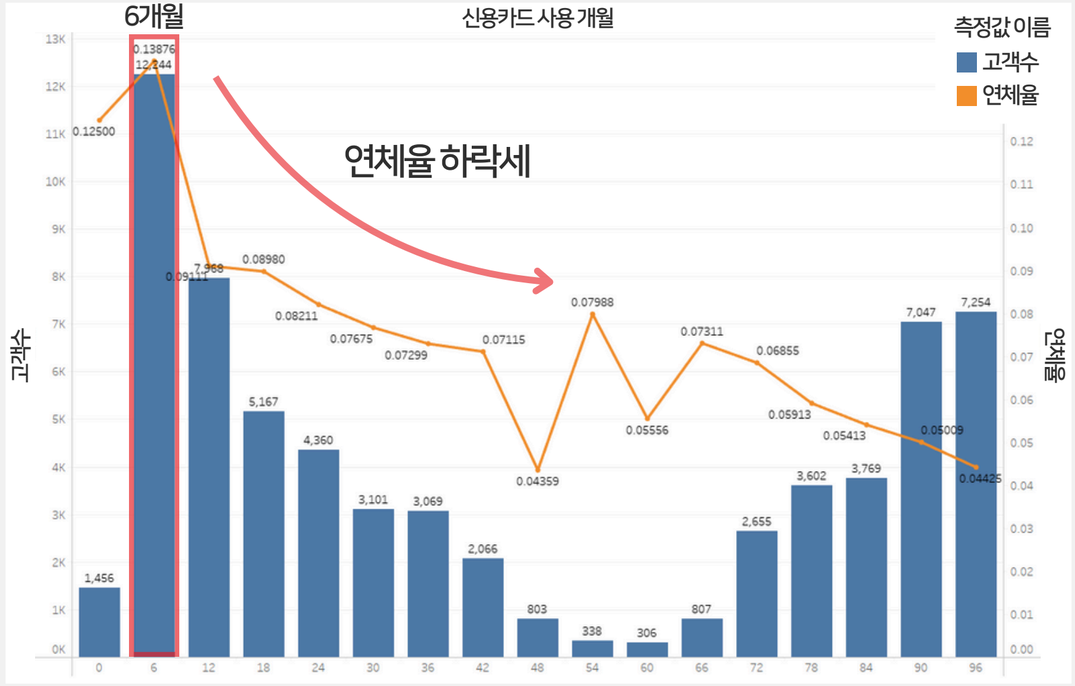


측정값 이름
- 고객수
- 연체율

완납 비중 0.3이후 부터 증가할수록 **연체율은 하락하는 추세**

➡ 완납 비중이 높은 고객은 신뢰성이 높음

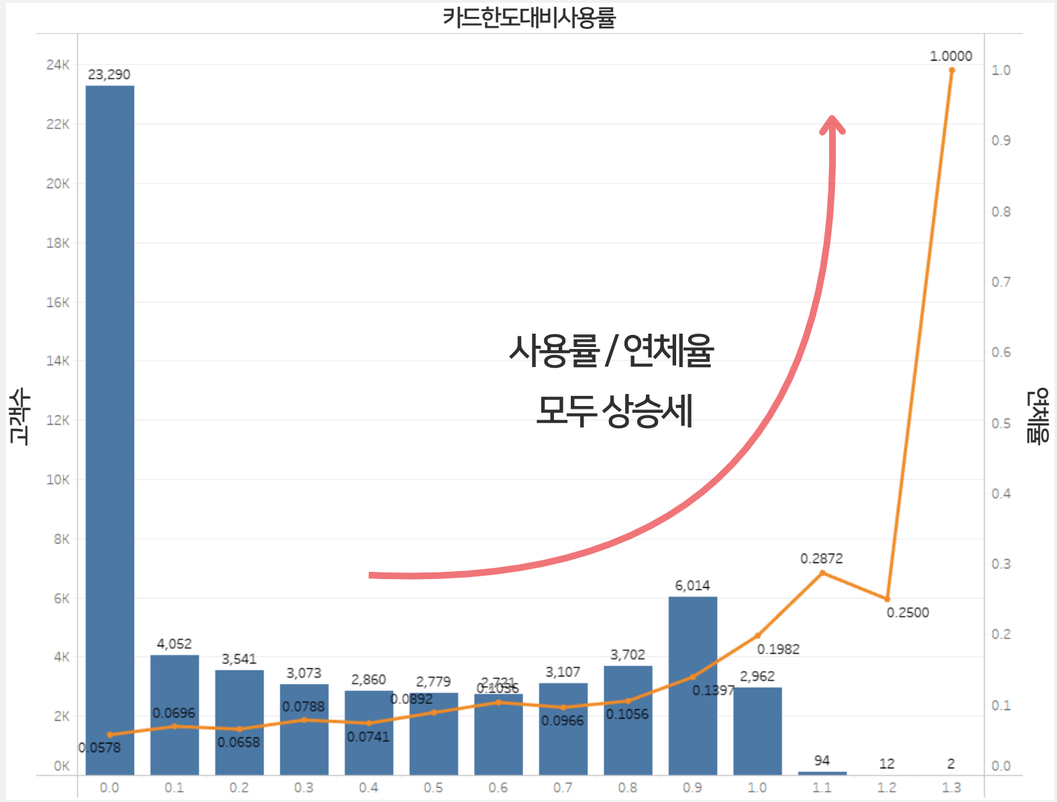
4 KPI 변수 생성 · 신용카드 사용 개월



신용카드 이용개월수가 **6개월을 넘으면**
연체율이 급격하게 감소

➡ 신용카드 이용이력 짧은 고객은 신뢰도가 낮음

4 KPI 변수 생성 · 신용카드 사용 적정성



신용카드 **한도대비 사용률**이
증가할수록 **연체율** 또한 **상승**

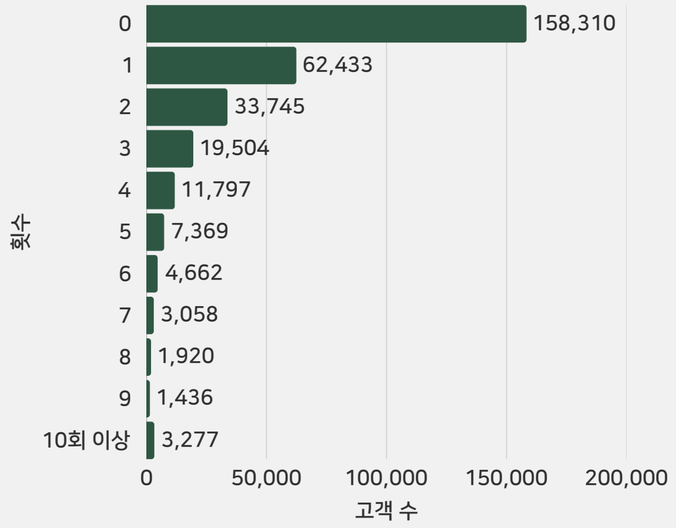
➡ 한도 대비 사용률이 높은 고객은 주의가 필요함

4 KPI 변수 생성 · 현금 대출 / 리볼빙 대출 / 할부 사용 횟수

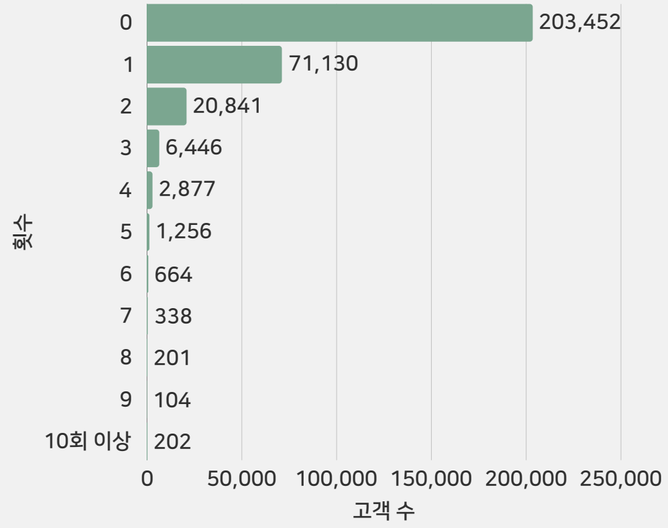
현금 / 리볼빙 대출

사용하지 않은 고객이 가장 많고, 이후 급격히 줄어드는 전형적인 장기 꼬리 분포

현금 대출 사용 횟수별 고객수 분포



리볼빙 대출 사용 횟수별 고객수 분포

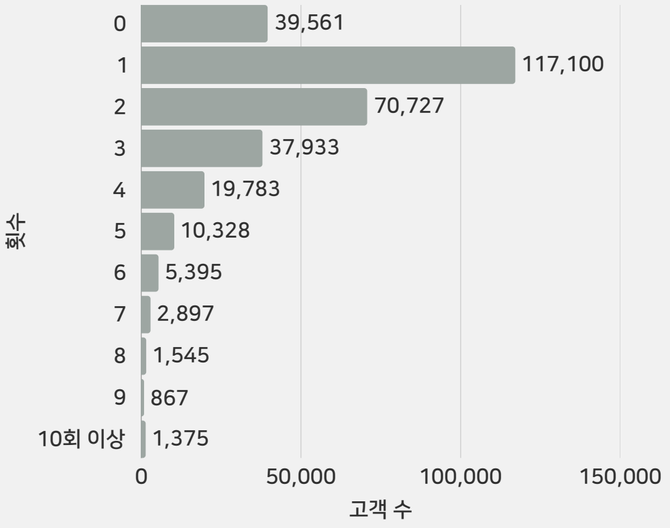


현금·리볼빙은 소수 집중 이용

할부 사용

1~2회 이용 고객이 많으며 횟수가 늘어도 **감소세가 완만하다.**

할부 사용 횟수별 고객수 분포



할부는 다수 반복 이용 패턴

5 스코어보드 설계

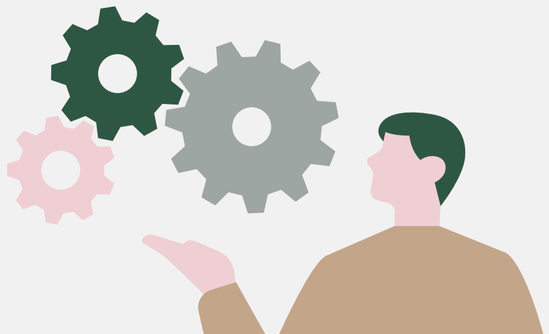
5 스코어보드 설계 도입

가설

신규 고객은 정보 부족으로
기존 고객보다 **점수가 낮을 것이다.**

전략

신규 고객의 점수를 비금융 요소로 보정한다.



5 스코어보드 설계 과정

5.1 선행 연구를 참고하여 WOE / IV 방법론 채택

Information Value (IV) · Weight of Evidence (WOE)

대출 연체 예측을 위해 각 변수의 영향력을 측정하고
이를 **점수화하는 데 활용되는 통계 기법**

구분	항목
WOE	$$WOE_i = \ln\left(\frac{\frac{\text{Good}_i}{\sum_{j=1}^{n}\text{Good}_j}}{\frac{\text{Bad}_i}{\sum_{j=1}^{n}\text{Bad}_j}}\right)$$ 각 구간별로 '연체 고객 비율'과 '정상 고객 비율'의 차이를 수치화한 값 **값이 높을수록 해당 구간이 연체 여부를 잘 구분함**
IV	$$IV = \sum_{i=1}^{n}\left(\frac{\text{Good}_i}{\sum_{j=1}^{n}\text{Good}_j} - \frac{\text{Bad}_i}{\sum_{j=1}^{n}\text{Bad}_j}\right) \times WOE_i$$ 변수 전체의 예측력을 수치화한 값 **값이 클수록 연체 예측에 유용한 변수**

출처: 유진아, 「개인 신용평가 모형을 위한 특성 상호작용 적용 연구」, 고려대학교 SW·AI 융합대학원, 2024.
김가연, 「합성데이터를 이용한 씬 파일러의 연체 특성에 관한 연구」, 서강대학교 , 2023.

5 스코어보드 설계 과정

5.2 점수 변환 과정

1 점수 변환

Z-스케일

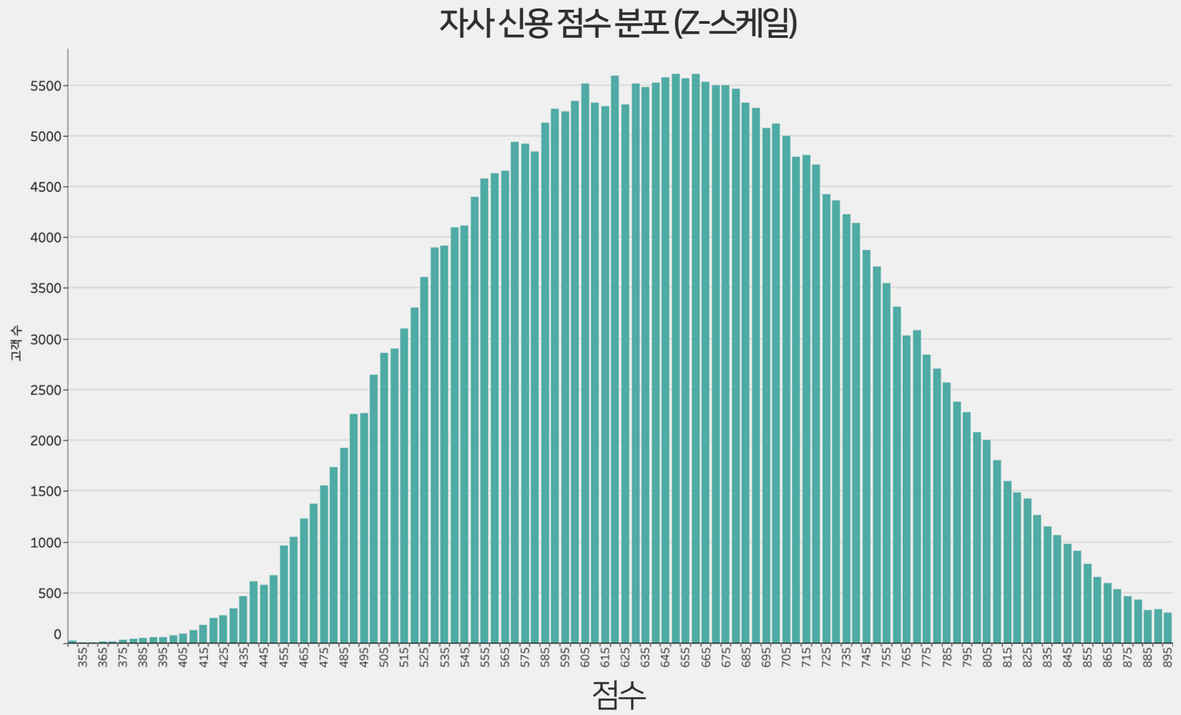
0 ~ 1 점수를 표준화해 0 ~ 900 범위로 변환

고객 등급

분위수 방식을 사용해 전체를 **20%씩**

구간화해 하위부터 **E ~ A 등급** 부여

2 그래프



5 스코어보드 설계 과정

5.3 첫 번째 가설 및 결과

1 가설 및 결과

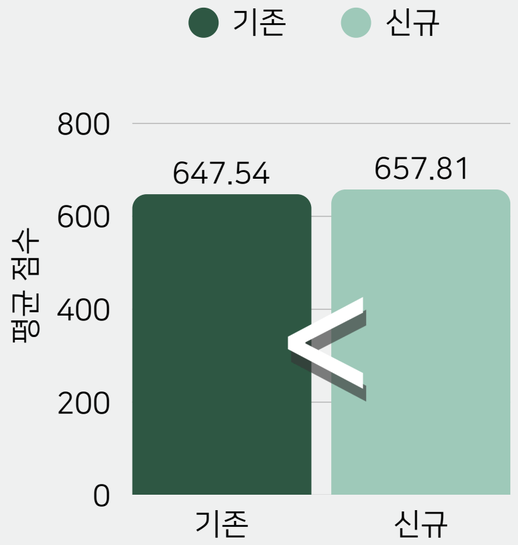
신규 고객은 정보 부족으로
기존 고객보다 **점수가 낮을 것이다.**

▼

예상과 달리 **신규 고객의 점수가 더 높았다.**

+ 13.12 ▲

2 성능 및 평균 점수



모델 성능

AUC: **0.7028** | KS: 3062

고객 위험도 **구분력 양호**

AUC : 모델이 연체 고객과 정상 고객을 얼마나 잘 구분하는지 나타내는 지표
KS: 두 집단(연체 vs 정상)의 점수 분포가 얼마나 차이나는지 보여주는 지표

5 스코어보드 설계 과정

5.4 첫 번째 결과의 원인 분석 및 보정

원인 분석 & 보정

① **Missing 값 과대 평가**

Missing의 **'정보 없음' = '리스크 없음'**으로 해석해 **가점**을 부여한 경우

해결 방안

Missing 범주에 **패널티 부여**

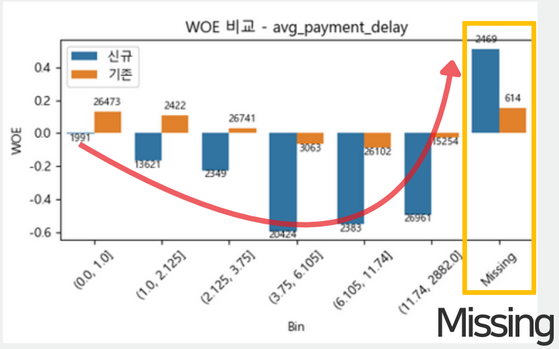
② **저빈도 bin 과대평가**

표본 적은 bin이 **우연히 불량율이 낮게** 나와 WOE가 양수로 계산되어 **가점** 부여한 경우

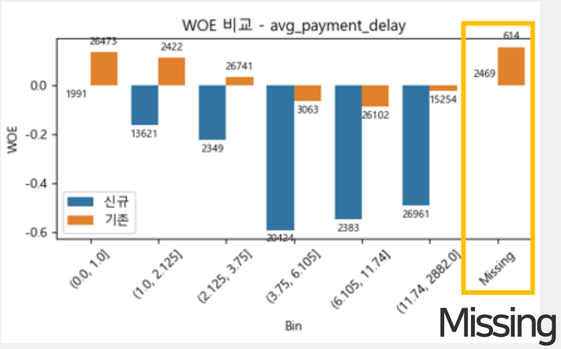
해결 방안

min_count 미만 & WOE > 0 인 구간에 WOE를 조정해 **과대 가점을 제거**

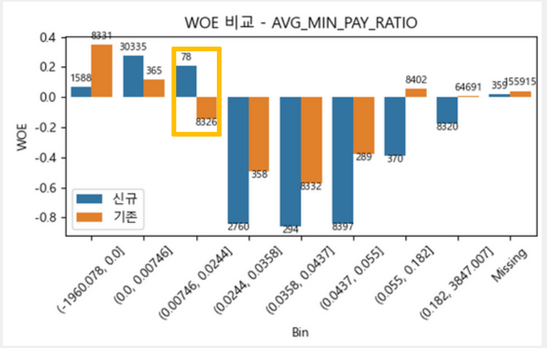
하락하는 추세를 보이다
Missing 범주에서만 양수로 전환 → Missing 범주에 패널티 부여



Missing



Missing



신규
기존

5 스코어보드 설계 과정

5.5 점수 보정 추가 시도

1 추가 시도

신규 점수 상승에 기여한 피처 **전체에 패널티**

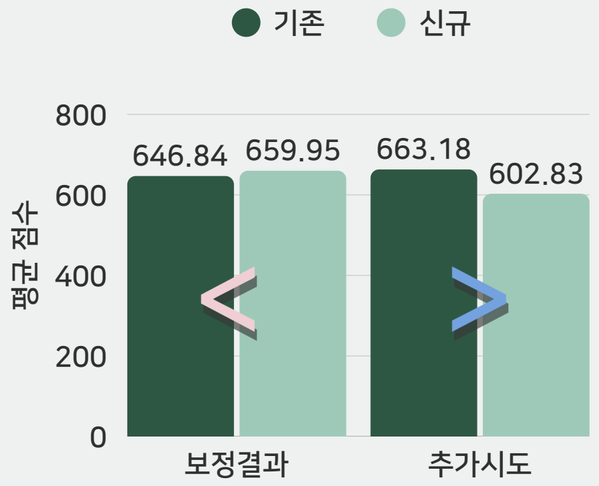
모델 성능 하락

AUC 0.70 → 0.68

KS 0.30 → 0.26

2 점수 비교 그래프

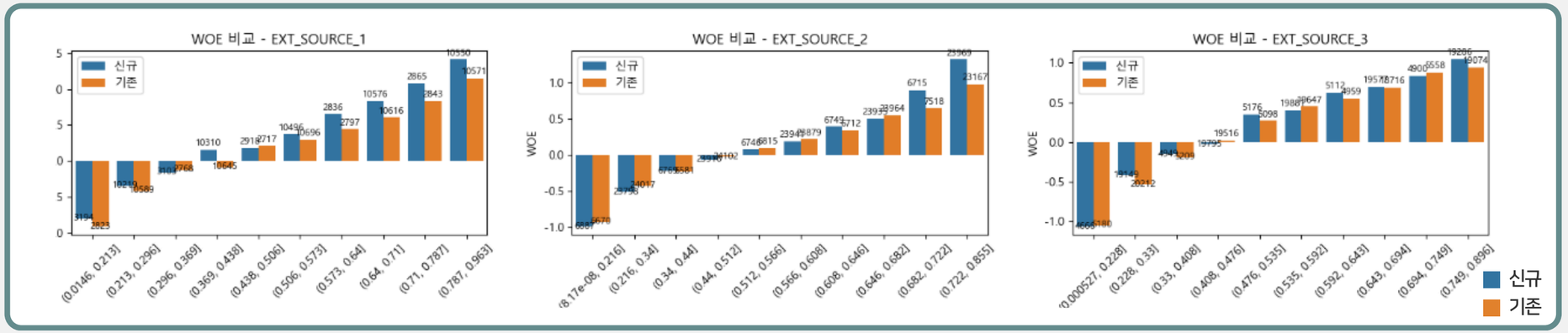
(신규) 모든 피처에 **일괄 패널티**를 적용하는 방식은
데이터적 근거가 부족하다고 판단



5 스코어보드 설계 과정

5.6 분석 결과 및 가설 기각

▲ 신규 고객이 실제로 더 우량하게 나타남



분석 결과

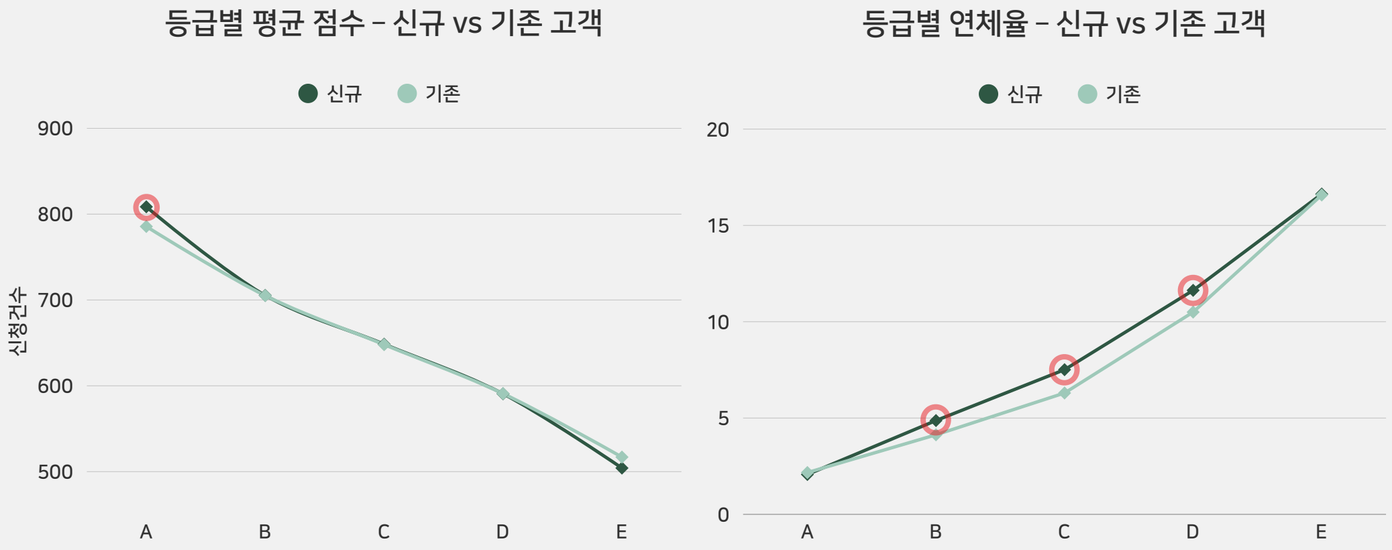
일부 변수인 **외부 신용 점수**에서 기존 고객 보다
신규 고객의 WOE가 높거나 비슷한 수준을 보인다.

가설 기각

신규 고객은 정보 부족으로 기존 고객보다 점수가 낮을 것이다.

5 스코어보드 설계 과정

5.7 새로운 가설을 위한 분석



기존 고객 = 다중 채무자

신규 고객 ≠ 신용 이력 없음

신용 이력에 따른 추가 분석 필요

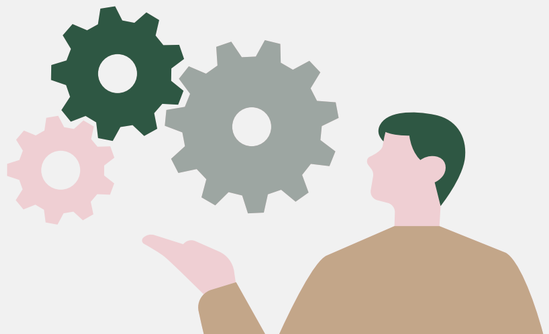
B~D 등급은 점수 차이는 미미하지만, **신규 고객**의 **연체율이 전반적으로 높음**

기존 고객은 다중 채무 비율이 높아, **상대적으로 낮은 평가를 받은 경향**

5.8 새로운 가설 수립

새로운 가설

신규 고객 중 **신용이력이 없는 고객**은 **점수와 연체율이 더 높을 것**이다.



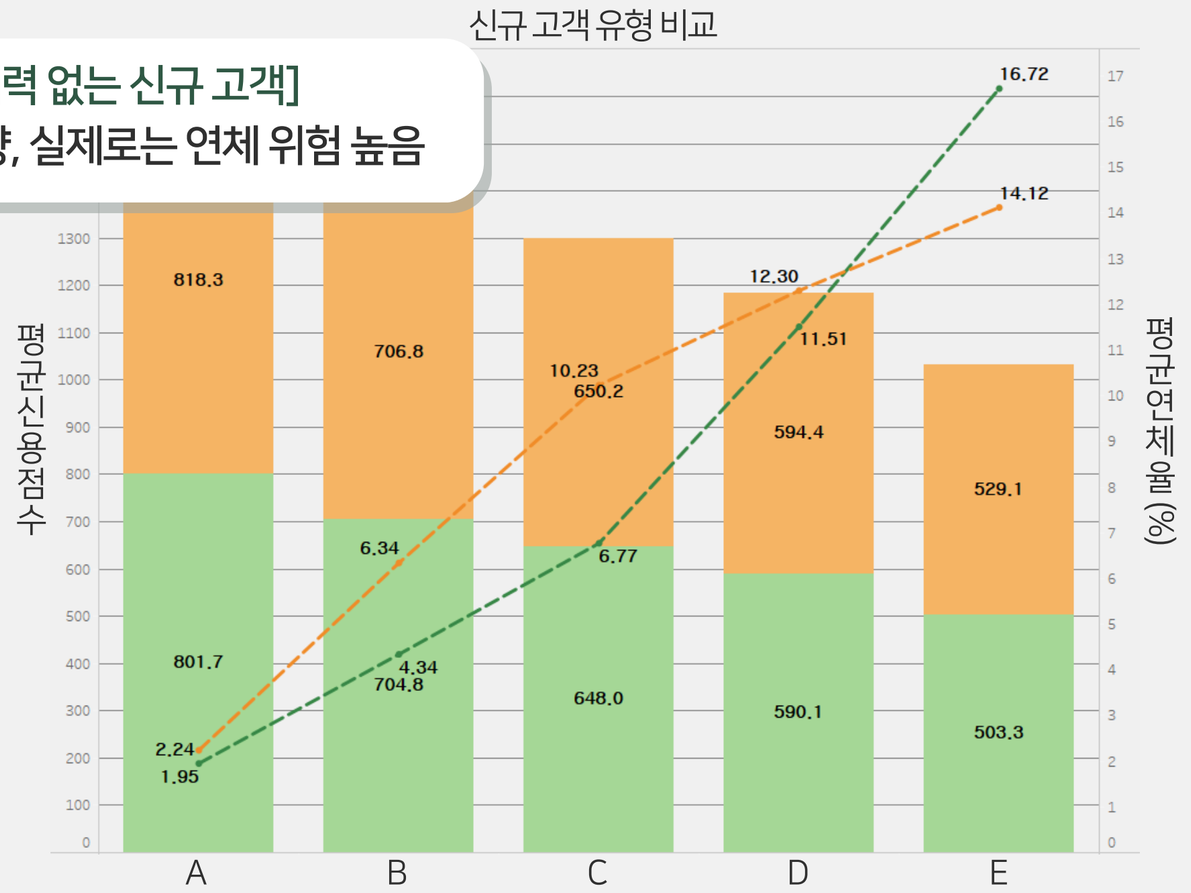
5 스코어보드 설계 과정

5.9 가설 분석 결과

신용점수 (막대)
- 신용이력없는신규
- 신용이력있는신규

연체율 (라인)
- 신용이력없는신규
- 신용이력있는신규

[신용 이력 없는 신규 고객]
표면적으로 우량, 실제로는 연체 위험 높음



해석 요약

전반적으로 **신용이력이 없는 고객**이
신용이력이 있는 고객보다 **평균 점수가 높게 나타남**

E 등급 제외 전 등급에서 신용이력이 없는신규 고객의
점수가 더 높음에도 불구하고 **연체율 또한 높게 나타남**

5 스코어보드 설계 과정

5.10 스코어카드 적용 예시 및 기대 효과

구분	외부점수	내부점수
데이터 결측률	최대 30% 이상 결측 (EXT_SOURCE_1 기준)	**모든 고객**에 대해 **산출 가능**
점수 산출 방식	외부 표준화 규칙 적용	자사 고객 특성 반영 맞춤형 규칙
적용 가능 고객	외부 데이터가 있는 고객만	전체 고객군
활용 목적	외부 기준에 따른 위험 평가	자사 영업·리스크 관리 전략 최적화
장점	업계 공통 비교 가능	결측 문제 없음, 내부 정책 반영 가능, 고객군 세분화 관리

5 스코어보드 설계 과정

5.11 스코어카드 적용 예시 및 기대 효과

01 | 모든 고객에게 적용 가능

02 | 자사 특성 반영 가능

03 | 자사 영업 · 리스크 관리 최적화

04 | 세분화된 고객 관리 가능

6 결론 및 시사점

6 결론 및 시사점

6.1 주요 인사이트 요약

01 **신규 고객군**의 **우량함**이 외부·내부 데이터 모두에서 확인

02 **기존 고객군**(다중채무자)에서 위험도가 높아 **심사 정책**에서 **주의 필요**

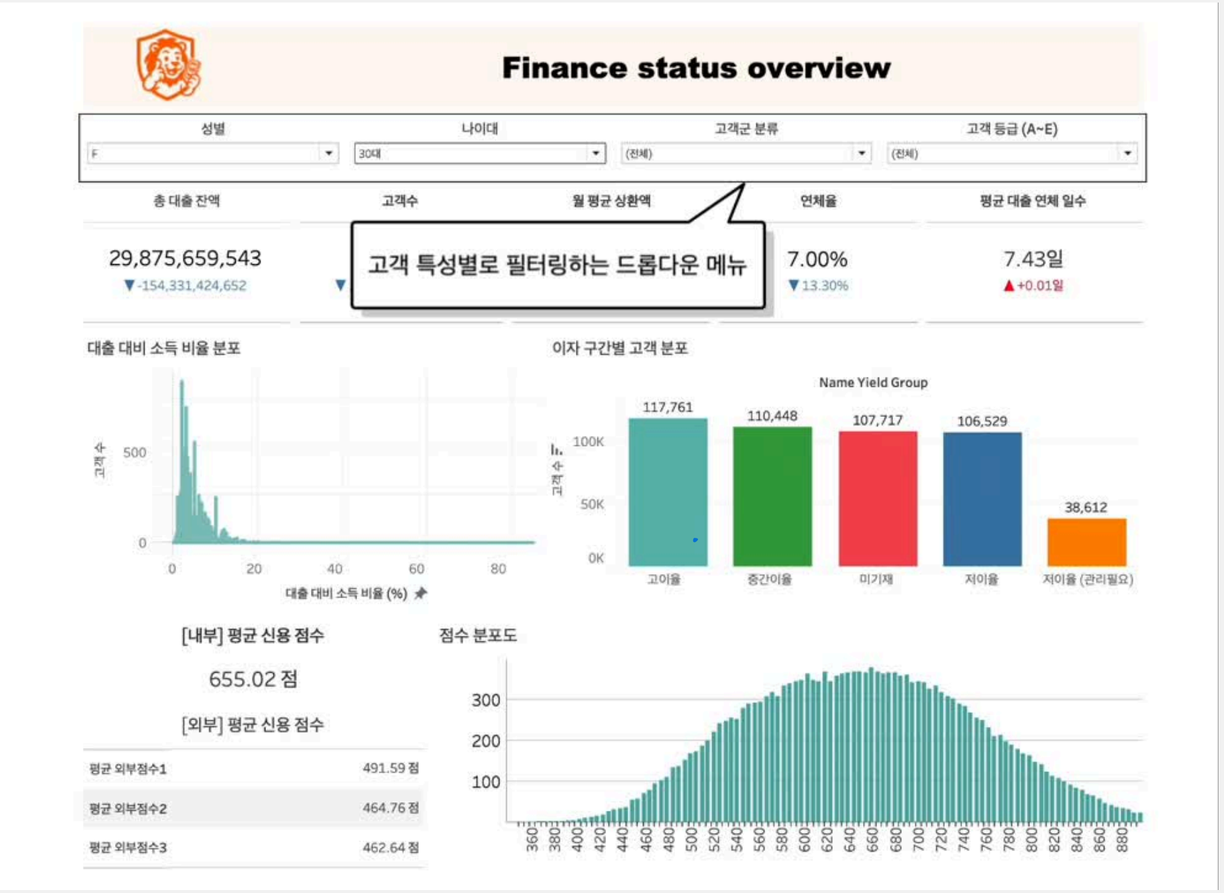
03 **자사 점수 체계**가 **외부 점수** 대비 **위험 구분력**이 **동등**하거나 **우수한 변수 존재**

➡ 외부 신용점수가 없는 고객에게도 비슷한 점수대 분포 형성

➡ **리스크 축소 우려**가 있어 **모니터링**과 **패널티 부여** 필요성 제시

6 결론 및 시사점

6.2 실제 업무 적용 기대효과



고객군 별 모니터링 대시보드

대시보드 개요

- 고객 대출 현황, 상환 능력, 신용 상태 종합 분석
- 성별·나이대·고객군·등급별 필터로 특정 집단 비교

핵심 지표

1. 총 대출 잔액, 고객 수
2. 월 평균 상환액
3. 연체율
4. 평균 대출 연체 일수

6 결론 및 시사점

6.3 한계점 및 개선방안

01

표본 불균형

일부 직업군·연령대·대출유형
비중이 과도하게 높아 통계 왜곡 가능

표본 가중치 조정, 분석 시 그룹별
비율 균형 맞추는 **층화분석 적용**

02

결측치 처리 한계

Missing 값이 특정 구간에서
과대평가되거나 과소평가됨

결측치 원인 분석 후 그룹별 대체
또는 별도 카테고리화로 **편향 최소화**

03

변수 범위 제한

금융 거래 중심 변수에 치중, **비금융 생활 데이터**(통신, 공과금, 전자상거래 등) **부재**

다양한 대안신용정보를 **포함**해
고객의 **다면적 신용도 평가** 가능

감사합니다

궁금한 점 있으시면 편하게 질문해주세요