finda

앱 사용성 데이터를 통한 대출신청 예측분석

데이터 분석분야 퓨처스 부문

팀명 : Import finda

박성수 tjdtn3248@naver.com

박준영 pjuny0@naver.com

백찬진 anback14@gmail.com

송창용 thdckdyd123@naver.com

Index.

01

개요

- 문제 정의

02

대출신청 예측

- 데이터 전처리
- Feature Engineering
- Modeling
- Result

03

고객 유형 분석

- RFM With Finda
- Modeling
- 고객 유형 분석
- 서비스 메시지 제안

01. 개요

- 문제 정의

1. 대출신청 예측

핀다 홈화면 진입 고객 중 특정기간 안에 대출신청 고객 예측

02. 고객 유형 분석

모델 기반 고객 군집분석 및 군집별 서비스 메시지 제안

02 대출신청 예측

user_spec 데이터 전처리

1. 결측치 처리

총 14개의 컬럼에 결측치 존재

birth_year	credit_score	desired_amount	purpose	gender	personal_rehabilitation_yn	existing_loan_cnt	yearly_income	p
1985.0	660.0	1000000.0	기타	1.0	0.0	4.0	108000000.0	
1968.0	870.0	30000000.0	대환대출	1.0	0.0	1.0	30000000.0	
1997.0	710.0	10000000.0	생활비	1.0	0.0	5.0	30000000.0	
1989.0	820.0	2000000.0	생활비	1.0	0.0	7.0	62000000.0	
2000.0	630.0	5000000.0	생활비	1.0	0.0	1.0	36000000.0	
...	...	...	...	...	...	...	...	
2000.0	590.0	5000000.0	사업자금	1.0	NaN	NaN	25000000.0	
1955.0	980.0	50000000.0	생활비	1.0	NaN	1.0	20000000.0	
1983.0	750.0	100000000.0	대환대출	1.0	NaN	8.0	75000000.0	
1975.0	640.0	10000000.0	대환대출	1.0	NaN	10.0	50000000.0	
1977.0	NaN	20000000.0	생활비	0.0	NaN	NaN	35000000.0	

02 대출신청 예측

- user_spec 데이터 전처리

결측치를 채워줌

변수명	방법
birth_year	평균으로 대체
company_enter_month	
credit_score	
desired_amount	최빈값으로 대체
purpose	
gender	많은 값인 1로 채움
personal_rehabilitation_yn	분석 결과 해당 열이 Na인 것은 미해당으로 판단하여 0으로 대체
yearly_income	
extsting_loan_cnt	
personal_rehabilitation_complete_yn	개인회생자가 아닌 것으로 채움
income_type	분석 결과 소득이 없거나 주거에 대한 정보를 판단할 수 없었기에 '없음'으로 채움
houseown_type	

02 대출신청 예측

loan_result 데이터 전처리

1. 중복 데이터 제거

총 26개의 데이터가 중복

```
loan_table[loan_table['application_id'] == 2160853]
```

	application_id	loanapply_insert_time	bank_id	product_id	loan_limit	loan_rate
1507073	2160853	2022-06-16 09:47:22	59	150	24000000.0	16.4
1507088	2160853	2022-06-16 09:47:24	63	226	15000000.0	18.7
1507090	2160853	2022-06-16 09:47:23	10	65	29000000.0	15.6
1507095	2160853	2022-06-16 09:47:23	10	65	29000000.0	15.6

중복된 데이터를 제거해줌

2. 결측치 처리

Loan_limit, loan_rate 변수 결측치 처리

	application_id	loanapply_insert_time	bank_id	product_id	loan_limit	loan_rate	i
461	1029177	2022-06-07 15:29:06	51	21	NaN	NaN	
4318	1086409	2022-06-07 18:11:31	13	262	NaN	NaN	
6971	1193826	2022-06-07 17:13:03	1	202	NaN	NaN	
8459	447492	2022-06-07 17:12:08	30	232	NaN	NaN	
9220	260060	2022-06-07 11:45:43	30	85	NaN	NaN	
...	...	...	...	...	...	...	
13521728	1222550	2022-06-03 16:51:24	13	262	NaN	NaN	
13522701	135727	2022-06-03 10:59:08	10	149	NaN	NaN	
13523316	687402	2022-06-03 12:12:31	1	102	NaN	NaN	
13523827	621491	2022-06-03 17:05:01	51	21	NaN	NaN	

해당 고객에게 추천된 제품의
최다 한도와 금리로 결측값 대체

02 대출신청 예측
Feature Engineering

1. 일반 파생 변수

파생변수 생성기준

1. 신청 대출과 미신청 대출을 잘 분류할 수 있는 파생변수 생성
2. 사용자의 특징을 잘 나타낼 수 있는 변수
3. 대출 상품의 특성을 잘 나타낼 수 있는 변수

약 30여 개의 변수를 생성함

대출 이력
연령대
신용 등급
소득 등급
근속 년수
.
.
.
기대출 평균 금액
대출 조회 건수
신용 평가 시간
Income type별 소득 수준

02 대출신청 예측
Feature Engineering

2. 중요 파생 변수

"사용자가 신청하는 대출의 특징을 분석하기 이전에 사용자에 대한 분석을 진행하고 사용자와 관련된 변수를 생성"

대출 이력 (loan_history)

이전에 Finda에서 대출을 신청한 이력이 있는 사용자는 상품을 신청할 때 거부감이 적을 것으로 판단하여 변수 생성

추천 은행 수 (n_bank)

추천 받은 은행의 수가 많은 사용자는 선택의 폭이 비교적 넓을 것으로 판단하여 대출 신청의 가능성이 높을 것으로 생각

추천 상품 수 (n_product)

추천 은행 수와 마찬가지로 상품 선택의 폭이 넓을 것으로 판단

02 대출신청 예측
Feature Engineering

2. 중요 파생 변수

“ 사용자가 Finda에서 대출을 선택하는 과정을 분석하여, 선택하는 상품에 대한 지수를 계산하고 변수로 사용함 ”

	is_applied	loan_limit	loan_rate	desired_amount
10269865	1.0	100000000.0	7.8	100000000.0
10269883	0.0	100000000.0	7.9	100000000.0
10269897	0.0	100000000.0	9.9	100000000.0

선택 대출과 미선택 대출의 예시

사용자는 본인이 원하는 금액과 추천 대출의 한도가 비슷하고, 대출의 금리가 낮은 상품을 선택하는 경향이 있음

02 대출신청 예측
Feature Engineering

2. 중요 파생 변수

“ 사용자가 Finda에서 대출을 선택하는 패턴을 분석하여, 선택하는 상품에 대한 지수를 계산하고 변수로 사용함 ”

대출 지수
(loan_coefficient)

수식
(사용자가 원하는 금액 - 승인 한도) * 승인 금리

즉, 사용자가 원하는 금액과 승인 한도가 비슷하고
승인 금리가 작을 수록 1에 가까운 값을 가짐.

선택 대출과 미선택 대출의 예시

또한, 사용자가 원하는 금액과 승인한도의 차이가 크고
금리가 클수록 큰 값을 가짐

사용자는 본인이 원하는 금액과 추천 대출의 한도가 비슷하고, 대출의 금리가 낮은 상품을 선택하는 경향이 있음

02 대출신청 예측
Feature Engineering

3. 변수 중요도 검증

Feature Importance(CART)?

변수의 중요도를 판단하는 방법이다.

특성을 기준으로 샘플들을 spit할 때,

불순도가 감소하는 양을 바탕으로 불순도 감소가 클 경우 중요한 변수라고 판단하는 방법

다른 변수들과 상관관계를 배제한 독립적인 변수의 중요도를 확인하지 못한다는 단점을 가짐

Permutation Importance?

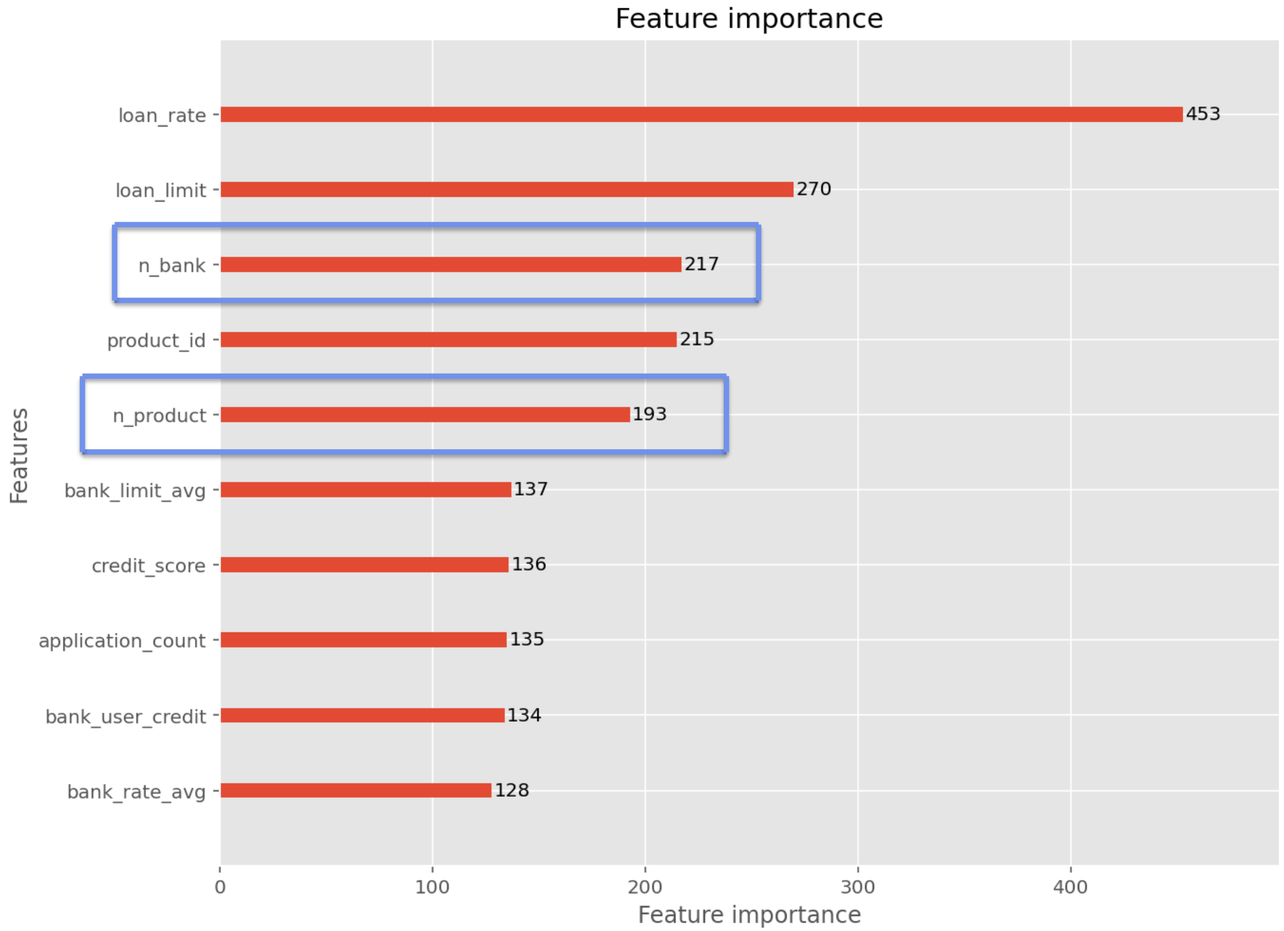
CART 방식이 가진 단점을 보완한 방법으로 변수들 간의 상호 관계를 배제하고 독립 변수의 중요도를 확인하는 방법

위의 두가지 방법을 모두 고려하여 변수의 중요도를 확인하는 것이 가장 바람직함!

02 대출신청 예측
Feature Engineering

3. 변수 중요도 검증

CART Feature Importance



Permutation Importance

Weight	Feature
0.2265 ± 0.0006	loan_history
0.2237 ± 0.0011	n_bank
0.1295 ± 0.0006	n_product
0.0737 ± 0.0009	bank_rate_avg
0.0629 ± 0.0014	loan_coefficient
0.0589 ± 0.0006	loan_rate
0.0540 ± 0.0007	credit_score
0.0218 ± 0.0005	product_id
0.0208 ± 0.0006	application_count
0.0173 ± 0.0004	bank_id

모든 변수 중 중요한 변수 상위 10개씩을 도출한 결과이고,

생성한 중요 파생 변수의 점수가 높음을 알 수 있음.

02 대출신청 예측

Encoding & Scaling

Encoding

Label Encoding

1. 데이터의 크기가 매우 크기 때문에
   모델 상용화시 OnehotEncoding 보다
   속도가 빠르다는 장점이 존재함

2. 범주형 변수의 **고유 값들이 매우 많았**기 때문에
   OneHotEncoding을 할 경우 데이터가
   매우 Sparse해진다는 문제점이 있기 때문에
   LabelEncoding을 사용함

Scaling

RobustScaler

1. 연간 소득, 대출 희망 금액 등의 변수에서 값의
   차이가 매우 큰 **이상치가 존재**하였기 때문에
   이상치의 영향이 적은 Scaling방법인
   RobustScaler를 사용함

2. 중앙값과 IQR을 사용하여 Scaling을 하기 때문에
   데이터의 분포를 더욱 넓게 나타낼 수 있고,
   이는 대출 신청 여부를 분류하는데 효과적일 것으로
   판단함

02 대출신청 예측

Modeling

1. 모델 설명 - 머신 러닝 모델

LGBM

큰 사이즈의 데이터를 효과적으로 다룰 수 있기 때문에
실제 모델로 사용 시 속도가 빠르다는 장점을 가짐

Catboost

이름에서도 알 수 있듯,
범주형 변수의 특징을 효과적으로 사용할 수 있는 모델

RandomForest

예측의 변동성이 낮음,
배깅 방식을 사용하기 때문에 결측치에 강건한 모델

DescisionTree

분석의 결과를 직관적으로 확인할 수 있는 장점을 가진 모델

XGB

과적합에 대한 규제가 있기 때문에 과적합에 강한 모델,
병렬처리로 속도가 빠르기 때문에 실무에 적용시 속도가 빠름

02 대출신청 예측

Modeling

1. 모델 설명 - 딥러닝 모델

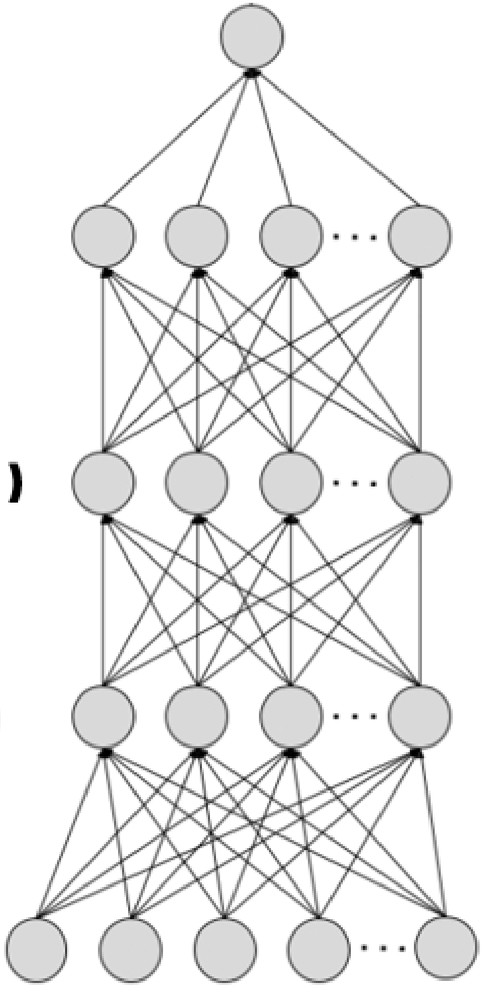
Output layer (1)
sigmoid

Hidden layer 3 (64)
relu

Hidden layer 2 (128)
relu

Hidden layer 1 (64)
relu

Input Layer (158)



Epoch: 3

Batch size: 1024

dropout: 0.4

Loss Function: BCELoss

Optimizer: Adam

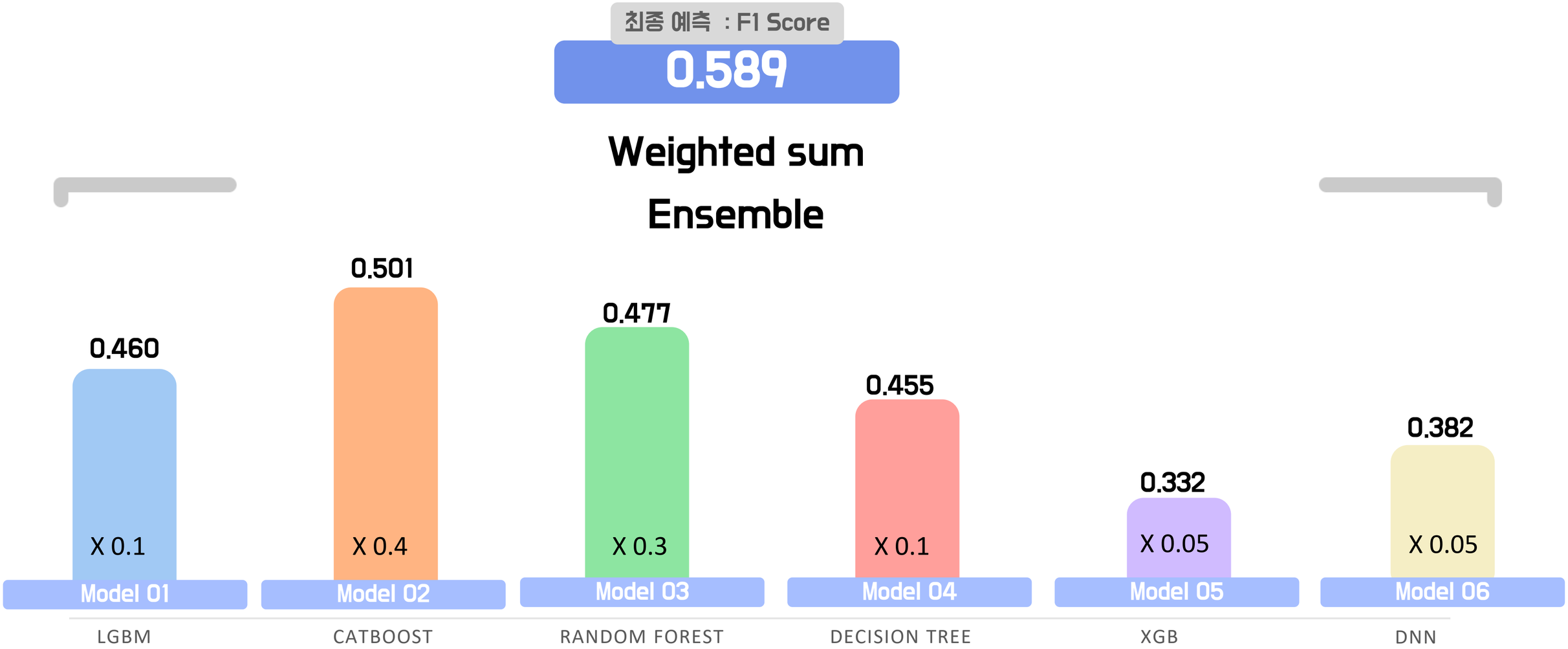
Learning Rate: 0.01

Embedding Size: 122

02 대출신청 예측
Modeling

2. 최종 모델

평가지표 : F1 Score



02 대출신청 예측

Modeling

2. 최종 모델

평가지표 : F1 Score

Weighted sum Ensemble

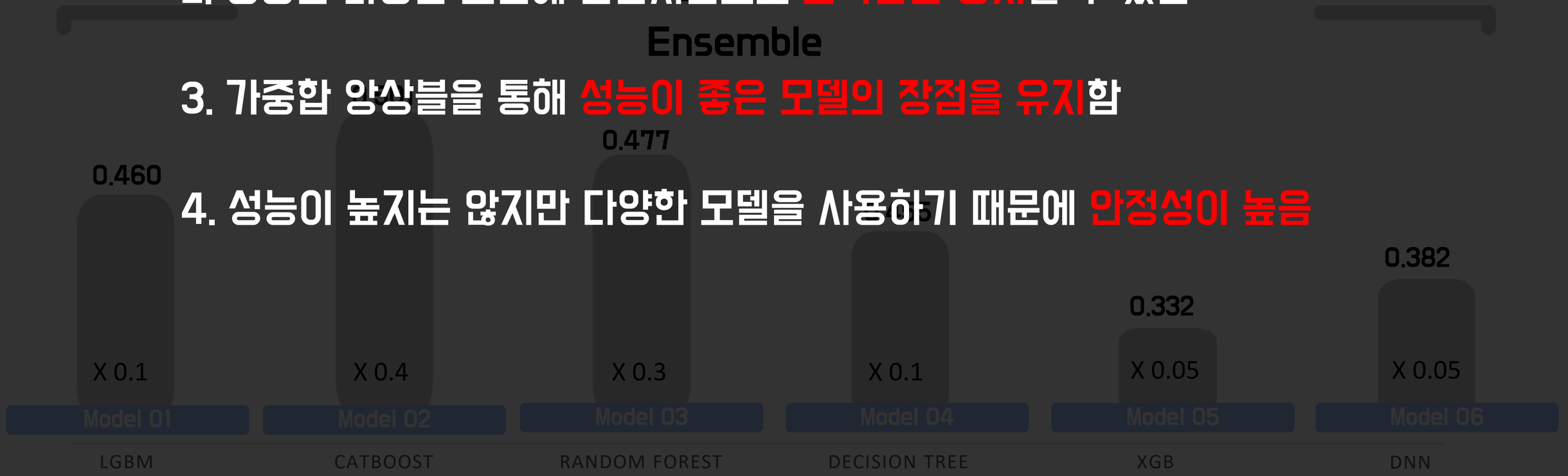
1. 다양한 장점을 가진 모델을 앙상블 함으로 **예측 변동성을 줄일 수 있음**

2. 성능을 다양한 모델에 분산시킴으로 **과적합을 방지**할 수 있음

Ensemble

3. 가중합 앙상블을 통해 **성능이 좋은 모델의 장점을 유지**함

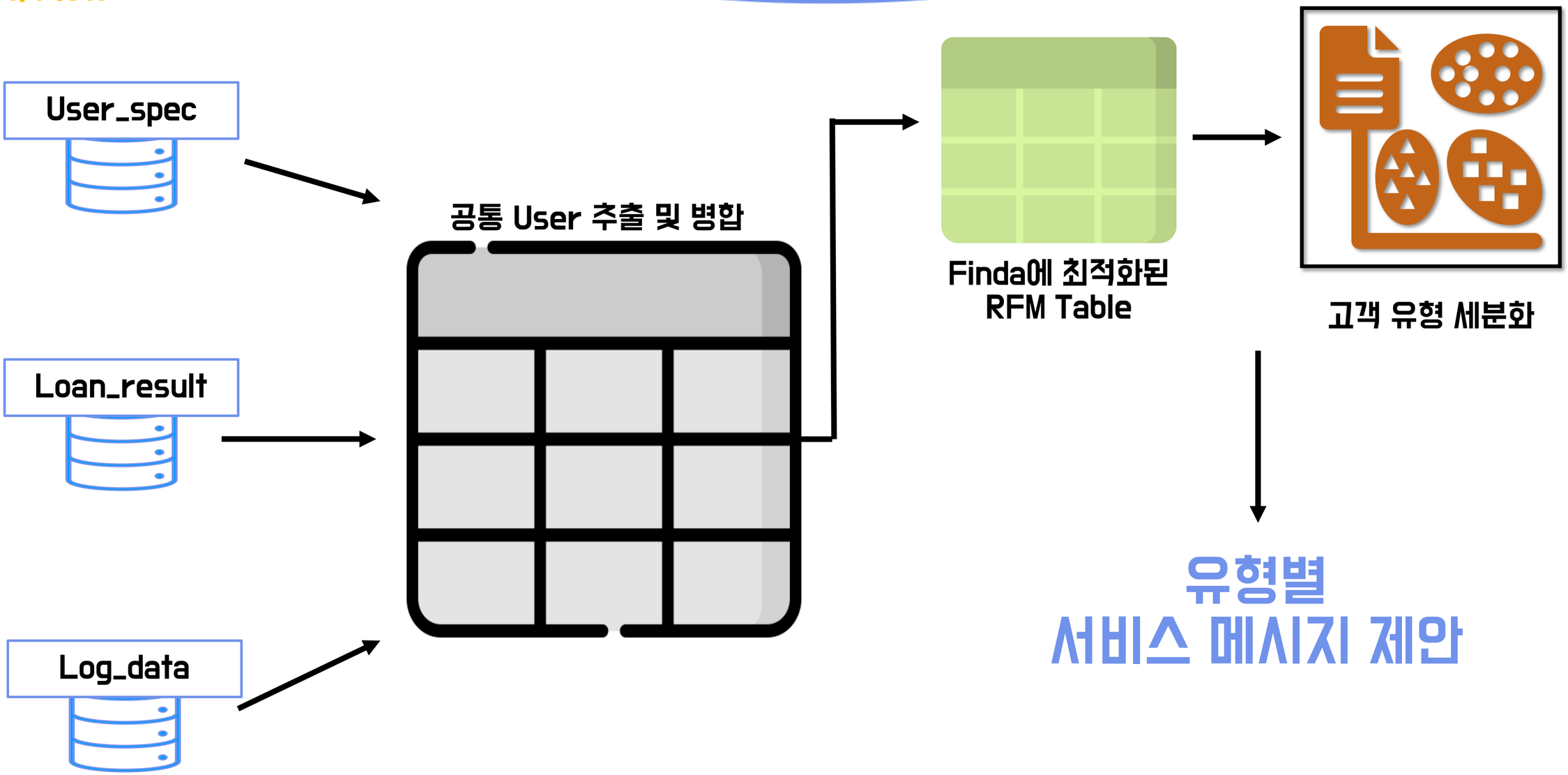
4. 성능이 높지는 않지만 다양한 모델을 사용하기 때문에 **안정성이 높음**



03 고객 유형 분석

RFM With Finda

1. Flow



03 고객 유형 분석

RFM With Finda

1. Flow

User_spec

Loan_result

Log_data

Pre-processing

1. User_spec

- 이전과 동일한 방법으로 결측치 처리

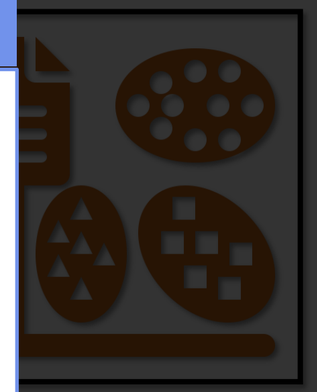
2. Loan_result

- 이전과 동일한 방법으로 결측치 처리 및 중복값 제거

3. Log_data

- mp_os 결측치
  -> 이전에 os기록이 있는 경우 이전 기록의 최빈값으로 대체
  -> 이전에 기록이 없는 경우 전체 사용자의 최빈값으로 대체

- mp_app_bersion
  -> 결측치가 존재하는 날짜별로 최빈값으로 대체



객 유형 세분화

제안

03 고객 유형 분석

RFM With Finda

1. Flow

User_spec

Loan_result

Log_data

공통 User 추출 및 병합

Finda에 최적화된
RFM Table

세가지 데이터에 모두 공통적으로
존재하는 사용자를 분석 대상으로 함
즉, 약 29만명의 사용자가 분석대상

유형별
서비스 메시지 제안

03 고객 유형 분석

With Finda

RFM with Finda

R

Recency

고객이 최근에 finda를 사용한 정보를 알기 위해 log_data를 이용하여 가장 최근에 앱을 사용한 날짜를 구함

최근에 얼마나 Finda에 관심이 있는지 알 수 있음

F

Frequency

핀다 앱에 얼마나 관심을 두고 앱을 사용하는지 알기 위하여 OpenApp(앱시작) 버튼을 누른 횟수를 계산하여 사용

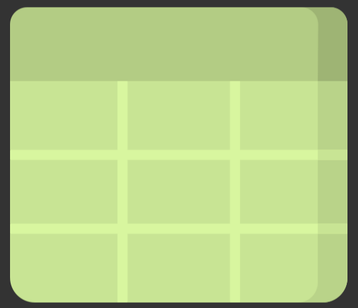
앱을 많이 사용하는 유형에 대한 분석이 가능함

M

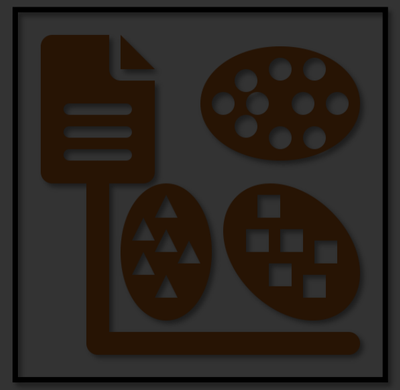
Monetary

해당 변수는 사용자의 구매력을 의미하는 변수인데, 이를 finda에 알맞게 변경하기 위하여 Credit_score 데이터를 사용함

신용 등급에 따른 유형별 차이를 확인하는 것이 가능함



Finda에 최적화된
RFM Table



고객 유형 세분화

RFM이란?

고객 구매 데이터를 바탕으로 고객을 군집화 하는데
일반적으로 많이 쓰이는 방법으로 사용자를
그룹(또는 등급)을 나누어 분류 하는 분석 기법

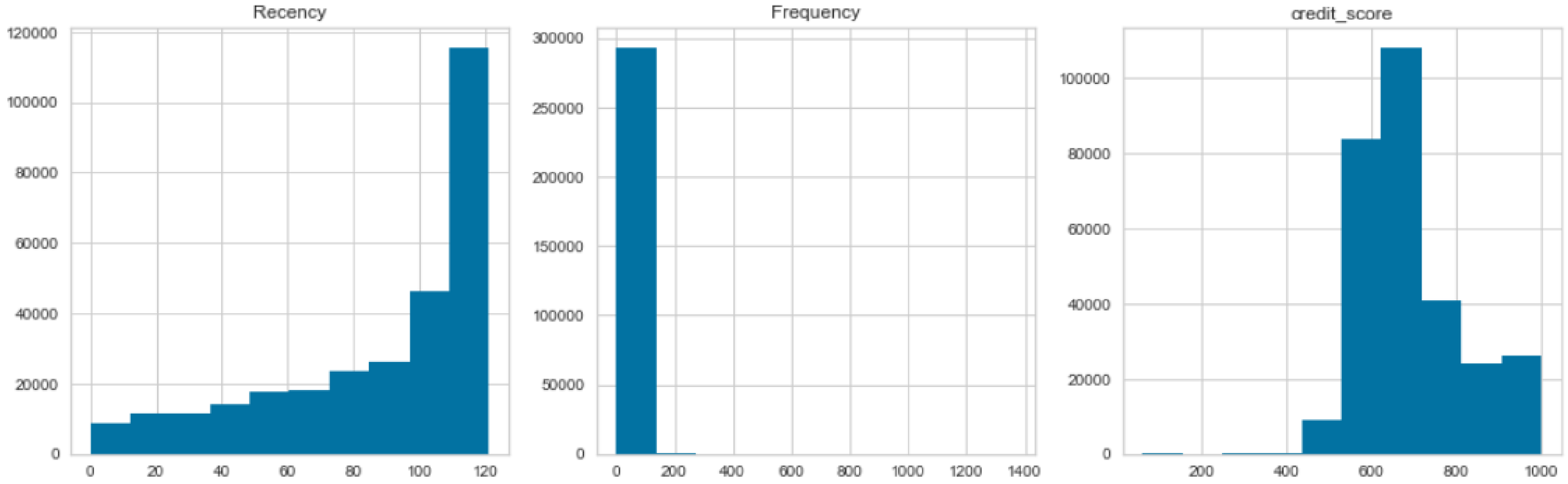
Recency (거래 최근성): 얼마나 최근에 구입했는가?
Frequency (거래빈도): 얼마나 빈번하게 우리 상품을 구입했나?
Monetary (거래규모): 구입했던 총 금액은 어느 정도인가?

-> 하지만 Finda의 데이터는 구매 데이터가 아니기 때문에
RFM방식을 차용하여 Finda에 최적화된 기준 변수를 생성

03 고객 유형 분석

RFM With Finda

2. 선정한 기준 변수의 분포 확인



Frequency 변수의 경우 값이 매우 큰 이상치가 존재하기 때문에 RobustScaler를 사용하여 Scaling을 진행한 후 군집화

03 고객 유형 분석

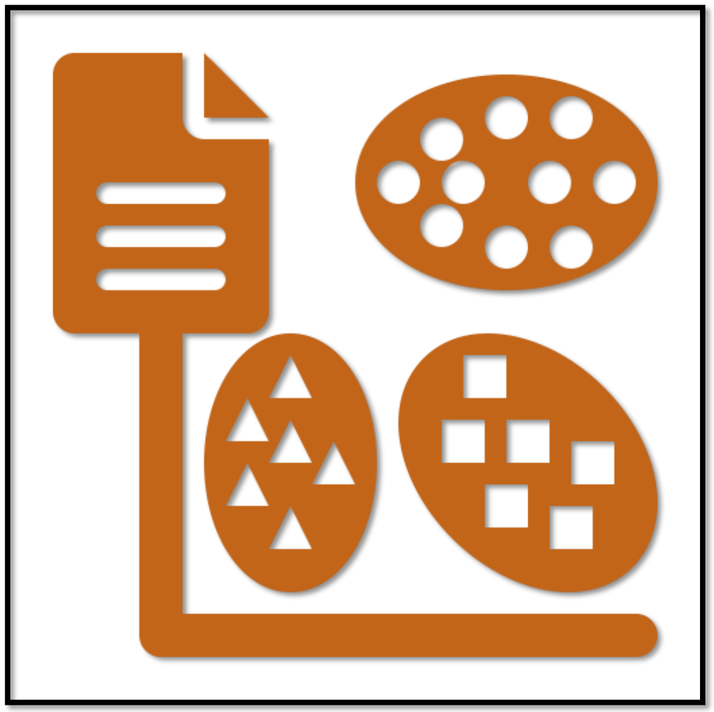
RFM With Finda

3. K-means Clustering

	user_id	Recency	Frequency	credit_score
0	9	-0.220843	-0.541150	0.993080
1	11	0.475164	-0.236046	-0.202462
2	14	0.989604	-0.541150	-0.396241
3	17	-2.490431	-0.358088	-1.195370
4	19	-0.432671	-0.541150	-0.396241
...	...	...	...	...
293511	879691	-2.278603	0.374160	-0.546940
293512	879692	-1.007634	-0.236046	0.830973
293513	879693	0.959342	0.008036	-0.157882
293514	879695	-0.039276	-0.541150	-0.396241
293515	879696	-2.278603	-0.480129	-0.627994

Scaling된
RFM 테이블

K-means Clustering

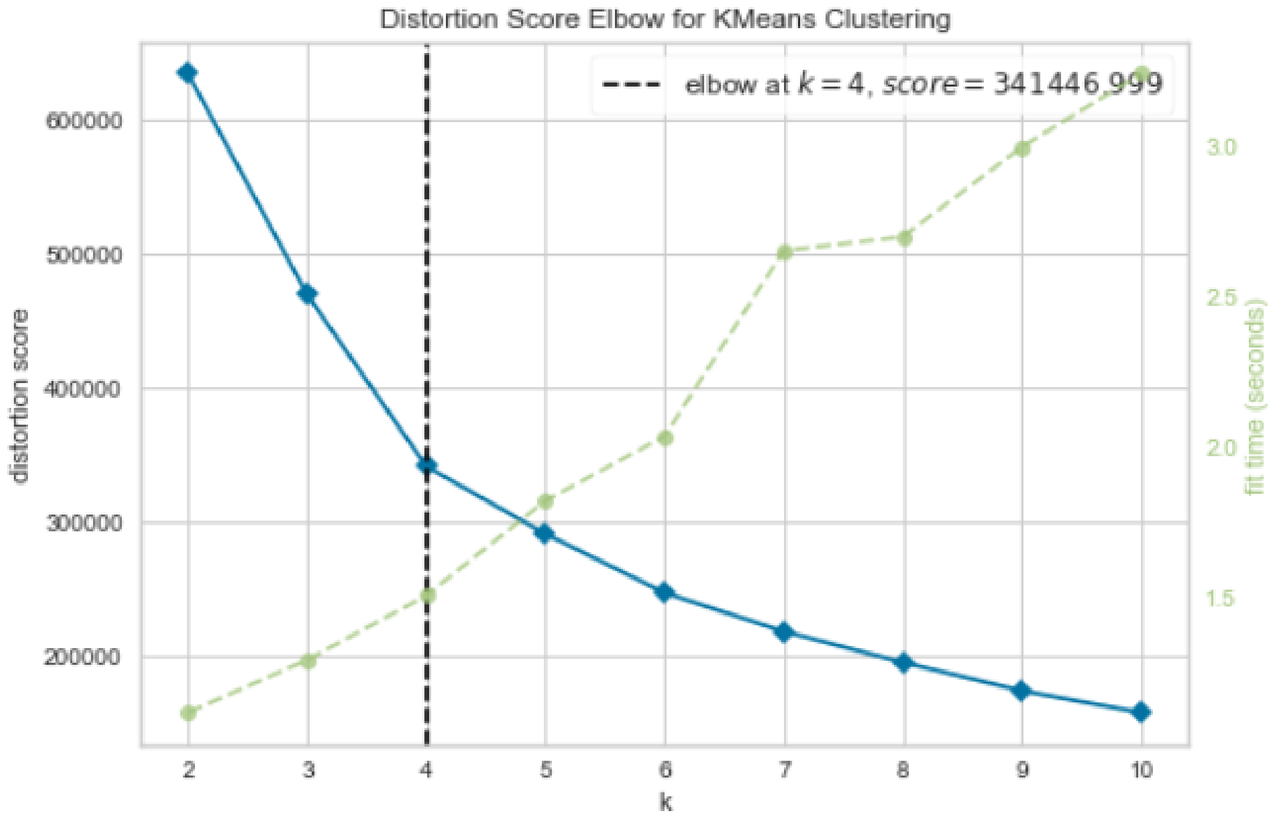


고객 유형 세분화

03 고객 유형 분석

RFM With Finda

3. K-means Clustering



Elbow point

KMeans는 주어진 데이터를 K개의 군집으로 그룹화하는 알고리즘

이때 군집의 갯수 K개는 하이퍼파라미터로, 각 군집간의 거리의

합인 inertia가 급격이 떨어지는 구간을 K로 설정해야 함

이를 Elbow Method라고 함

고객별 RFM 테이블의 K-means Clustering 실행 결과,

Elbow point가 K=4 구간이므로,

고객 유형을 **4개 군집**으로 세분화함

03 고객 유형 분석

RFM With Finda

3. K-means Clustering

실루엣 계수?

각 데이터 포인트와 주위 데이터 포인트들과의 거리 계산함

군집 안에 있는 데이터들은 잘 모여있는지, 군집끼리는

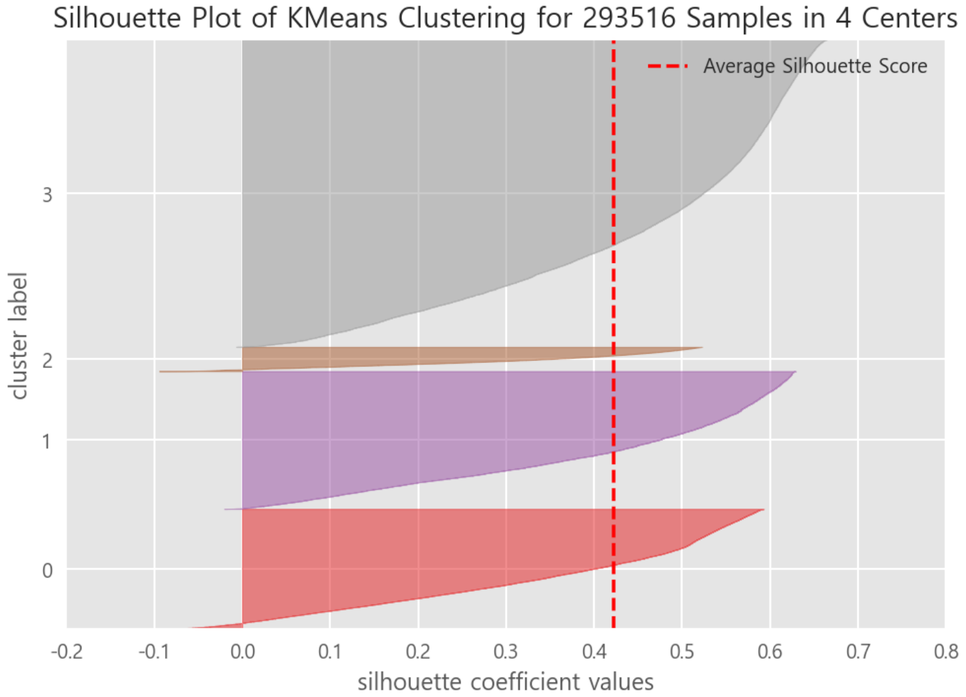
서로 잘 구분되는지 클러스터링을 평가하는 척도

Clustering을 위해 k-means, k-medoids, DBSCAN,

Gaussian Mixture Model, 등 다양한 방법 사용

실루엣 계수와 각 군집의 요소 수를 고려하여

최종 분석 모델로 k-means방식을 사용함



이전의 분석 결과를 실루엣 계수를 통해 검증함

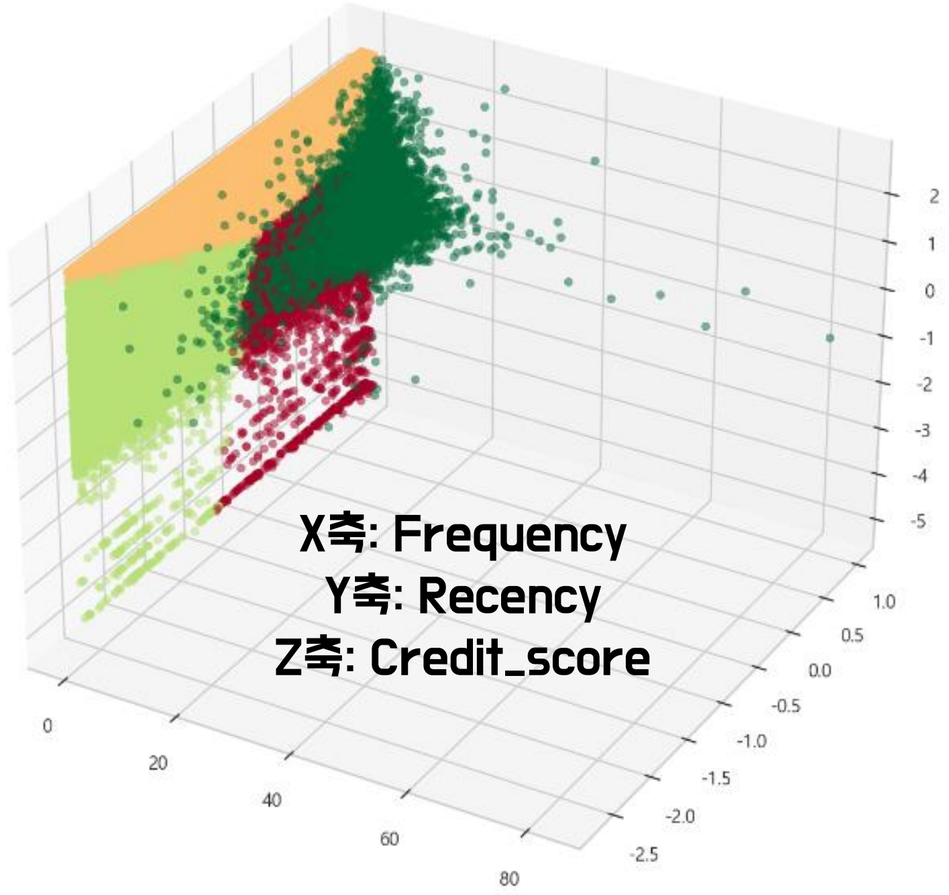
검증 결과 실루엣 계수는 0.4로 양호한 편이고,

각 군집에 적절하게 사용자가 분포함을 확인

03 고객 유형 분석

RFM With Finda

4. 군집화 결과를 시각적으로 확인



0	153291
2	68581
1	59612
3	12032

떨어져 있는 데이터 포인트들은
서로 다른 군집으로 적절하게 군집화 됨을 시각적으로 확인함

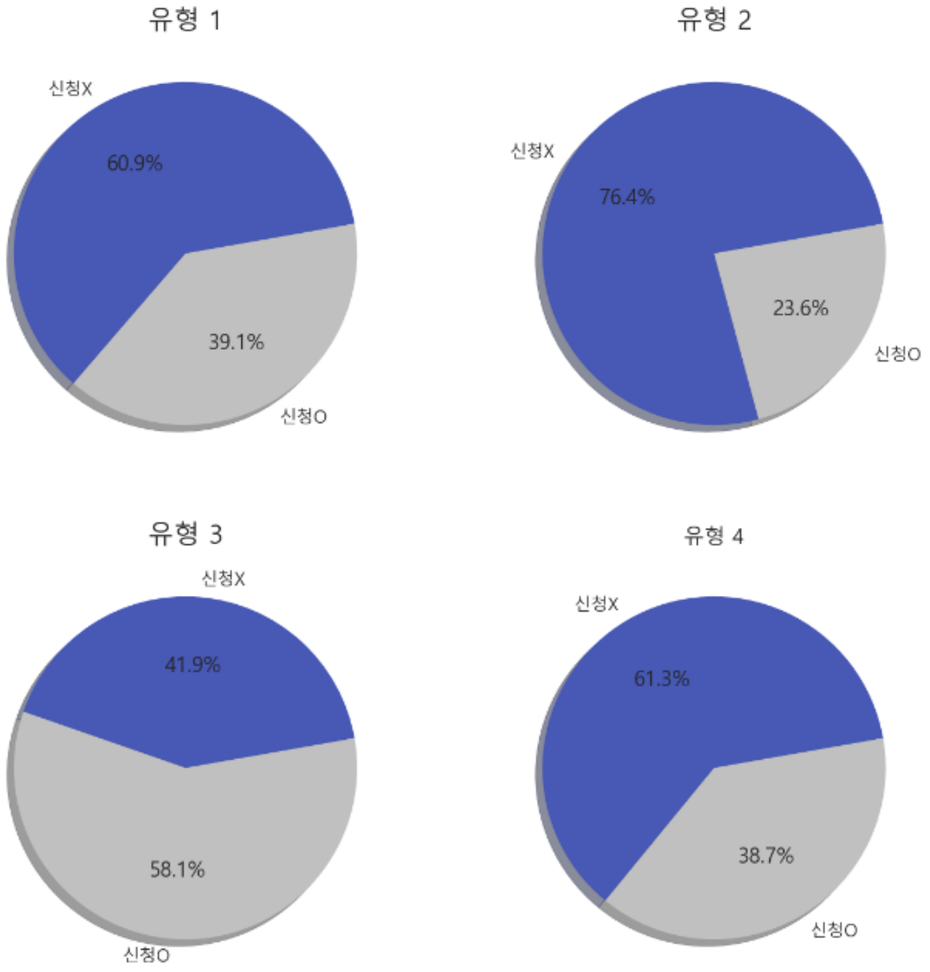
각 유형에 사용자들이 적절한 수치로 분포하고 있음을 확인함

03 고객 유형 분석

고객 유형 분석

1. 유형별 특징 분석

유형별 대출 신청 이력 분포



“ 유형3을 제외한 모든 유형은
대출 이력이 없는 사용자의 수가 많음
특히, **유형2**의 경우 그 특징이 두드러짐
유형3은 신청 이력이 있는 고객이 더 많음 ”

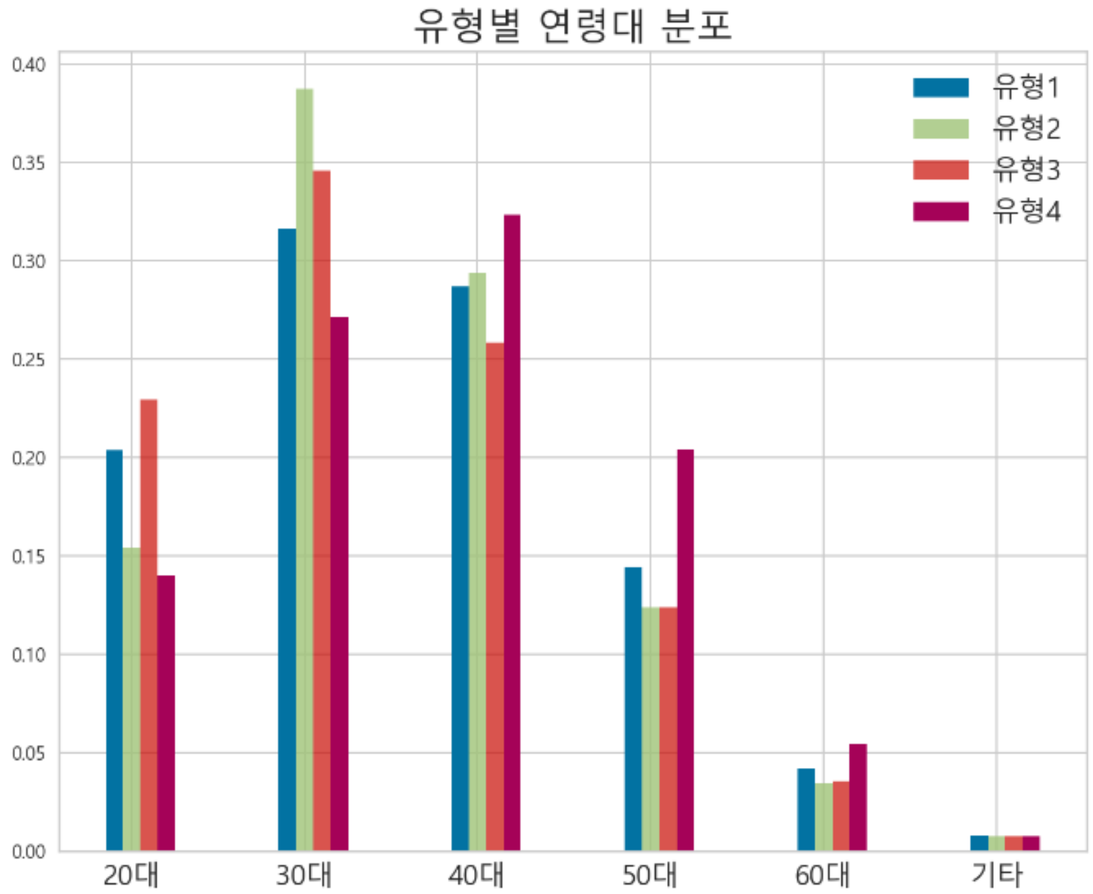
추가적인 분석을 통해 **대출 신청을**
유도할 수 있는 서비스 메시지 제안이 필요함

03 고객 유형 분석

고객 유형 분석

1. 유형별 특징 분석

연령대



“ 유형2. 30대의 분포 많음
유형 1.3. 비교적 20대의 분포가 많음
유형4는 40. 50대의 비율이 높음 ”

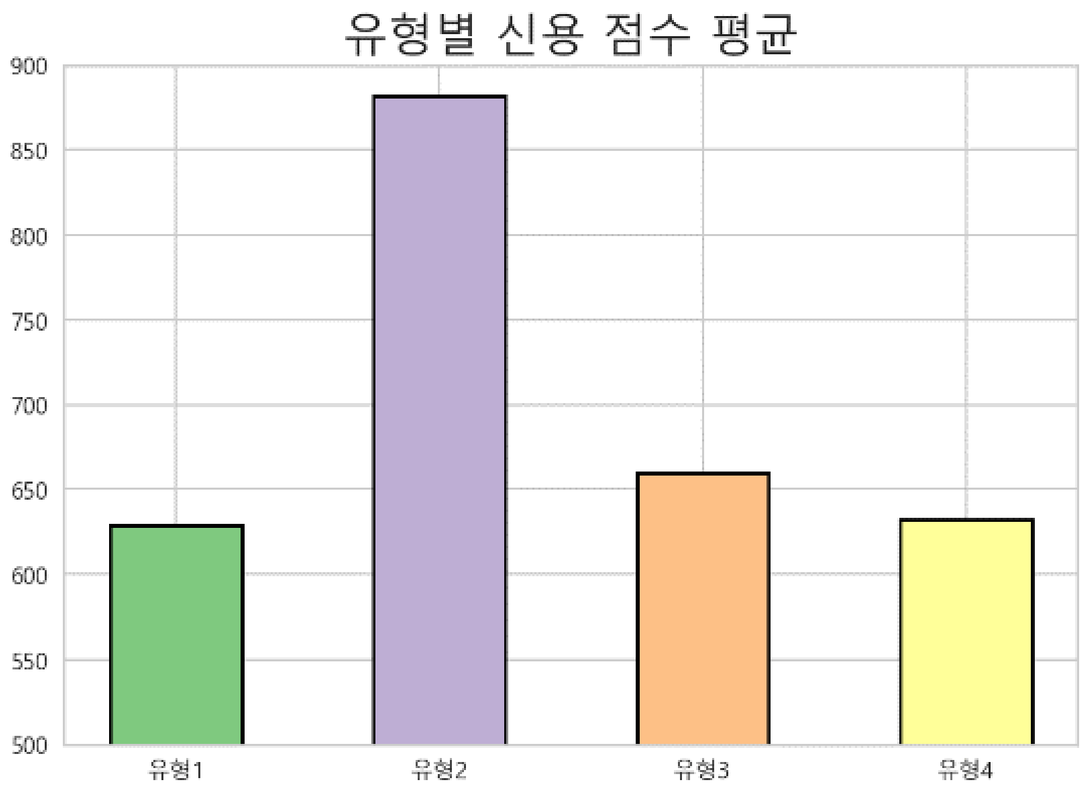
각 유형에 많이 분포하는 연령대별로 각기 다른 서비스 메시지 제안을 할 수 있음

03 고객 유형 분석

고객 유형 분석

1. 유형별 특징 분석

신용점수



“ 유형2의 경우 다른 유형들과
두드러지는 특징을 보임
유형별 신용 점수의 평균 값이 매우 큼 ”

신용 점수가 높기 때문에 **대출에 이점**이 있음
즉, 이를 활용한 서비스 메시지 제안이 필요

03 고객 유형 분석

고객 유형 분석

1. 유형별 특징 분석

유형별 대출 목적

보편적으로 많이 분포하는 '생활비', '대환대출'을
제외하고 각 유형에서 많은 수를 차지하는
대출 신청 목적을 살펴봄

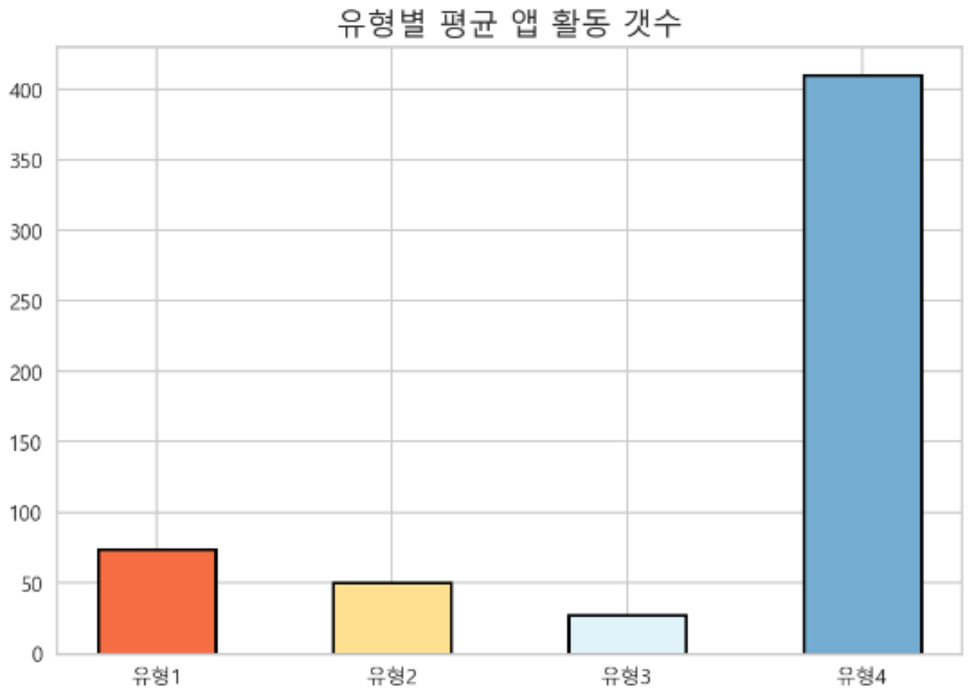
타 유형과 다른 특징을 가지는 대출 목적을 확인하고
이에 따라 서비스 메시지를 제안하는 것이 바람직함

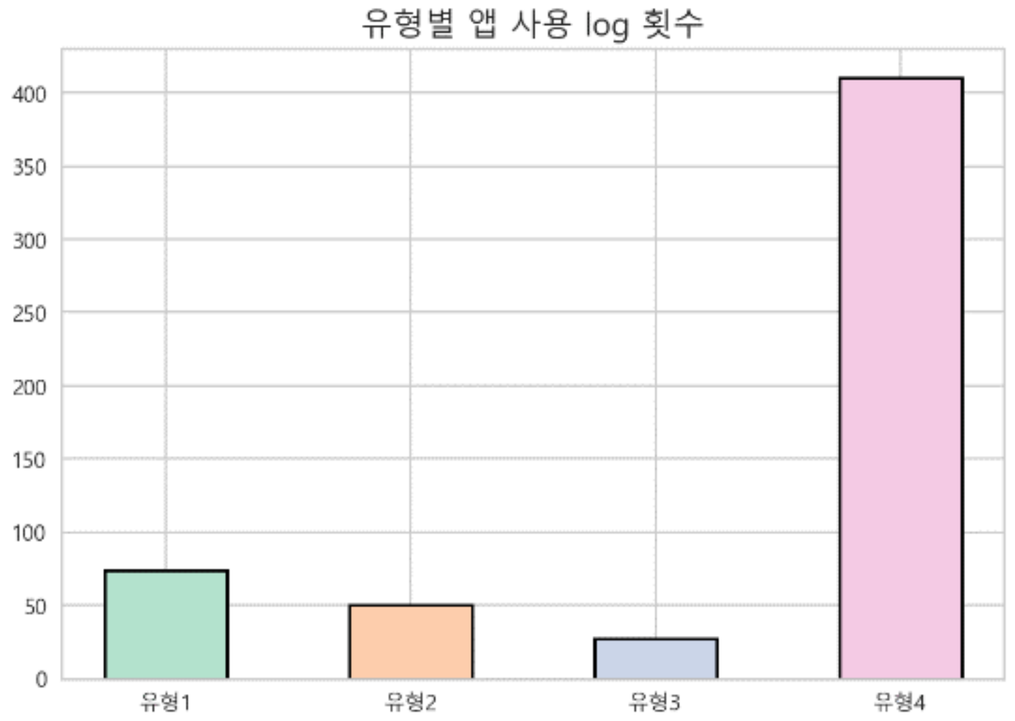
03 고객 유형 분석

고객 유형 분석

1. 유형별 특징 분석

Log횟수





“ 유형4의 경우 다른 유형과는 다르게 앱 사용량이 매우 많음을 알 수 있음 ”

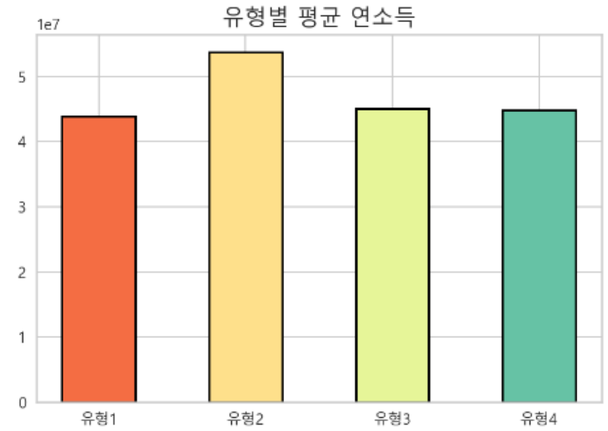
앱에서 활동이 가장 많음 즉, 대출에 관심이 많음을 알 수 있음

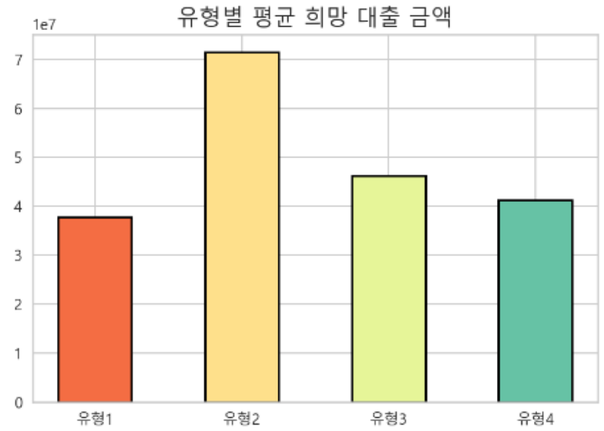
03 고객 유형 분석

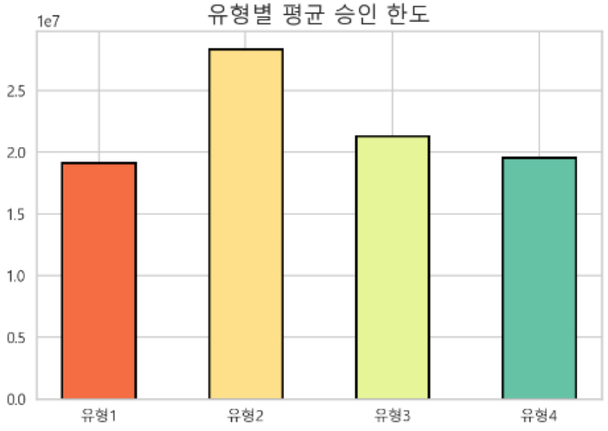
고객 유형 분석

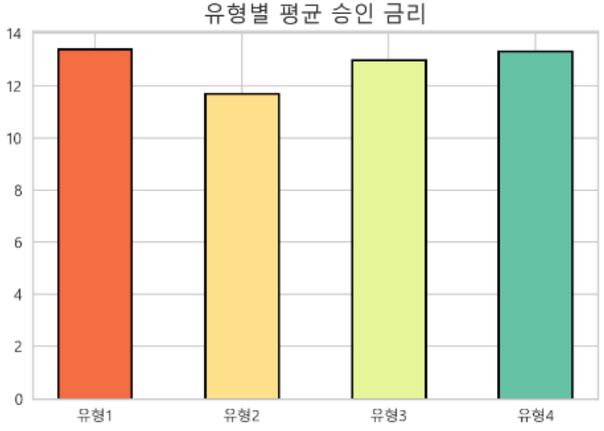
1. 유형별 특징 분석

여러 정보를 바탕으로 분석









“여러 그래프를 바탕으로 보았을 때,”

유형2의 경우 비교적 자본이 많은 유형임을 알 수 있음

또한, **유형1**의 경우는 유형2와 반대로 **자본이 적은** 유형임을 알 수 있음

해당 결과를 바탕으로 finda에서 이용할 수 있는

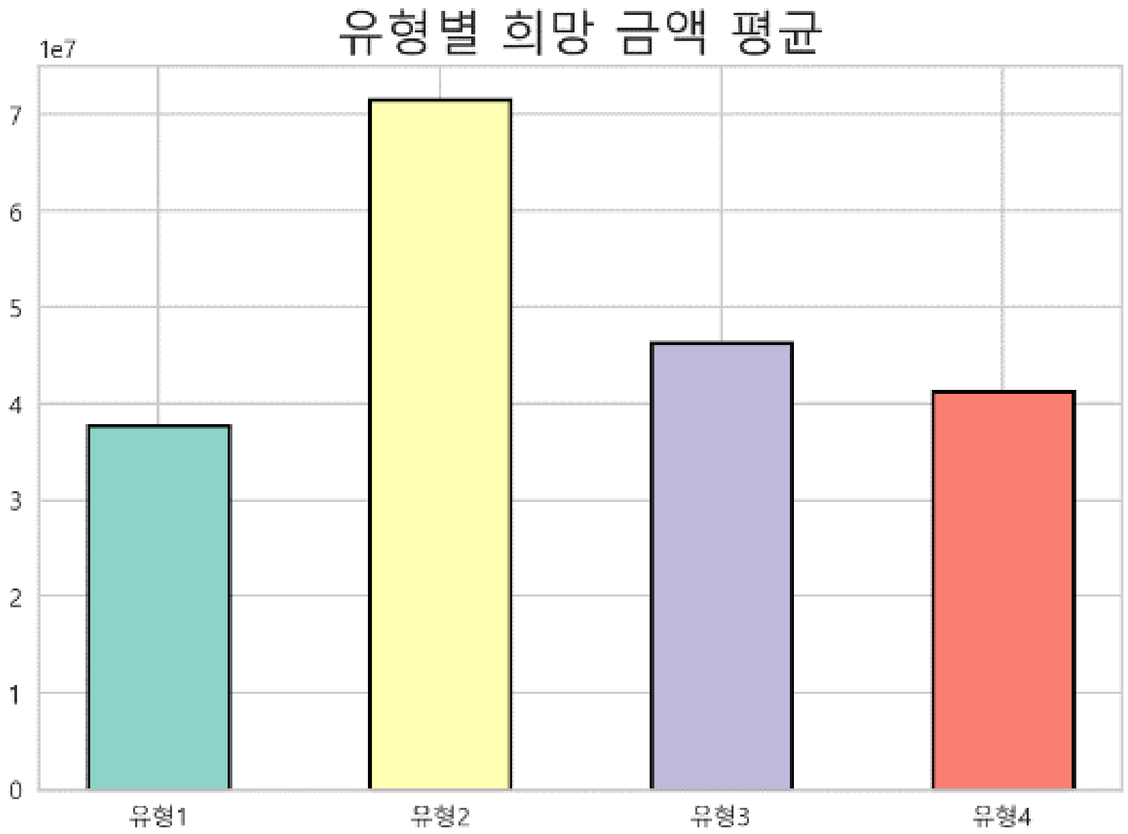
다양한 서비스에 대한 제안이 필요함

03 고객 유형 분석

고객 유형 분석

1. 유형별 특징 분석

희망 대출 금액



“

이전 분석의 연장으로

유형1의 경우는 희망 대출 금액이 낮고,

유형2의 경우는 높음을 알 수 있음

”

희망 금액에 따라

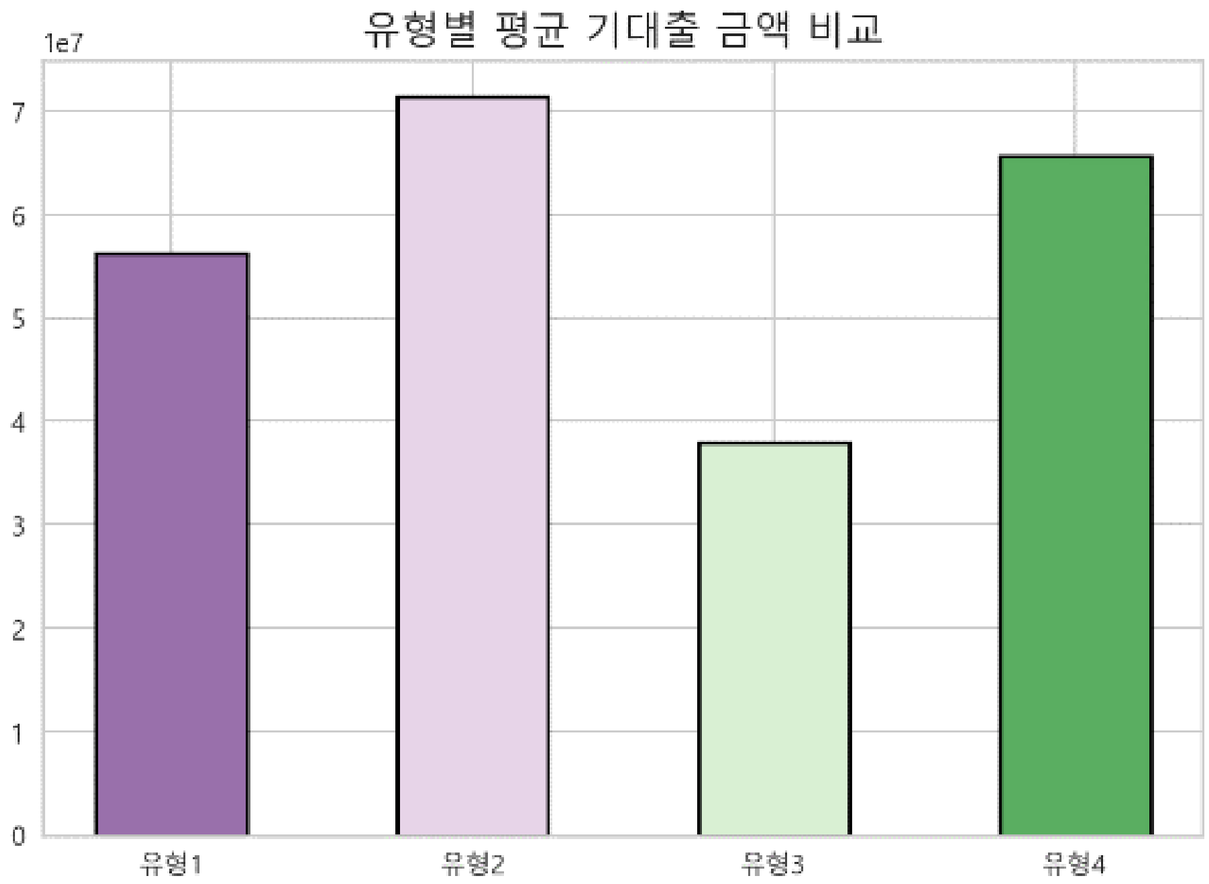
효과적인 서비스 메시지 제안이 가능함

03 고객 유형 분석

고객 유형 분석

1. 유형별 특징 분석

기대출금액



“유형2의 경우 자본이 많지만 기대출 금액 또한 많음
유형3의 경우 기대출 금액이 가장 낮음
즉, 유형2는 대출의 여유가 없지만
유형3은 대출에 여유가 있음”

유형3의 경우 대출 이력이 있는 고객들이
과반수 이상이었음, 즉, finda에 대한 홍보를 한다면
앱을 사용하여 대출을 신청할 확률이 높음

서비스 메시지 제안

유형1

대출 이력 없는 고객 약 60%
비교적 20대 분포 많음
신용점수 비교적 낮음
비교적 자본이 적은 고객
기대출 금액 평균
대출 희망 금액 낮음

유형2

대출 이력 없는 고객 약 75%
비교적 30대 분포 많음
신용점수 비교적 높음
비교적 자본이 많은 고객
기대출 금액 많음
대출 희망 금액 높음

유형3

유일하게 대출 이력 고객 과반수
20대, 30대 젊은층 많음
Finda 앱 사용량 적음
기대출 금액 낮음
대출에 여유가 있음

유형4

대출 이력 없는 고객 약 60%
40,50대 연령층 많음
신용점수 비교적 낮음
앱 활동량 매우 많음
대출 관심 많음

서비스 메시지 제안

유형1

대출 이력 없는 고객 약 60%
비교적 20대 분포 많음
신용점수 비교적 낮음
비교적 자본이 적은 고객
기대출 금액 평균
대출 희망 금액 낮음

유형2

대출 이력 없는 고객 약 75%
비교적 30대 분포 많음
신용점수 비교적 낮음
비교적 자본이 많은 고객
기대출 금액 많음
대출 희망 금액 높음

유형3

유일하게 대출 이력 고객 과반수
20대, 30대 젊은층 많음
신용점수 비교적 낮음
기대출 금액 낮음
대출에 여유가 있음

유형4

대출 이력 없는 고객 약 60%
40,50대 연령층 많음
신용점수 비교적 낮음
앱 활동량 매우 많음
대출 관심 많음

부족한 생활비 저금리로 대출받자!

지금 핀다에서 대출 받으면 조건 없이

최대 4.4%까지 우대금리 적용가능!

핀다에서 우대 금리로 대출 받으면 어깨 finda!

서비스 메시지 제안

유형2

대출 이력 없는 고객 약 75%
비교적 30대 분포 많음
신용점수 비교적 높음
비교적 자본이 많은 고객
기대출 금액 많음
대출 희망 금액 높음

유형1

대출 이력 없는 고객 약 60%
비교적 20대 분포 많음
신용점수 비교적 낮음
비교적 자본이 적은 고객
대출 희망 금액 낮음

유형3

유일하게 대출 이력 고객 과반수
Finda 앱 사용량 적음
기대출 금액 낮음

유형4

대출 이력 없는 고객 약 60%
신용점수 비교적 낮음
앱 활동량 매우 많음

인기 차종 사전예약부터 할부신청까지 한번에!

본인 인증만 하면 여윳돈 관리부터 대출관리까지!

핀다에게 대출 관리 맡기면 걱정 finda!

서비스 메시지 제안

유형3

유일하게 대출 이력 고객 과반수
20대, 30대 젊은층 많음
Finda 앱 사용량 적음
기대출 금액 낮음
대출에 여유가 있음

유형1

대출 이력 없는 고객 약 60%
비교적 20대 분포 많음
신용점수 비교적 낮음
기대출 금액 평균

유형2

대출 이력 없는 고객 약 75%
비교적 30대 분포 많음
신용점수 비교적 높음
기대출 금액 많음

유형4

대출 이력 없는 고객 약 60%
40,50대 연령층 많음
신용점수 비교적 낮음
대출 관심 많음

1분만에 대출조건 비교하고 5분내로 대출금 입금까지!

당신에게 딱 맞는 대출 받고 근심 finda!

서비스 메시지 제안

유형4

대출 이력 없는 고객 약 60%
40,50대 연령층 많음
신용점수 비교적 낮음
앱 활동량 매우 많음
대출 관심 많음

유형1

대출 이력 없는 고객 약 60%
비교적 자본이 적은 고객

유형2

대출 이력 없는 고객 약 75%
비교적 자본이 많은 고객

유형3

유일하게 대출 이력 고객 과반수
기대출 금액 낮음
대출에 여유가 있음

기존 대출보다 더 낮은 금리의 상품이 존재해요.

62개 금융기관, 200여개 대출상품 비교하고

더 좋은 조건으로 갈아타면 얼굴 finda!

Thank you!

세상에 없던
대출 비교 플랫폼

finda