[신한금융그룹 빅데이터 해커톤]

신한카드 ESG로 미래세대와 소통하다

> 2022.09.15. (목) - 2022.10.07. (금)
> 신한금융그룹 빅데이터 해커톤 데이터 분석 트랙
> 우수상 수상
>
> 팀 숙데콘(숙명데이터유니콘)
> 강수연, 김세은, 임다미

📂 분석 배경과 목표

▼ 미래세대 정의

"MZ 세대 뿐 아니라 다가올 미래에 적응하고 새로운 삶을 실현해 나갈 모든 세대"

▼ 미래세대를 위한 사업 방향성과 신한카드의 실용적 문제 해결을 위한 솔루션 고안

- 다가올 미래에 적응하는 미래세대에 발맞춰 미래 세대와 공생관계에 있는 기업 또한 미래를 대비하는 경영, 지속가능한 경영을 지향해야한다고 판단함. 따라서 ESG 지

수를 높일 수 있는 사업을 고안함.
⇒ 내부통제를 위한 고객 신용 리스크 예측 서비스

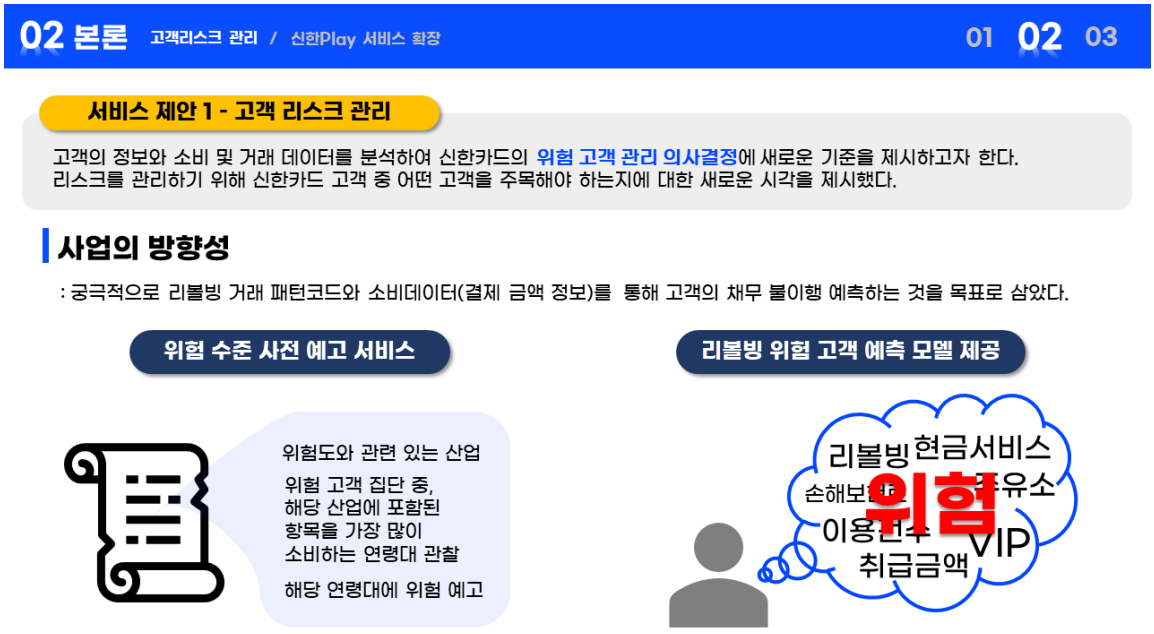
- 신한 Play를 직접 이용하며 느꼈던 아쉬운 점을 보완하기 위한 실용적인 솔루션을 고안함.
⇒ 플랫폼 통합을 통한 고객 맞춤형 서비스

📂 사업 요약

"알파를 플레이하다"

▼ 비용 절감 측면의 고객 신용 리스크 관리 서비스 제안

💡 데이터 기반의 내부통제 사업 기획



💡 기대효과

고객 유형별 신용리스크 시그널 관리 효과	더욱 정교한 미래 대손 규모 예측 효과
예상위험 고객에게 향후 채무불이행 가능성을 사전에 예고함.	빅데이터를 이용하여 사업 보고서 및 지속가능경영보고서 작성 과정에서 리볼빙 채무불이행 고객을 예측함.
신용카드사 고객이 경험하는 정보의 비대칭성을 해소함.	보다 정확한 대손규모를 적립함으로써 비용을 절감함.
ESG 중 사회공헌 지수(Social sector)를 높일 수 있음.	ESG 중 지배구조 지수(Governance sector)를 높일 수 있음.

▼ 새로운 가치 창출을 위한 신한 Play 확장 서비스 제안

💡데이터 기반의 플랫폼 통합 사업 기획

02 본론 고객리스크 관리 / 신한Play 서비스 확장 01 02 03

서비스 제안 2 - 신한 Play 서비스 확장

고객의 소비 및 거래 데이터를 기반으로 신한 Play 내에서 이용 가능한 기능을 확대하여 신규 고객 유치와 기존 고객 유지를 위한 서비스를 제안하고자 한다. 더불어 ESG 플랫폼과의 연동을 통해 ESG로의 소비를 유도할 수 있는 방법을 제시했다.

사업의 방향성

: 궁극적으로, 분산되어 있는 신한 플랫폼을 '신한Play' 내에서 한 번에 이용할 수 있도록 하여 앱 활용도를 높이는 것을 목표로 삼았다.

소비 유도

Hmall 롯데ON 신세계몰 emart 11번가

고객 맞춤형 제휴 쇼핑몰 추천

소비성향에 맞는 제휴 플랫폼 추천, 연동

Shinhan My Car / Allthat

새로운 가치 창출

신한 Play — 소비 데이터 → 펀드 추천

신한카드 데이터를 그룹사로 확장하여 새로운 가치를 창출
소비자 → 투자자 전환 기회 제공

고객의 소비 유형에 따라 개별 고객의 소비 패턴을 발견한다.

각 소비 패턴에서 고객의 니즈를 파악하고 그를 충족시킬 수 있는 맞춤형 서비스를 제안한다.

* 맞춤형 서비스 : 개인화한 고객의 소비 패턴에 맞는 제휴 쇼핑몰 추천, 펀드 추천, 신한 플랫폼 추천 서비스

신한 금융 지주사의 모든 플랫폼 통합

- 궁극적으로 신한 PLAY를 중심으로 신한 *올댓 쇼핑, My Car, 신한 알파, 신한라이프* 등의 **신한 금융 지주사의 모든 플랫폼을 통합**한다. - 신한카드에서 **모든 신한 금융 계열사로 서비스를 확장**한다.

💡 기대효과

앱 활용도 향상을 통한 이익 창출	미래 세대 고객의 결제 유치 강화	핀테크 기업과 차별화된 경쟁력
분산되어 있던 플랫폼 간의 연결성을 높이면서 앱의 활용도를 향상시킴으로써 고객의 앱 이용 시간을 늘리고 이익 창출까지의 과정을 유도할 수 있다.	편익 소비, ESG 이슈 관련 가치 소비 등 미래 세대의 가치관 변화에 대한 신용카드업을 영위하는 금융 회사의 결제 유치를 강화할 수 있다. ESG 산업 소비를 선호하는 미래 세대에게 선택의 폭을 확장해줌으로써 친환경(E) 소비를 유도할 수 있다.	신한 그룹의 그룹사를 활용하여, 다양한 금융 서비스를 제공할 수 있다. 플랫폼의 확장으로 그룹사 간의 연동을 가능하게 함으로써 핀테크 기업이 가지고 있지 않은 경쟁력을 확보할 수 있을 것이다.

📂 분석 요약

카이제곱 독립성 검정	연령대 변수가 신용 위험수준 변수에 영향을 미친다는 가설 검정
이항 로지스틱 회귀분석	고객의 신용 위험도 관련된 산업 변수 추출
의사결정 분류나무	신용카드 리볼빙* 대금의 채무불이행 고객의 특성 예측 모델링 *) 카드대금 일부 이월결제약정
K-means 군집분석	신용카드 이용건수 및 결제금액에 따른 소비성향이 유사한 고객 군집 8가지 형성 (VIP 여부에 따른 사전 분류)
Apriori 연관규칙분석	8가지 고객 군집별 26개의 결제지출이 있는 산업 중 연관규칙이 있는 산업 집합 추출

📂 의사결정 분류나무 기반의 신용 리스크 발생 고객의 특성 예측

카드이월대금을 상환하지 않는다고 예측된 고객 특성	취급금액, 이용건수, 신한카드 VIP 여부 및 주유소, 손해보험료 결제금액이 일정수준보다 높은 고객
리볼빙을 연간 3회 이상 연속으로 이용하며, 1회당 100만 원 이상 이용하는 고객의 거래패턴	월 1회 이상 현금서비스 (카드단기대출) 이용하는 경향 장기론 (카드장기대출) 이용률 저조, 할부금융 서비스 이용률 전무
신용 리스크 관리 서비스 기대효과	사업 보고서 및 지속가능경영(ESG) 보고서 작성 시 빅데이터를 이용한 리볼빙 채무불이행에 대한 예측 수행 보다 정확한 대손 충당금 규모를 설정하고 적립함으로써 비용을 절감하고 수익을 증가시키는 효과 주주, 채권자, PEF 등 이해관계자에 대하여 정확한 정보를 제공하여 지배구조(G sector) 지수를 향상

📂 연관규칙분석 기반의 산업간 집합 분류

군집분석 및 연관규칙분석 기반의 서비스	신한 Play에서 고객의 소비 패턴에 맞는 **신한 플랫폼과 제휴쇼핑몰**을 추천하여 분산되어 있던 앱 연결성과 앱 활용도를 높이는 서비스 기획한 신사업에 대한 예시 하단에 기재
VIP 집단 중 샌드위치 유형 (이용건수와 취급금액이 모두 중간인 군집) (외출 시 외식을 자주 하는 집단으로 판단)	**- 소비유도** 신한카드 종합생활금융 플랫폼, 신한 Play의 프리미엄 쿠폰 추천 제휴사와의 협업을 통해 제휴사 추천 예) VIP 집단으로 다이닝 위주 레스토랑 추천 **- 투자유도** 문화예술, 외식 산업을 영위하는 회사의 주식 종목이 포함된 펀드 추천 → **신한투자증권 알파 앱**으로 연동 예) 신한 K-컬쳐 증권 자투자신탁 (K-이커머스, 헬스케어, 미디어)
신한카드 결제플랫폼 (신한 Play) 확장 서비스 기대효과	고객의 라이프사이클 내 점유시간을 늘리고 이익 창출까지의 과정을 유도 ESG 산업 관련 가치 소비를 선호하는 미래세대의 선택의 폭을 확장하여 친환경 소비를 유도하고 환경 (E sector) 지수를 향상

목차

1. 데이터 수집과 전처리

📂 데이터 수집

데이터 명	출처	내용
data_007.csv	신한카드	- 총 473,225명의 고객 데이터 - 고객정보 7개 변수 - 결제정보 168개 변수 - 거래패턴코드 6개 변수 - 세부 사항은 보안상 공개불가

📂 데이터 전처리

▼ 데이터 결측값 및 이상치 처리

- 결측값 처리 기준

  모르는 값(NULL)이 아니라 고객이 해당 소비항목에 대한 지출을 하지 않았다는 의미로 결측값을 제거하거나 대체하지 않음.

- 이상치 처리 기준

  이상치라고 판정할 수 있는 수준의 결제금액이 높은 고객의 소비지출 데이터를 제거하는 것은 기획의 취지에 맞지 않기 때문에 제거하거나 대체하지 않음.

- 해결과정

  각 소비지출항목의 결제금액을 동일 산업별로 합산함. 즉, 산업코드 변수 26개 생성하여 결측값(0)을 최소화함.

▼ 파생변수 생성

- 신용카드사의 여신금융서비스 이용 여부에 대한 패턴코드의 파생변수를 생성한 이유

  연간 월별 거래 서비스 이용 여부(0, 1)에 따른 요약 변수 형태의 원본 데이터를 분석 목적*과 편의를 위해 새로운 **거래 위험등급 변수** 파생함.
  *)신용 리스크 있는 고객 관리 목적

- 해결과정

  각 패턴코드 변량(value)의 1(서비스 이용한 경우)의 개수를 추출하는 방식을 고안함. 카드 리볼빙 서비스를 12개월 중 3개월 이상 연속으로 이용한 고객을 초고위험등급으로 설정함.

  - 거래 이용횟수 연속적으로 연 3회 이상 : HH등급(high risk high level)

  - 거래 이용횟수 불연속적으로 연 3회 이상 : HL등급(high risk low level)

  - 거래 이용횟수 연 1회 ~ 2회 : L등급(low risk)

  - 거래 이용횟수 연 0회 : N등급(no risk)

  카드 리볼빙은 익월로 대금을 이월하더라도 최소결제금액을 납부하지 않을 경우 연체로 처리되고 신용점수가 내려갈 수 있음.

  90일 연체 시 신용등급이 8 ~ 9등급으로 하락하고 이른바 신용불량자*로 확정됨. 금융거래가 제한된다는 점에서 상기된 기준으로 거래 위험등급 파생변수 생성함.

  *) 금융거래 등 상거래에서 발생한 대금 또는 대출금 등의 채무에 대하여 정당한 사유 없이 약정된 기일 내에 변제를 이행하지 아니한 자

2. 고객 신용 리스크 관리 서비스

📂 사업 개요

- 사업 기획 배경

  리볼빙이란 지급할 카드대금을 일부만 납부하고, 20% 내외의 높은 수수료를 조건으로 연체처리 없이 익월로 이월하는 신용카드사의 여신금융서비스를 말함.

  리볼빙 서비스 대상고객은 소액의 급전이 필요한 저신용자이자 상품 특성상 대출심사 또는 승인기간을 요하지 않는 불황형 대출로 잠재 부실위험이 존재함.

  따라서 리볼빙 부실화에 대비해 신용카드사에서 대손충당금 적립액을 더 쌓아야 하는 부담을 완화하는 솔루션을 제시함.

- 사업의 방향성

  카드 돌려막는 신용 리스크 고객 색출하는 머신러닝 알고리즘 (regression & tree)

📂 사업 내용

고객 신용 리스크 관리 서비스

1. **리볼빙 이용이 많은 고객 특성 분석**

   📌 분석결과

   이항 로지스틱 회귀분석 결과 신용 위험등급이 높아질 때 결제금액이 증가함에 따라 양의 상관관계가 있는 산업을 추출함.

   📌 인사이트

   - 이동 산업, 전자통신 산업에 대한 결제금액이 증가할수록 리볼빙 위험등급이 L(low)에서 HL(high risk low level)로 높아짐.

   - 고정지출 관련 산업, 고가품 산업, 가스 산업에 대한 결제금액이 증가할수록 리볼빙 위험등급이 HL(high risk low level)에서 HH(high risk high level)로 높아짐.

📌 사업 연결

- 금융소비자보호법 제19조 제1항에 의거한 카드대출상품 조건 및 리볼빙 서비스 금리와 대출한도를 비교하는 고지함으로써 설명의무를 이행하는 과정에서 사업을 제안함.

  잠재적 신용 리스크가 있는 연령대의 고객에게 본인의 위험수준과 이를 높이는 소비지출항목이 있는 산업유형을 안내함.

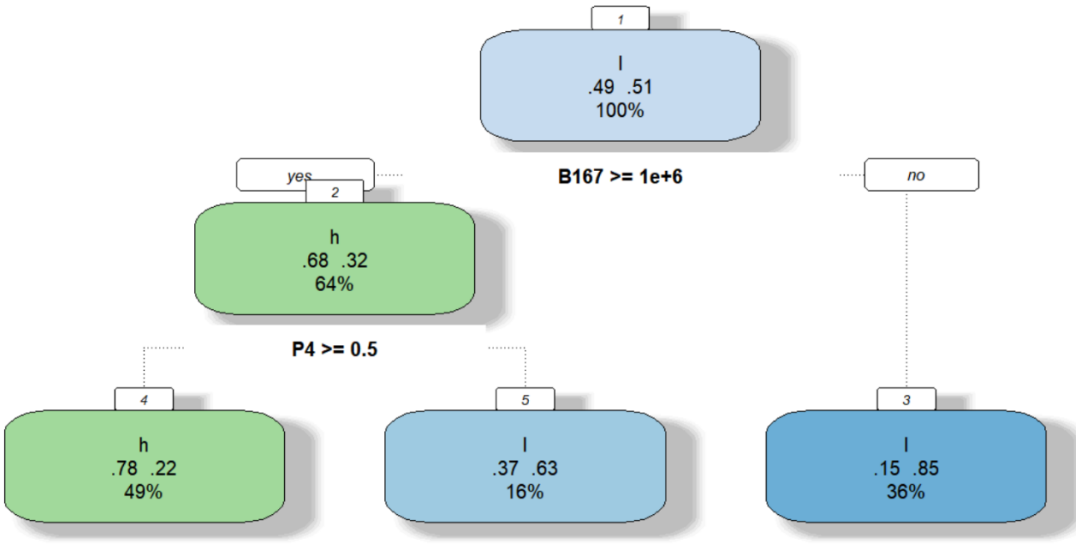
  향후 채무불이행 가능성을 사전 예고함으로써 고객이 신용카드사 대비 부족한 데이터로 인하여 경험하는 정보의 비대칭성을 완화함.

산 업	사전 예고 서비스 대상
C5 (고정지출)	60대 후반(세금지출), 60대 초반(생명보험), 20대 후반(손해보험)
C18 (고가품)	50대 초반(고가 시계) , 60대 초반(귀금속), 50대 후반(악세사리)
C27 (가스)	50대 초반(주유소), 60대 후반(LPG가스), 30대 후반(가정용 연료)
C4 (이동)	20대 후반(택시), 50대 초반(차량용품), 60대 후반(오토바이), 50대 후반(자전거)
C16 (전자통신)	20대 후반(통신요금_이동시내전화) , 40대 후반(PC통신), 30대 초반(TV유선)

2. 연체 위험 고객 예측 모델

📌 분석 결과

- 리볼빙 초고위험등급(HH) 고객의 대금상환 여부를 예측함.
  고객의 카드이월대금 연체에 따른 신용 리스크 가능성을 high 등급과 low 등급으로 구분.



📌 인사이트

- 예측된 카드이월대금을 상환하지 않을 고객의 특징을 찾아냄.

  신용카드 결제금액, 이용건수, 카드사 VIP 여부, 주유소 결제금액, 손해보험료가 일정 수준보다 높은 고객

  월 1회 이상 현금서비스 즉, 카드단기대출을 이용하는 경향이 있는 고객

  연령대에서 높은 비중 차지한 40대 초반, 40대 후반, 50대 초반대의 고객

📌 사업 연결

- 잠재 부실위험이 존재하는 고객의 특징과 신용 리스크를 기업의 내부통제 차원에서 관리

  머신러닝 모델링 결과 연체 예측된 신용 리스크 고위험 고객의 이월대금의 일정 비율을 대손충당금으로 적립하여 이해관계자에게 ESG 보고서에 공시하는 사업을 제안함.

  신용 리스크 고위험 고객의 카드단기대출 시 승인기간 또는 심사절차를 마련하는 사업을 제안함.

📂 분석과정

▼ 신용 리스크가 있는 리볼빙 서비스 이용 고객의
연령대별 산업항목에 대한 분석

▼ 이원 카이제곱 독립성 검정 (교차분석)

결과

p-value < 0.05, Ha 채택

→ **연령과 위험등급이 관련이 있다**고 할 수 있다

1. 변수 선택 및 가설 수립

   범주형 종속변수 : 리볼빙 위험등급(N, L, HL, HH)

   범주형 독립변수 : 연령대(20대 초반 ~ 60대 후반)

   귀무가설(H0) : 연령대와 리볼빙 위험등급은 상호 독립적이다.

   대립가설(Ha) : 연령대와 리볼빙 위험등급은 상호 영향을 미칠 수 있는 연관성이 있다.

2. 교차 분할표(contingency table)를 작성하여 연령대 변수와 리볼빙 위험등급 변수의 빈도수를 정리함.

   [교차 분할표]

```
> xtabs(~ P2 + G1, data=sbh_H)
          G1
P2           HH    HL
  20대_초  1924    25
  20대_후  8661    76
  30대_초  9833    58
  30대_후 10296    45
  40대_초 13662    68
  40대_후 13010    59
  50대_초 12623    63
  50대_후  8843    44
  60대_초  6872    27
  60대_후  2695    20
```

   [비례 배분]

```
> prop.table(xtabs(~P2+G1, data=sbh_H), margin=1)
          G1
P2                  HH          HL
  20대_초 0.987172909 0.012827091
  20대_후 0.991301362 0.008698638
  30대_초 0.994136083 0.005863917
  30대_후 0.995648390 0.004351610
  40대_초 0.995047342 0.004952658
  40대_후 0.995485500 0.004514500
  50대_초 0.995033896 0.004966104
  50대_후 0.995048948 0.004951052
  60대_초 0.996086389 0.003913611
  60대_후 0.992633517 0.007366483
```

3. 결론적으로 **연령대는 리볼빙 위험등급에 통계적으로 유의한 영향을 미칠 수 있음.**

   카이제곱 분포의 값과 검정통계량 47.685를 비교한 결과 검정통계량이 더 크기 때문에 유의수준 a*보다 유의확률 p-value가 낮으므로 대립가설 채택함. 즉, 검정통계량보다 카이제곱 값이 극단값일 가능성은 10,000번 중 27번**임.

   *) 두 변수가 실제 무관하지만 연관성이 있다고 결론을 내릴 위험을 5% 수준에서 감수하는 결정을 말함.

   **) p-value = 2.925 / (epsilon의 7제곱) = 0.002667

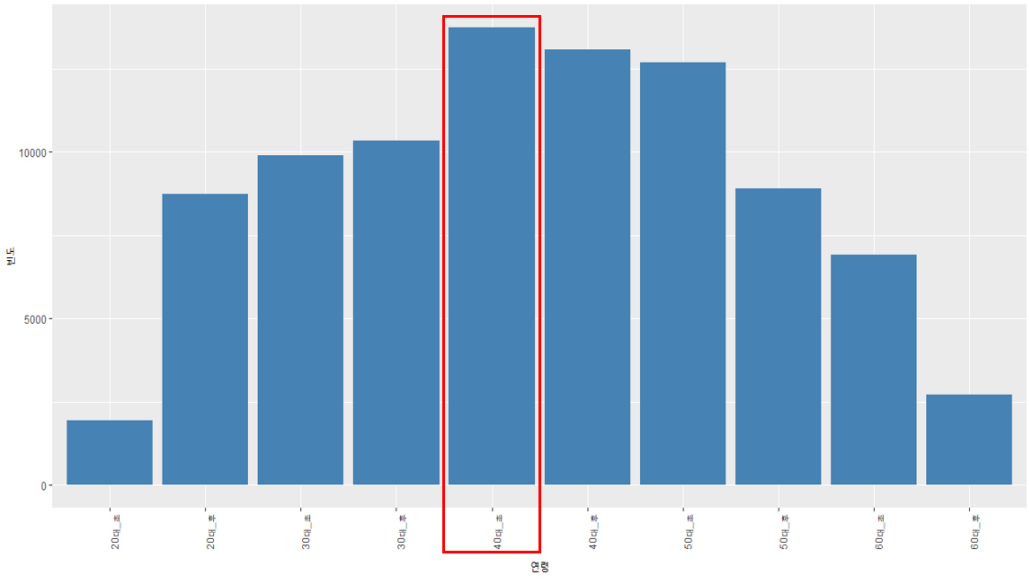
$$\chi^2_0 \geq X^2_\alpha[(m-1)(n-1)]$$

```
> chisq.test(xtabs(~P2+G1, data=sbh_H))

        Pearson's Chi-squared test

data:  xtabs(~P2 + G1, data = sbh_H)
X-squared = 47.685, df = 9, p-value = 2.925e-07
```

4. 리볼빙 서비스를 연간 3회 이상 이용하는 고객 중 신용카드사가 주목해야할 연령대 집단은 가장 높은 빈도의 **40대 초반 고객**으로 나타남.



▼ 이항 로지스틱 회귀분석과정

- 선형 대신에 시그모이드 곡선(S자형태의 곡선)을 이용한 회귀분석

1. 변수 선택 및 가설 수립

i. 이항 종속변수 : 리볼빙 위험등급 (HL, HH)

독립변수 : 산업항목 (26개)

귀무가설 : 특정 산업의 결제금액은 리볼빙 위험등급을 HH로 상승 [HL로 하락]과 무관하다.

대립가설 : 특정 산업의 결제금액은 리볼빙 위험등급을 HH로 상승 [HL로 하락]과 상관관계가 있다.

ii. 이항 종속변수 : 리볼빙 위험등급 (L, HL)

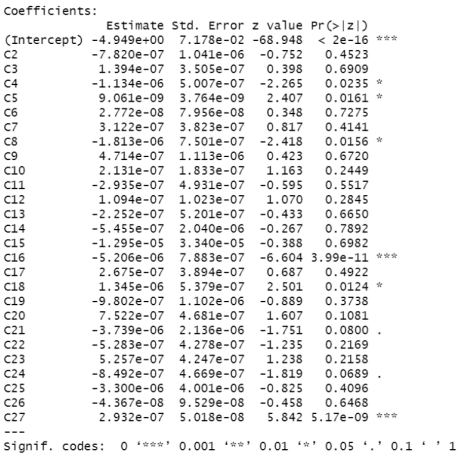
독립변수 : 산업항목 (26개)

귀무가설 : 특정 산업의 결제금액은 리볼빙 위험등급을 HL로 상승[L로 하락]과 무관하다.

대립가설 : 특정 산업의 결제금액은 리볼빙 위험등급을 HL로 상승[L로 하락]과 상관관계가 있다.

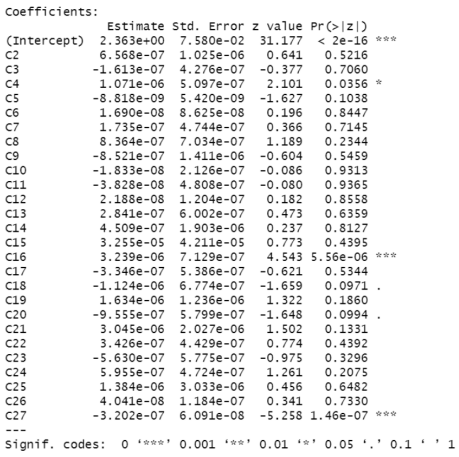
2. 이항 로지스틱 회귀분석 결과

[HL~HH등급으로 위험 높이는 산업]



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.949e+00  7.178e-02 -68.948  < 2e-16 ***
C2          -7.820e-07  1.041e-06  -0.752   0.4523
C3           1.394e-07  3.505e-07   0.398   0.6909
C4          -1.134e-06  5.007e-07  -2.265   0.0235 *
C5           9.061e-09  3.764e-09   2.407   0.0161 *
C6           2.772e-08  7.956e-08   0.348   0.7275
C7           3.122e-07  3.823e-07   0.817   0.4141
C8          -1.813e-06  7.501e-07  -2.418   0.0156 *
C9           4.714e-07  1.113e-06   0.423   0.6720
C10          2.131e-07  1.833e-07   1.163   0.2449
C11         -2.935e-07  4.931e-07  -0.595   0.5517
C12          1.094e-07  1.023e-07   1.070   0.2845
C13         -2.252e-07  5.201e-07  -0.433   0.6650
C14         -5.455e-07  2.040e-06  -0.267   0.7892
C15         -1.295e-05  3.340e-05  -0.388   0.6982
C16         -5.206e-06  7.883e-07  -6.604 3.99e-11 ***
C17          2.675e-07  3.894e-07   0.687   0.4922
C18          1.345e-06  5.379e-07   2.501   0.0124 *
C19         -9.802e-07  1.102e-06  -0.889   0.3738
C20          7.522e-07  4.681e-07   1.607   0.1081
C21         -3.739e-06  2.136e-06  -1.751   0.0800 .
C22         -5.283e-07  4.278e-07  -1.235   0.2169
C23          5.257e-07  4.247e-07   1.238   0.2158
C24         -8.492e-07  4.669e-07  -1.819   0.0689 .
C25         -3.300e-06  4.001e-06  -0.825   0.4096
C26         -4.367e-08  9.529e-08  -0.458   0.6468
C27          2.932e-07  5.018e-08   5.842 5.17e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

[L~HL등급으로 위험 높이는 산업]



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.363e+00  7.580e-02  31.177  < 2e-16 ***
C2           6.568e-07  1.025e-06   0.641   0.5216
C3          -1.613e-07  4.276e-07  -0.377   0.7060
C4           1.071e-06  5.097e-07   2.101   0.0356 *
C5          -8.818e-09  5.420e-09  -1.627   0.1038
C6           1.690e-08  8.625e-08   0.196   0.8447
C7           1.735e-07  4.744e-07   0.366   0.7145
C8           8.364e-07  7.034e-07   1.189   0.2344
C9          -8.521e-07  1.411e-06  -0.604   0.5459
C10         -1.833e-08  2.126e-07  -0.086   0.9313
C11         -3.828e-08  4.808e-07  -0.080   0.9365
C12          2.188e-08  1.204e-07   0.182   0.8558
C13          2.841e-07  6.002e-07   0.473   0.6359
C14          4.509e-07  1.903e-06   0.237   0.8127
C15          3.255e-05  4.211e-05   0.773   0.4395
C16          3.239e-06  7.129e-07   4.543 5.56e-06 ***
C17         -3.346e-07  5.386e-07  -0.621   0.5344
C18         -1.124e-06  6.774e-07  -1.659   0.0971 .
C19          1.634e-06  1.236e-06   1.322   0.1860
C20         -9.555e-07  5.799e-07  -1.648   0.0994 .
C21          3.045e-06  2.027e-06   1.502   0.1331
C22          3.426e-07  4.429e-07   0.774   0.4392
C23         -5.630e-07  5.775e-07  -0.975   0.3296
C24          5.955e-07  4.724e-07   1.261   0.2075
C25          1.384e-06  3.033e-06   0.456   0.6482
C26          4.041e-08  1.184e-07   0.341   0.7330
C27         -3.202e-07  6.091e-08  -5.258 1.46e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

i. 고정지출 관련 산업, 고가품 산업, 가스 산업에 대한 결제금액이 증가할수록
리볼빙 위험등급이 HL(high risk low level)에서 HH(high risk high level)로 높아짐. (양의 상관관계)

ii. HL ~ HH등급의 고객 집단에서 이동 산업, 전자통신 산업에 대한 결제금액과 리볼빙 위험등급이 음의 상관관계가 있다는 결과를 재검정함.

iii. L ~ HL등급의 고객 집단에서 가설 검정한 결과 다음과 같은 결론 도출함.

이동 산업, 전자통신 산업에 대한 결제금액이 증가할수록 리볼빙 위험등급이 L(low)에서 HL(high risk low level)로 높아짐. (양의 상관관계)
단, 해당 산업은 리볼빙 위험등급을 HH까지 높이지 않는다는 점에 유의함.

3. 리볼빙 위험등급을 높이는 산업에서 소비지출 빈도가 높은 고객 연령대 확인

각 산업에 속한 세부 지출항목에 소비한 금액의 평균을 연령대별로 비교함.

산업별로 신용카드사에서 주목해야 할 연령층에 대한 시각화로 탐색적 데이터 분석.

▼ 리볼빙 채무불이행 위험 고객 예측 모델 구현

▼ 의사 결정 분류 나무

- **의사결정 분류나무 기반의 채무불이행 고객 예측 분석과정**

1. 리볼빙 초고위험등급(HH) 고객*의 대금상환 여부를 예측하기 위한 종속변수를 새롭게 파생함. 연간 월별 100만 원 이상 리볼빙을 이용한 횟수가 **5회 이하인 경우** 및 **5회 초과 12회 이하인 경우**로 기준을 수립함. 5회를 기준으로 한 이유는 중간값이 5로 클래스 불균형 문제를 사전 방지하기 위함.

   이용금액 100만 원 넘는 경우 연 5회 이하 : l등급(low default)

   이용금액 100만 원 넘는 경우 연 5회 초과 12회 이하 : h등급(high default)

   *) 거래 이용횟수 연속적으로 연 3회 이상 : HH등급(high risk high level)

2. 시드 설정하고, 훈련용 및 검증용 데이터 셋을 7:3으로 분리함. 단, 모형 적합(fitting)을 위한 추후 성능 평가 시 3가지 데이터 셋 분리기준으로 모델링, 9회 시드 고정을 풀고 모델링 수행함.

3. 의사결정 분류나무 모형에 들어가는 설명변수 (하이퍼파라미터 튜닝)

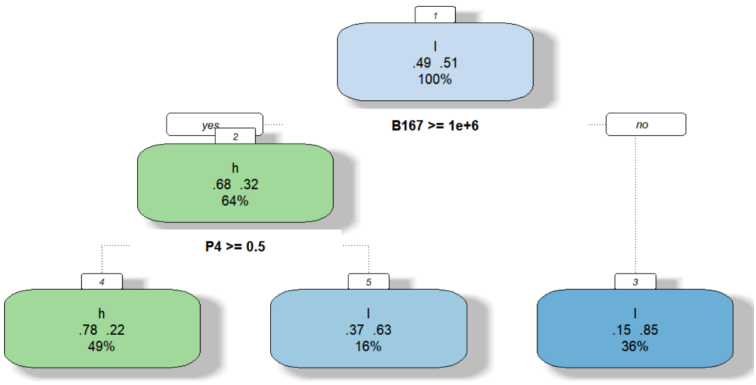
   리볼빙 100만 원 이용등급, 연령, 카드VIP 여부, 금융투자활동 여부, 결제금액,
   이용건수 및

   위 가설검정에서 추출한 위험수준과 양의 상관관계가 있는 산업 내 지출 항목
   (세금공과금, 생명보험료, 손해보험료, 고가 시계, 악세사리, 금, 주유소, 가스, 유류도매, 가정용 가스)

4. xerror(오차의 추정값 기준)가 가장 낮은 깊이 3에서 트리의 노드 분리 종료함.

- **의사결정 분류나무 분석결과**

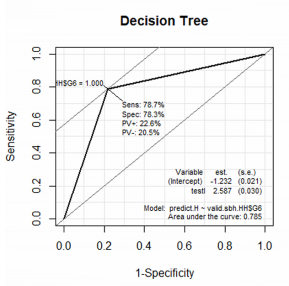
  초고위험등급 고객 62,232명 중 채무불이행(연체, 상환 불가능) 고객의 특성을 예측.

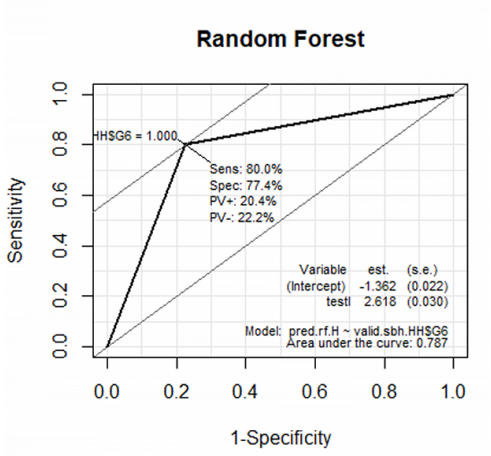


[변수 중요도와 노드 분리(split) 기준에 따라 예측된 고객의 특성]

1. 신용카드 결제금액, 이용건수, 카드사 VIP 여부, 주유소 결제금액, 손해보험료가 일정 수준보다 높은 고객
2. 월 1회 이상 현금서비스 (카드단기대출) 이용하는 경향이 있었으며, 장기론 (카드장기대출) 이용률 저조, 할부금융 서비스 이용률 전무함.
3. 리볼빙 대금 채무불이행 예측 고객 중 40대 초반 > 40대 후반 > 50대 초반 순으로 비중이 높았음.

   연령대가 위험등급에 영향을 미친다는 카이제곱 독립성 검정 결과의 통계적 유의성을 트리모델로 재차 확인함.





[의사결정 분류나무와 랜덤 포레스트 모델 비교]

1. ROC(Receiver Operation Characteristic) 곡선과 AUC(Area Under Curve) 값을 통해 트리와 랜덤 포레스트 모델의 성능을 비교함.

2. 곡선이 Y축의 1 부근에서 직각으로 꺾여야 좋은 분류모델을 만들었다고 평가할 수 있음. 카드리볼빙대금을 실제로 상환하였는지 여부에 관한 이진 변수 데이터가 없음*을 고려할 때 각 AUC 값은 78.3%, 78.7%로 분류 모형이 적합하다고 판단함.

   *) 보안상 공개 불가능한 데이터이며 현업에서 모델링할 경우에도 상환 여부라는 답이 주어지지 않은 상황일 가능성 있음.

3. 채무불이행(대금 상환 불가능) 고객의 특성을 예측하는 모형을 구현한다는 데이터 분석 목적에 따라 예측력이 다소 낮지만 설명력이 높은 의사결정 분류나무를 채택함.

▼ 의사결정 분류나무 모형 적합을 위한 성능평가

- **모델 성능평가 과정**

  1. 3가지 데이터 셋 분리기준(7:3, 8:2, 6:4)으로 모델링, 9회 시드 고정을 풀고 모델링

  2. 12개 시나리오별 의사결정나무 및 랜덤포레스트 모델의 혼동행렬 내 평가지표 비교함.

- **모델 성능평가 결과**

  1. 고객의 신용 리스크가 없다고 예측하였는데 실제로 있을 오분류율(1-TP rate : 분석목적에 따라 가장 주목해야할 오차비용)이 22.59%로 시드 고정한 7:3 분류 모델이 가장 낮음.

  2. 실제로 리스크 없는 고객을 없다고 예측하거나 실제 리스크 있는 고객을 있다고 예측할 비율, 모형의 정확도는 78.48%로 시드 고정한 7:3 분류 모델이 가장 높음.

  3. 정밀도, 재현률의 조화평균*인 F1 점수가 시드 고정한 7:3 분류 모델이 가장 높음.

     *) 0 ~ 1 사이이며 클수록 두 값 골고루 큰 경우로 모델 성능 높다고 판단함.

  4. 시드 고정한 7:3 분류 모델은 맥니마 검정* 시 유의수준 5%에서 유의확률(p-value)이 0.06으로 H0 채택함. 즉, 실제값과 예측값의 차이가 통계적으로 유의하지 않으며 타 시나리오 대비 예측의 정확도가 높다는 결론을 내림.

     *) 맥니마 검정은 머신러닝에서 2x2 혼동행렬에 대한 예측모델의 정확도를 비교하는 방법으로 사용됨.

지표의 개념	세부 모형 성능평가 지표	7:3 set.seed	8:2 set.seed	6:4 set.seed	random 1	random 3
옳은 예측의 비율	Accuracy (정확도)	78.48%	78.15%	78.27%	78.17%	78.3%
실제로 채무불이행 고객(P)인 사람 중 채무불이행할 것(P)으로 예측된 비율	Sensitivity (Recall, TP Rate) (재현율)	77.41%	76.47%	77.24%	77.12%	77.2%
실제로 채무불이행 고객(P)인 사람 중 상환할 것(N)으로 예측된 비율	1 - TP Rate (오분류율)	22.59%	23.53%	22.76%	22.88%	22.8%
실제로 상환할 고객(N)인 사람 중 채무불이행할 고객(P)으로 예측된 비율	FP Rate(1 - Specificity) (오분류율)	20.5%	20.27%	20.74%	20.83%	20.65%
채무불이행 고객(P)으로 예측한 사람 중 실제로 채무불이행 고객(P)인 경우의 비율	Pos Pred Value(precision) (정밀도)	78.26%	78.19%	78.17%	77.82%	78.19%
정밀도와 재현도 두 지표의 조화평균	F-1 Score	77.83%	77.32%	77.70%	77.47%	77.69%
예측 모델의 정확도를 평가하는 지표로 유의수준(α) 0.05 이상이어야 모형 적합	Mcnemar's Test P-value	6%	0.24%	1.85%	13.18%	3.01%

예측값 / 실제값	Y	N
Y	True Positive(TP)	False Positive(FP)
N	False Negative(FN)	True Negative(TN)

3. 플랫폼 통합을 통한 고객 맞춤형 서비스

📂 사업 개요

- 사업 기획 배경
  - 플랫폼의 분산으로 신한 금융 지주사 고객이 신한 금융 지주사의 혜택이나 편리함을 경험하지 못한다는 문제점을 발견함.
  - 신한 Play를 중심으로 모든 신한 플랫폼의 통합이 이루어지면 좋겠다고 생각함.
- 사업의 방향성
  - 소비성향별 고객 세그먼트 분류
  - 고객 맞춤형 종합금융 추천 알고리즘

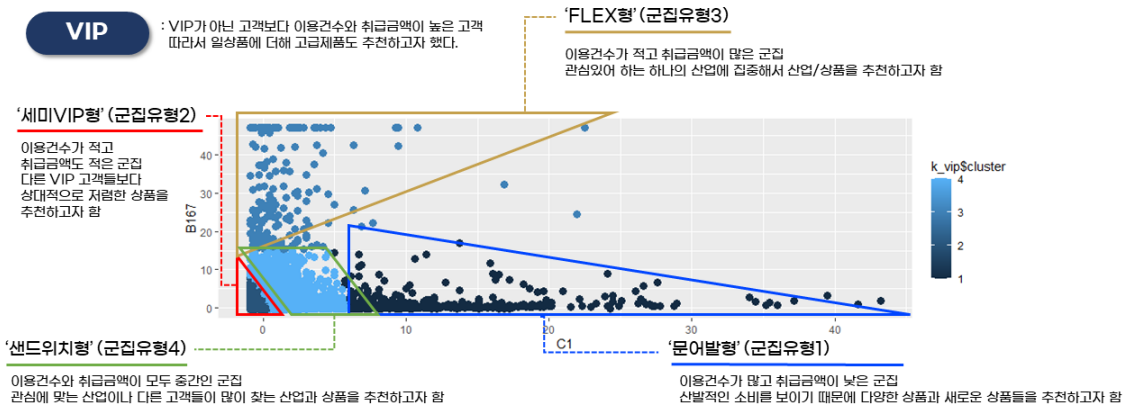
📂 사업 내용

소비 패턴, 개인화, 맞춤형 서비스

1. **카드 이용 건수와 취급금액을 기준으로 고객 군집 분류**

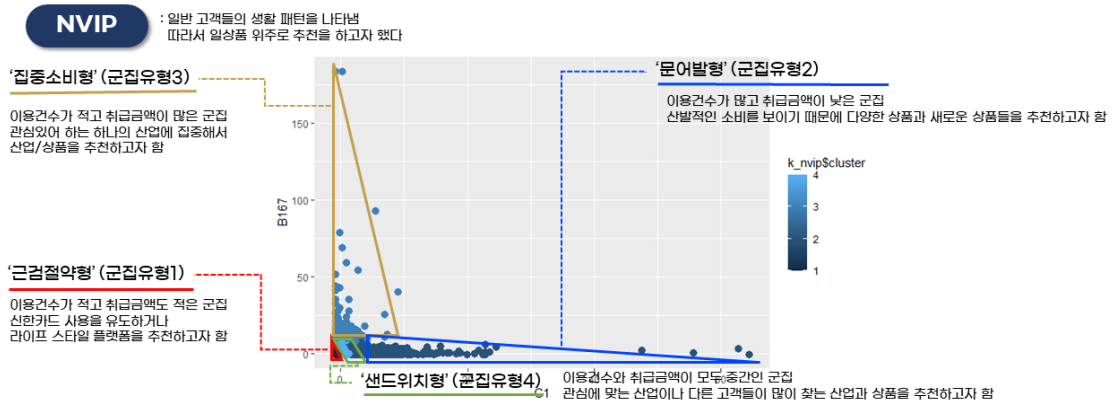
📌 분석 결과 및 인사이트

- 신한카드 VIP 고객 중 월별 이용 건수와 취급금액을 기준으로 분류 생성



문어발형	세미 VIP형
이용건수가 많고 취급금액이 낮은 군집 산발적인 소비를 보이기 때문에 다양한 상품과 새로운 상품들을 추천하고자 함	이용건수가 적고 취급금액도 적은 군집 다른 VIP 고객들보다 상대적으로 저렴한 상품을 추천하고자 함
Flex형	**샌드위치형**
이용건수가 적고 취급금액이 많은 군집 관심있어 하는 하나의 산업에 집중해서 산업/상품을 추천하고자 함	이용건수와 취급금액이 모두 중간인 군집 관심에 맞는 산업이나 다른 고객들이 많이 찾는 산업과 상품을 추천하고자 함

- 신한카드 고객 중 VIP가 아닌 고객을 월별 카드 이용건수와 취급금액을 기준으로 분류



근검절약형	문어발형
이용건수가 적고 취급금액도 적은 군집 신한카드 사용을 유도하거나 라이프 스타일 플랫폼을 추천하고자 함	이용건수가 많고 취급금액이 낮은 군집 산발적인 소비를 보이기 때문에 다양한 상품과 새로운 상품들을 추천하고자 함
집중소비형	**샌드위치형**
이용건수가 적고 취급금액이 많은 군집 관심있어 하는 하나의 산업에 집중해서 산업/상품을 추천하고자 함	이용건수와 취급금액이 모두 중간인 군집 관심에 맞는 산업이나 다른 고객들이 많이 찾는 산업과 상품을 추천하고자 함

2. **연관성 분석을 이용한 고객 소비 패턴 추출**

📌 분석 결과

(* 비밀 유지 서약 문제로 자세한 분석 결과는 기재하지 않음)

📌 인사이트

- 생활필수적 유형의 소비패턴(e.g. {고정지출, 이동}) 이외에 고객 유형별 취향과 가치관이 반영된 2가지 연관규칙 있는 산업 간 집합이라는 패턴을 발견함.
- 생활에 필수적인 소비 패턴 이외의 소비 패턴 발견 후 각각 군집의 특성과 소비 간 연관 규칙을 반영하여 인사이트를 도출함.

  예) 카드사 VIP가 아닌 집단 중 문어발형의 {중고품 판매, 이동} 연관규칙
  - 중고품 거래를 선호하는 집단이라고 판단
  - 중고품 거래를 선호할수록 리사이클, 업사이클 등의 환경 이슈에 관심이 많을 것이라고 판단
  - 신한 올댓 쇼핑 중 'ESG 몰' 추천, 연동 서비스 제공

📌 사업 연결

- 앞서 분류한 VIP 집단과 NVIP 집단 내 4가지 소비 유형별로 소비 규칙 발견, 각 소비 패턴에서 고객의 니즈를 찾고 그를 만족시킬 수 있는 맞춤형 서비스를 제안
- *맞춤형 서비스 종류 : 고객의 소비 패턴을 개인화하고 각 패턴에 맞는 제휴 쇼핑몰 추천, 펀드 추천, 신한 플랫폼 추천 서비스
- 궁극적으로 **신한 PLAY를 중심으로 신한 올댓 쇼핑, My Car, 신한 알파, 신한라이프 등의 신한 금융 지주사의 모든 플랫폼을 통합**하고 **신한카드 뿐만 아니라 모든 계열사**

로 사업을 확장한다.

- 고객의 생애 주기에 금융을 제공하는 점유시간을 확장하는 플랫폼 내 서비스를 구축하는 한편 고객에게 소비에서 파생된 새로운 투자 대안을 제공함

- 단순히 특정 산업에 대한 소비지출 빈도가 높은 고객 유형을 분류하는 게 아니라 금융활동 소비습관이라는 1차 분류 후 소비패턴 유사한 고객 유형 2차 분류한다는 점에서 하향식 접근 방식(backward approach)으로 비즈니스 기획함

** 핵심 사업 위주로 작성*

▼ 세미 VIP형

	식료품 - 유통
특징	식료품을 구매하면서 다른 산업의 제품과 서비스를 소비하는 것을 선호하는 유형
맞춤형 추천 서비스	**[Hmall, 신세계 몰 등 백화점과 같은 제휴 쇼핑몰 추천]** - 세미 VIP 유형은 다른 vip 유형보다 취급금액이 적은 유형이기 때문에 백화점 중에서도 종합 쇼핑몰 위주로 추천 - 한 번에 다양한 품목을 소비하는 경향이 있는 유형이기 때문에 종합 쇼핑몰을 추천 **[대형 유통업 테마 펀드 추천]** - 유통 산업의 소비로 이어지는 유형이므로 대형 유통업 영위 주식회사의 지분 증권 종목 펀드 추천 *예) 신한 K-컬쳐 증권 자투자신탁 (K-이커머스, 헬스케어, 미디어)*
신한 금융 지주사 플랫폼으로의 확장	**[신한 올댓쇼핑]** : 종합 쇼핑 플랫폼으로 이동할 수 있도록 연동 **[신한 알파(신한투자증권 플랫폼)]** : 고객에게 맞춤형 펀드를 추천한 뒤 해당 펀드 상품을 확인할 수 있도록 신한 알파 플랫폼으로 연동 : 신한카드 고객이 신한 투자 증권을 통한 금융 거래까지 도달할 수 있도록 플랫폼 간 연결성을 높이고 플랫폼 기능을 확장하여 접근성을 높임

▼ Flex형

	농산품 - 이동
특징	- 거리가 멀더라도 신선제품을 구매하러 이동할 가능성이 높은 유형 - Flex 형은 취급금액이 높고 이용 건수는 적어 관심있는 곳에 많은 지출을 보이는 고객 유형이기 때문에 신선 제품에 큰 관심을 보이는 유형이라고 판단

	농산품 - 이동
맞춤형 추천 서비스	**[신한 ‘올댓 쇼핑’ 내 친환경 몰 추천]** - 고객이 선호하는 신선식품과 친환경 제품을 온라인으로도 구매할 수 있도록 신한 올댓 쇼핑 내 친환경 몰을 추천 **[에너지 및 농산물 관련 펀드 추천]** - 신선 제품에 관심이 많고 큰 지출을 보이는 집단이기 때문에 에너지 및 농산물 관련 펀드 추천
신한 금융 지주사 플랫폼으로의 확장	**[신한 올댓 쇼핑]** : 종합 쇼핑 플랫폼으로 이동할 수 있도록 연동 **[신한 알파(신한 금융 투자 플랫폼)]** : 고객에게 맞춤형 펀드를 추천한 뒤 해당 펀드 상품을 확인할 수 있도록 신한 알파 플랫폼으로 연동 : 신한카드 고객이 신한 투자 증권을 통한 금융 거래까지 도달할 수 있도록 플랫폼 간 연결성을 높이고 플랫폼 기능을 확장하여 접근성을 높임

	고정지출 - 이동
특징	고정지출(보험료, 고정비)와 교통비에 많은 돈을 쓰는 유형
맞춤형 추천 서비스	**[신한 ’My Car’ 플랫폼 추천]** - 전기차 할부금융 / 렌트 / 오토리스 등 거래 추천 - 이동 산업으로의 큰 지출이 많은 유형이기 때문에 전기차로의 관심을 유도하고 신한 자동차 라이프 플랫폼 사용을 추천한다. **[자동차 손해보험상품 추천]** - My car 플랫폼 내 신한그룹사 자동차 보험 상품 가입 추천 - 신한 카드 이용고객에 한하여 보험료 할인 - 보험료 지출이 높고, 이동 산업으로의 연속적인 소비 금액도 높은 유형이기 때문에 이동 산업과 관련된 고정 지출 소비에 혜택을 부여하면서 신한 자동차 손해보험 상품으로 유도
신한 금융 지주사 플랫폼으로의 확장	**[신한 My Car]** : 신한 Play에서 My Car 플랫폼으로 확장해나갈 수 있도록 연동 **[신한 EZ 손해보험]** : 고객이 신한 play에서 신한 보험 앱으로 이동할 수 있도록 연동 : 신한카드 고객이 신한 보험 이용까지 도달할 수 있도록 플랫폼 간 연결성을 높이고 플랫폼 기능을 확장하여 접근성을 높임

▼ 샌드위치형

	(유통, 생활서비스, 의료, 가스, 이동, 전자통신, 식료품, 문화예술, 고정지출) - 외식
특징	외출 시 외식을 자주 하는 고객 유형
맞춤형 추천 서비스	**[신한 Play 프리미엄 쿠폰 추천]** -신한카드 종합생활금융 플랫폼, 신한 플레이의 프리미엄 쿠폰 추천 *예) VIP 집단이기 때문에 가격대가 있는 다이닝 위주 레스토랑 추천* **[리테일 관련 펀드 추천]** - 다양한 문화생활을 위해 외출이 잦고, 외식 산업에 대한 소비로 이어지는 유형이기 때문에 문화예술, 외식 산업을 영위하는 주식회사의 지분증권 종목 펀드 추천 *예) 신한 K-컬쳐 증권 자투자신탁 (K-이커머스, 헬스케어, 미디어)*

	(유통, 생활서비스, 의료, 가스, 이동, 전자통신, 식료품, 문화예술, 고정지출) - 외식
신한 금융 지주사 플랫폼으로의 확장	**[신한 알파]** : 고객에게 맞춤형 펀드를 추천한 뒤 해당 펀드 상품을 확인할 수 있도록 신한 알파 플랫폼으로 연동 : 신한카드 고객이 신한 투자 증권을 통한 금융 거래까지 도달할 수 있도록 플랫폼 간 연결성을 높이고 플랫폼 기능을 확장하여 접근성을 높임

	중고 - 외식 | 중고 - 유통
특징	재사용과 같은 환경 이슈에 관심이 있을 가능성이 상대적으로 높거나, 개성과 자기다움이 중요한 소비 가치관을 가진 집단일 가능성이 높은 유형
맞춤형 추천 서비스	**[맞춤형 제휴 쇼핑몰 추천]** - 비건 음료 가맹점 추천 - 개성과 자기다움을 표현하기 위해 패션에 신경을 쓸 가능성이 높기 때문에 LF몰, AK몰, Cjonstyle 등 리테일 몰 내 백화점 의류 브랜드를 추천 **[신한 '올댓쇼핑' ESG관 추천]** - ESG 전용 쇼핑몰 - '친환경관' 플랫폼 연동 **[My Car 플랫폼 추천]** - 전기차로의 관심 유도 **[ESG 관련 펀드 추천]** - ESG 펀드 추천 예) *신한 K-컬쳐 증권 자투자신탁(K-이커머스, 헬스케어, 미디어)* 예) *신한 글로벌 탄소중립솔루션 증권 자투자신탁 (수소, 모빌리티, 푸드테크 중 탄소중립의 사회 패러다임의 주식회사 종목 펀드)*
신한 금융 지주사 플랫폼으로의 확장	**[신한 올댓쇼핑]** : 신한 Play에서 신한 올댓쇼핑으로 이동이 가능하게끔 연동 **[My Car]** : 신한 Play에서 My Car 플랫폼으로 이동이 가능하게끔 연동 **[신한 알파]** : 고객에게 맞춤형 펀드를 추천한 뒤 해당 펀드 상품을 확인할 수 있도록 신한 알파 플랫폼으로 연동 : 신한카드 고객이 신한 투자 증권을 통한 금융 거래까지 도달할 수 있도록 플랫폼 간 연결성을 높이고 플랫폼 기능을 확장하여 접근성을 높임

▼ 근검절약형

	고정지출 - 이동
특징	매달 고정적으로 지출하는 금전(공과금과 교통비 납부)만 신한 카드에서 이체할 가능성이 높은 유형

	고정지출 - 이동
맞춤형 추천 서비스	**[신한 플레이 제휴사 할인 혜택 및 푸시 알림]** - 고정 지출 이외의 소비 지출도 신한카드로 결제 하는 경험을 제공하기 위한 할인 쿠폰 추천 - 할인 정보를 홍보하여 신한카드 이용을 유도 **[신한 EZ 손해보험 상품 추천]** - 신한 카드 이용고객에 한하여 보험료 할인 - 신한카드로 납부하는 고정치출 중 보험료를 절감할 수 있는 혜택을 부여하고 홍보함 - 신한 금융 서비스 이용을 유도할 수 있음
신한 금융 지주사 플랫폼으로의 확장	**[신한 EZ 손해보험]** : 고객이 신한 play에서 신한 보험 앱으로 이동할 수 있도록 연동 : 신한카드 고객이 신한 보험 이용까지 도달할 수 있도록 플랫폼 간 연결성을 높이고 플랫폼 기능을 확장하여 접근성을 높임

▼ 문어발형

	고정지출, 반려동물 - 이동
특징	반려동물과 함께 드라이브를 즐기는 고객 유형
맞춤형 추천 서비스	**[신한 '올댓 쇼핑' 플랫폼 추천]** - 차에 설치할 반려동물 안전 장치에 대한 수요가 있을 것이므로 리빙 제품을 구매할 수 있는 종합 쇼핑 플랫폼을 추천함
신한 금융 지주사 플랫폼으로의 확장	**[신한 '올댓 쇼핑']** : 신한 Play에서 신한 올댓쇼핑으로 이동이 가능하게끔 연동

▼ 샌드위치형

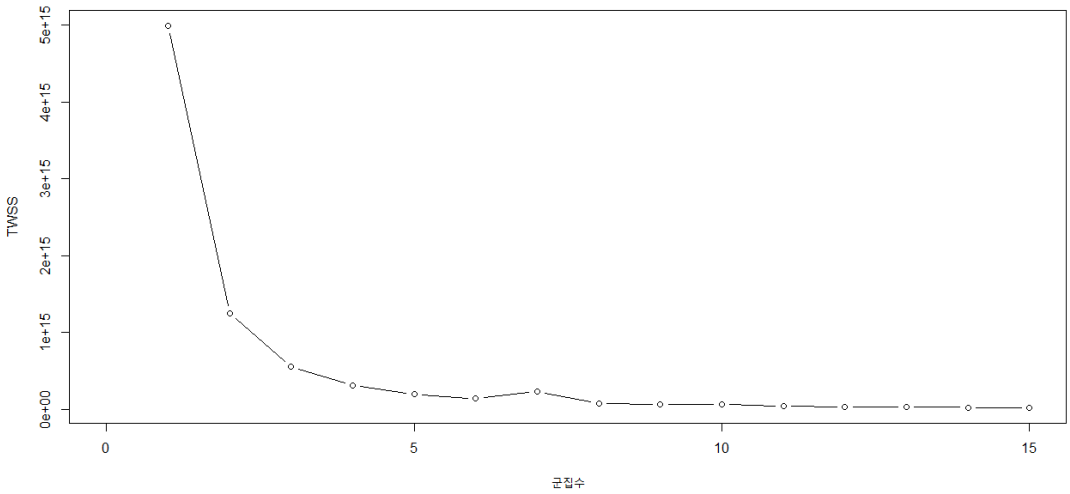
	(유통, 생활서비스, 의료, 이동, 전자통신, 식료품, 중고, 문화예술, 고정지출) - 외식
특징	외출 시 외식을 자주 하는 고객 집단
맞춤형 추천 서비스	**[외식 쿠폰 추천]** - 신한플레이의 "맛있는 쿠폰"에서 추천 *예) NVIP인 만큼 VIP와 다르게 가성비 또는 가심비 있는 외식 프렌차이즈 추천* **[리테일 관련 펀드 추천]** - 다양한 문화생활을 위해 외출이 잦고, 외식 산업에 대한 소비로 이어지는 유형이기 때문에 문화예술, 외식 산업을 영위하는 주식회사의 지분증권 종목 펀드 추천 *예) 신한 K-컬쳐 증권 자투자신탁 (K-이커머스, 헬스케어, 미디어)*
신한 금융 지주사 플랫폼으로의 확장	**[신한 알파]** : 고객에게 맞춤형 펀드를 추천한 뒤 해당 펀드 상품을 확인할 수 있도록 신한 알파 플랫폼으로 연동 : 신한카드 고객이 신한투자증권을 통한 금융 거래까지 도달할 수 있도록 플랫폼 간 연결성을 높이고 플랫폼 기능을 확장하여 접근성을 높임

📂 분석 과정

▼ 비슷한 소비 성향 유형으로 고객 군집 생성

K-means 군집분석을 이용한 유사 소비 성향별 고객 군집 분석 과정

1. 신한카드 Vip 고객 여부(P4)로 고객을 분류하여 각각 새로운 데이터 프레임을 생성함
   이 때, 리볼빙 이용 횟수가 많아 소비보다 상환에 신경써야하는 “HH”, “HL” 위험 고객 집단은 분석에서 제외함.
2. Vip, Nvip 각각의 데이터 프레임 에서 이용건수와 취급금액 표준화 진행
3. 최적 cluster 개수 찾기 - 엘보우(Elbow) 기법



최적의 k(군집 수)는 3이라는 결과가 나왔지만, k=4일 때 군집이 더 명확하기 분리되었기 때문에 k를 4로 설정함.

3. K-means 클러스터링
   a. 표본추출하지 않고 각각 전수로 군집을 생성함.
      *전체 Vip 고객, 267378명 | 전체 Nvip 고객, 205,847명
   b. 각각 4개의 군집으로 클러스터링을 진행함.

▼ 고객 개별 소비패턴 도출 및 맞춤형 서비스 제공

Apriori 알고리즘에 의한 연관규칙분석 기반의 산업 간 집합 분류 과정

1. 메모리 사용량 고려하여 개별 소비지출변수(167개)가 아닌 산업변수(26개)로 분석함.
2. 전수로 최초 모델링 수행 시 생활필수적 지출에 해당하는 산업 간 연관규칙 집합만 도출됨. 예) {이동, 고정지출}
   따라서 군집별 고객 100명의 표본을 랜덤 추출하여 소비패턴을 파악함.

3. 특정 산업에서 소비지출금액 있으면 “yes”, 없으면 “no”로 설정 후 데이터를 더 이상 분리할 수 없는 단위인 트랜잭션으로 변환함.
2. 빈발집합에 대한 상대적 지지도 및 신뢰도 임계값 설정함.
   (frequent item sets of minimum support threshold)

   모든 경우의 수인 (2의 n제곱 - 1)은 시간복잡도를 지수적으로 높이기 때문에 속도와 메모리 사용량 문제를 완화하기 위해 연관규칙의 효용성을 나타내는 지표의 임계값을 다소 높게 설정함.

	상대적 최소지지도 = 30% 최소신뢰도 = 75%	**상대적 최소지지도 = 50% 최소신뢰도 = 75%**
VIP집단	세미 VIP형 FLEX형 샌드위치형	문어발형 *) 이용건수가 많고 취급금액이 낮은 군집으로 산발적인 소비를 보이기 때문에 산업 간 연관규칙집합 추가 필터링
NVIP집단	근검절약형 집중소비형 샌드위치형	문어발형

4. 총평

📂 프로젝트 수행 시 발생한 문제점 및 해결 과정

문제점 발생 지점	**내용**
데이터 전처리 시 데이터 확인 과정	**[문제]** 데이터 확인 시, 데이터의 패턴코드(E1~E6)이 월간 거래 서비스 이용 여부에 대한 요약변수 형태(12개월)로 제시되었음. **[해결과정] :** 파생변수 생성 • nE1~nE6 변수를 생성 - 연간 이용 횟수를 보기 위해서 각 패턴코드 중 ‘1(이용)’의 개수를 추출 - 모든 변수를 일일이 이항 변수로 추출하기에는 `2^(12*473225)*6` 의 경우의 수가 발생하기 때문에 시간 복잡도 측면에서 비효율적이라고 판단 • sE1 변수 생성 - E1(리볼빙 이용 여부)에서 ‘1(이용)’이 연속으로 3번 이상인 고객을 True, 아닌 고객을 False로 나타내는 변수 • G1 변수 생성 - 리볼빙 이용 횟수에 따른 - nE1 == 0 ~ "N", - nE1 <= 2 ~ "L", - nE1 <= 12 & sE1 == FALSE ~ "HL", - sE1 == TRUE ~ "HH”

문제점 발생 지점	내용
고객 성향 분석 시 군집분석 과정	**[문제]** • K-medoids 군집분석은 결제금액 변수의 0(결측)이 너무 많아 군집의 대푯값이 0으로 도출되어 모델 활용 불가능함. • 계층적 군집분석은 빅데이터로 n의 크기가 커서 일반적으로 시간복잡도가 O(n의 세제곱) 높은 문제 발생함. **[해결과정]** • K-means 군집분석을 이용하여 고객 소비지출 성향 분석 수행하여 시간복잡도를 O(nlogn)으로 개선함. • 신용카드 이용건수 및 결제금액이 이상치라고 판단될 정도로 소비 규모가 높은 군집을 찾기 위해 K-means 군집분석이 적합함.
고객 소비 패턴 분석 시 분석 방법 선택 과정	**[문제]** 고객의 소비 패턴을 분석한 후 각각의 소비 패턴을 가진 고객에게 맞춤형 서비스를 추천하는 모델을 만들고자 했으나 네트워크 분석과 연관성 분석 중 모델 결정에 시행착오를 겪음. **[해결과정]** • 네트워크 분석 - 네트워크 분석은 빈도를 기준으로 인접 노드를 형성함. - 따라서 결제 금액이 있을 경우를 '1'로 설정함. - 하지만 산업마다 소비지출항목의 수가 상이하기 때문에 빈도 기준으로 추천 알고리즘을 설정하기에는 적합한 모델이 아니라고 판단함. - 표준화를 하게 되면 빈도가 1 미만으로 작아지기 때문에 결과를 확인하기 어려울 것으로 판단함. * *데이터 커서 모델 구현이 어려움* • 연관 분석 - 고객의 연속적인 소비를 통해 소비 패턴을 도출하는 것이 적합하다고 판단함. - 데이터 분석을 통해 범용적인 지출 이외의 새로운 연관 규칙을 찾고, 해당 규칙을 이용하여 맞춤형 서비스를 제안하는 것이 적합하다고 판단함.

📂 프로젝트 한계점 및 개선 계획

이항 로지스틱 회귀분석 모형 적합을 위한 성능 평가 미흡	**[한계점]** 이항 로지스틱 회귀분석 성능 평가 시 ANOVA 검정을 진행하였지만 검정 결과를 반영한 성능 개선에 미흡함. * *추출한 산업이 검정력을 크게 증가시키진 못한다고 결과가 나옴.* **[개선 계획]** • 이탈도 결정계수 및 수정된 이탈도 결정계수 측정함. • 독립변수(산업항목) 간 다중공선성 파악함. • 모형에 영향력 있는 변수 내림차순 시각화함. • 혼동행렬로 오분류율(제1, 2종 오류 이른바 오차비용), F1 등으로 모형 적합을 위한 성능 평가.
Apriori 알고리즘의 시간복잡도 문제	**[한계점]** • 연관규칙분석에 이용한 Apriori 알고리즘의 지수형 시간복잡도 O(2의 n제곱) 문제을 해결하고자 알고리즘 개선 시도하였으나 미흡함. • 1개월 간 수십만 명의 고객 데이터에서 100명 표본을 추출하여 모델링함. 시계열이 단기이며 고객 수가 많았다는 점에서 한계가 있다는 결론을 내림. **[개선 계획]** • 연관규칙분석의 시간복잡도를 O(n)까지 내려야 보조기억장치가 아닌 주기억장치에 데이터가 올라갈 수 있다는 점에서 보완 필요함. • Frequent Pattern Growth 알고리즘으로 연관규칙분석 개선 필요함.

▼ 이항 로지스틱 회귀분석 성능 평가 (ANOVA)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			88903	6022.2		
C2	1	0.733	88902	6021.4	0.392039	
C3	1	0.098	88901	6021.3	0.754296	
C4	1	3.328	88900	6018.0	0.068092	.
C5	1	2.964	88899	6015.1	0.085127	.
C6	1	0.154	88898	6014.9	0.694568	
C7	1	0.063	88897	6014.8	0.802543	
C8	1	7.284	88896	6007.6	0.006957	**
C9	1	0.053	88895	6007.5	0.818706	
C10	1	0.326	88894	6007.2	0.567841	
C11	1	0.204	88893	6007.0	0.651291	
C12	1	0.594	88892	6006.4	0.440819	
C13	1	0.398	88891	6006.0	0.528299	
C14	1	0.069	88890	6005.9	0.793484	
C15	1	0.336	88889	6005.6	0.562109	
C16	1	69.036	88888	5936.5	< 2.2e-16	***
C17	1	0.323	88887	5936.2	0.569728	
C18	1	3.621	88886	5932.6	0.057054	.
C19	1	1.476	88885	5931.1	0.224351	
C20	1	1.827	88884	5929.3	0.176506	
C21	1	5.699	88883	5923.6	0.016971	*
C22	1	1.856	88882	5921.7	0.173058	
C23	1	0.917	88881	5920.8	0.338201	
C24	1	4.736	88880	5916.1	0.029539	*
C25	1	1.278	88879	5914.8	0.258356	
C26	1	0.375	88878	5914.4	0.540086	
C27	1	21.474	88877	5893.0	3.586e-06	***

📂 업무 분배

팀원	업무
강수연	데이터 전처리 의사결정나무 의사결정나무 성능평가 연관규칙분석 시스템 구현 (신사업 기획)
김세은	데이터 전처리 군집분석 이항로지스틱 회귀분석 연관규칙분석 시스템 구현(신사업 기획)

📂 프로젝트 소감

팀원	프로젝트 소감
강수연	• 데이터 분석 프로젝트의 목표 달성에 필요한 3가지 역량 파악 및 활용 1. 데이터 분석 결과를 해석하여 **실제 비즈니스에서 문제를 해결**하는 실용성 (그로스 해킹) 2. 머신러닝 모델 적합(fitting)을 위한 **성능 평가지표 우선순위 결정**의 중요성 3. 머신러닝 모델이 실용적인 시간 안에 구현될 수 있도록 알고리즘의 **시간복잡도 (BigO)에 대한 피드백**의 필요성 • 데이터 분석 협업 과정에서 얻은 의사 전달력 1. 팀원들 간 프로젝트 관련 요구사항을 명확하게 전달하기 위한 고민과 프로젝트가 얼마나 진행되었는지 중간보고를 철저히 함 2. 타인에게 사업 아이디어를 피칭할 때 모든 이해관계자에게 공유가능할 수준의 직관적인 전달력에 대한 협의

팀원	프로젝트 소감
김세은	• 데이터 분석 과정에서 얻은 역량 및 개선점 1. 실제 고객 데이터를 가지고 데이터 분석부터 신사업 기획까지 모든 과정에 주도적으로 참여하면서 **데이터 기반의 의사결정 역량**을 키움. 2. 새롭게 배운 점과 앞으로 공부해야 할 점을 알게 됨. - Apriori 알고리즘의 **시간분석도(BigO)에 대한 개선의 필요성 -** 분석 모델링 성능평가의 중요성 - 머신러닝과 알고리즘 공부가 필요함 • 데이터 분석 협업 과정에서 얻은 의사소통 역량 1. 내가 해석한 분석 결과를 팀원들에게 설명하는 과정에서 어떻게 하면 더 쉽게 의미를 전달할지 고민함으로써 전달력과 표현법을 배움. 2. 최종 사업 아이디어를 실무진에게 피칭할 때 본 프로젝트의 목표 달성 과정을 효과적으로 전달하기 위해서 사업의 배경, 방향성, 분석 과정, 도출한 인사이트, 인사이트를 활용한 사업 설명 순으로 구조화하여 전달하는 스토리텔링 능력을 키움.