기계학습을 활용한 카드사 신용도 평가

- 리스크 최소화의 관점에서

오태경

김유정

김경진

박지혜

CONTENTS

1. 프로젝트 개요 (Gantt Chart)

1. 프로젝트 개요 (업무)

업무
탐색적 데이터 분석
그래프 시각화
변수선정
통계 기반 분석
모델 생성
파라미터 튜닝, 평가
데이터 준비
데이터 전처리 (Data Cleaning)

1. 프로젝트 개요

배경

'영끌 가계부채' 사상 최대...GDP 맞먹는 1765조 돌파

-사상최대 가계부채, 2030은 은행대출, 6070은 비은행대출

배경 : 코로나로 인한 '생활고' '영끌' '빚투'로 가계 부채 최대치
-> 신용 카드사에 연체율이 높으면 당기 순이익에 악영향을 미침

*** 금융권의 관심사 이동**

영업 이익 최대화 ⇨ 리스크 최소화

*** 카드사의 관심사 이동**

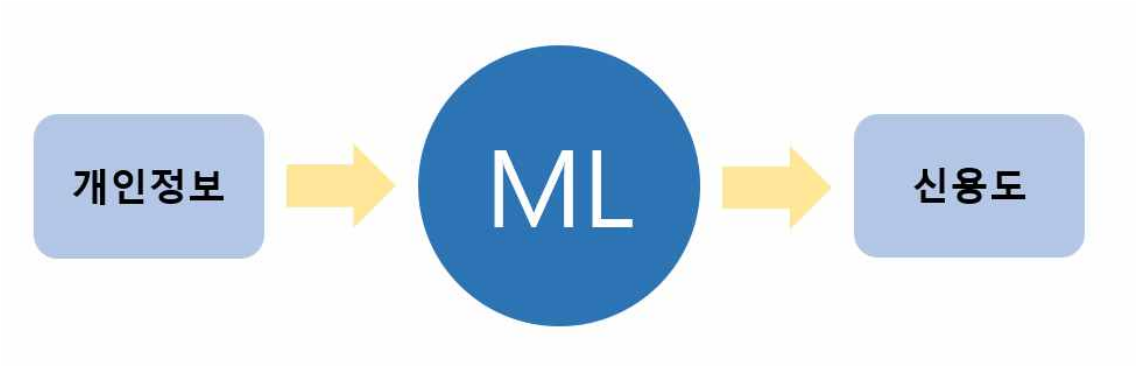
신용도 외 마이 데이터 사용 측면

데이터 분석 및 머신러닝

카드 이익 최대화 ⇨ 연체 비용 최소화

목적 : 개인 정보를 활용한 정확한 신용도 등급 부여로 미래 연체 확률 감소

2. 개발 요약



정형 데이터를 활용한
지도학습 머신러닝 분류 모델 학습 및 예측

3. 아키텍처 설계

구분	소프트웨어		버전
Language	R		4.0.5
IDE	R studio		1.4.1103
라이브러리	데이터 변환	stringr	1.4.0
		mltools	0.3.5
		data.table	1.14.0
		dplyr	1.0.6
	시각화	ggplot2	3.3.3
		corrplot	0.88
		Epi	2.44
		caret	6.0-88
	학습모델	nnet	7.3-16
		randomForest	4.6-14
		kernlab	0.9-29
		xgboost	1.4.1.1

4. 데이터 개요

*실제 중국 카드사 고객 개인 정보를 익명화하여 학습용으로 정제한 데이터

Value-name	Value-discription	Unique-value	Data-type
Index	인덱스	36457	integer
gender	성별	2	character
car	차량 여부	2	character
reality	부동산 여부	2	character
child_num	자녀 수	9	integer
income_total	총수입	265	numeric
income_type	수입 형태	5	character
edu_type	교육 형태	5	character
family_type	가족 구성	5	character
house_type	주거 형태	6	character
DAYS_BIRTH	출생일	7183	integer
DAYS_EMPLOYED	업무일수	3640	integer
FLAG_MOBIL	휴대폰 여부	1	integer
work_phone	업무용 전화 여부	2	integer
phone	전화 여부	2	integer
occyp_type	직업	19	character
family_size	가족수	10	numeric
begin_month	신용카드 발급월	61	numeric
credit	신용도	3	character

출처: 신용카드 사용자 연체 예측 AI 경진대회 - DACON
熊学堂 · 课程介绍 | 在线实习 之《信用卡申请评分模型》(qq.com)

5. 데이터 전처리

Ⅰ. 결측치

변수	개수
Occyp_type	11323개
credit	10000개

Ⅱ. 각 변수별 전처리

(1) 명목형 변수

변수	전처리 방법
gender	형변환 (factor)
car	형변환 (factor)
reality	형변환 (factor)
income_type	형변환 (factor) 재범주화 ('Student' -> 'State servant')
edu_type	형변환 (factor) 재범주화 ('Higher education' -> 'Academic degree')
family_type	형변환 (factor)
house_type	형변환 (factor)
work_phone	형변환 (factor)
phone	형변환 (factor)
email	형변환 (factor)
occyp_type	결측치 처리 (DAYS_EMPLOYED 변수와 조합하여 범주 생성(not in work), 나머지 결측치는 따로 범주 생성 (missing value))

5. 데이터 전처리

(2) 순서형 변수

변수	전처리 방법
child_num	재범주화 (4 이상의 값들을 새로운 범주)
family_size	재범주화 (4 이상의 값들을 새로운 범주)

(3) 연속형 변수

변수	전처리 방법
income_total	IRQ 범위에 벗어나는 값 이상치로 판별 -> Log Scale 후 극단적인 이상치 완화
DAYS_BIRTH	Min Max Scale
DAYS_EMPLOYED	이상치 조정 (265243 -> 0), 절대값, IRQ 범위에 벗어나는 이상치 확인 -> 정규화 어려움, 파생변수 생성
begin_month	Min Max Scale

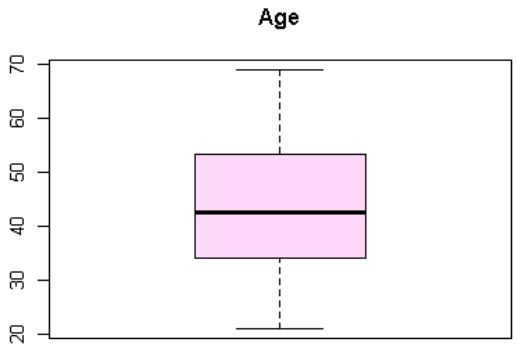
Ⅲ 파생변수 및 요약변수 생성

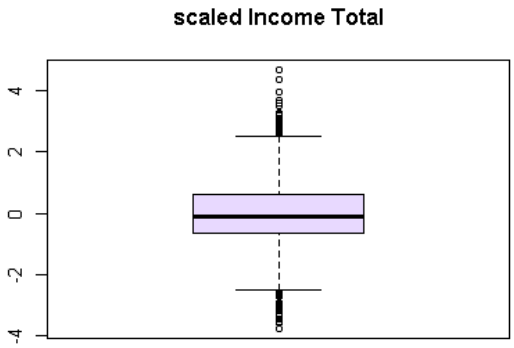
변수	생성 방법	설명
out_child	family_size - child_num	자녀 외 가족 수
age	DAYS_BIRTH / 365	나이
iswork	DAYS_EMPLOYED 이상치 여부	근무 여부
YEARS_EMPLOYED	DAYS_EMPLOYED / 365	근속년수
DAYS_UNEMPLOYED	DAYS_BIRTH - DAYS_EMPLOYED	실업 기간
YEARS_UNEMPLOYED	DAYS_UNEMPLOYED / 365	실업 년수
scale_income_per_year	scale(log(income_total / (YEARS_EMPLOYED+1)+1))	근속년수 당 연봉
isna	occyp_type : 0 or 1	결측치 여부
기타 연속형 변수	Min Max scale	scale 요약 변수

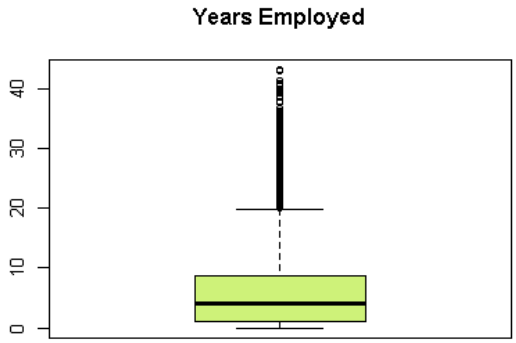
Ⅳ 변수 제거

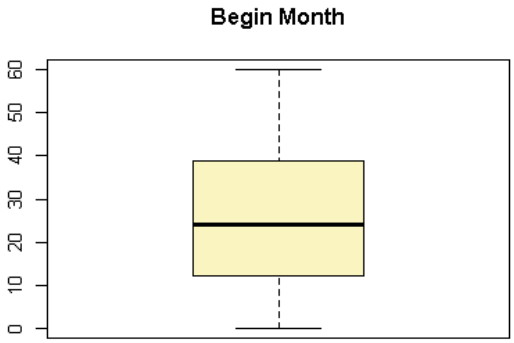
변수	전처리 방법
index	제거
FLAG_MOBIL	제거

6. EDA (연속형 변수의 이해)







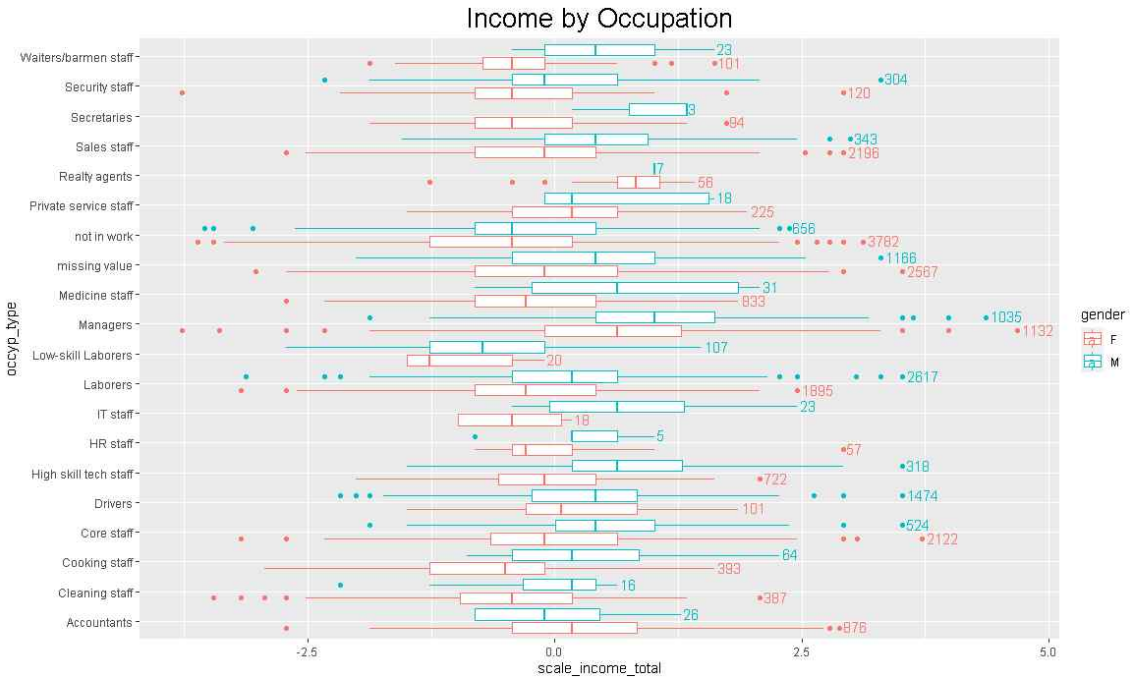


변수	Age	Scaled Income_Total	Years Employed	Begin Month
Min	21.11	-3.78	0.00	0.00
1st Qu.	34.10	-0.65	1.12	12.00
Median	42.59	-0.11	4.26	24.00
Mean	43.72	0.00	6.03	26.16
3rd Qu.	53.24	0.63	8.64	39.00
Max	68.91	4.68	43.05	60.00
IRQ	19.14	1.28	7.52	27

해당 카드사의 주요 고객층은 34~53세의 근속 연수가 평균 6년인 직장인이다. 이들의 대다수는 평균 연봉에 차이가 없으며 2년 전후의 카드이용 실적을 보유하고 있다.

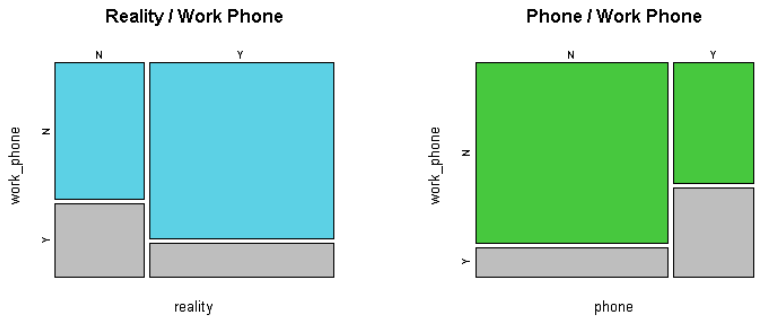
6. EDA (변수간의 관계 분석)

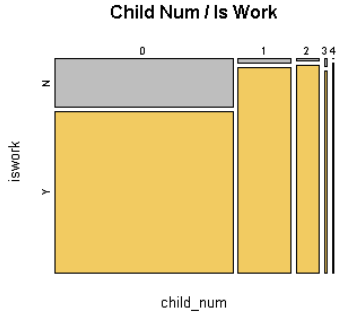
직업과 성별에 따른 총 소득

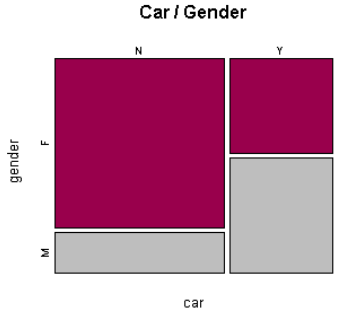


직업군에 따라 소득의 차이가 뚜렷하게 나타나고, 직업에 따라 성별의 비율이 다르며 대부분의 직업에서 남성의 소득이 여성보다 높다. 이는 독립변수의 조합에 따라 소득구간이 구분됨을 의미한다. 따라서 Income Total에 대해서는 정보 손실을 줄이기 위해 이상치를 제거하거나 임의로 값을 조정하지 않는 것을 고려했다.

범주형 변수 모자이크 플롯

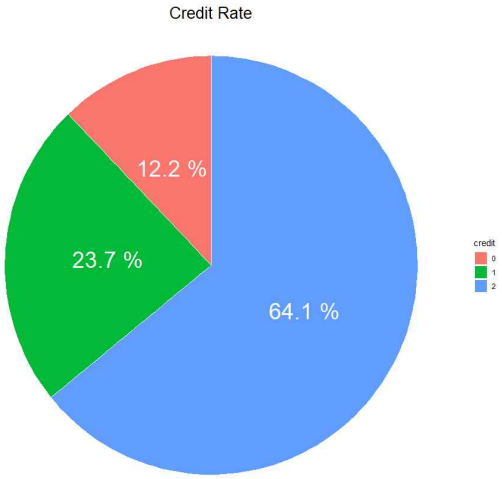






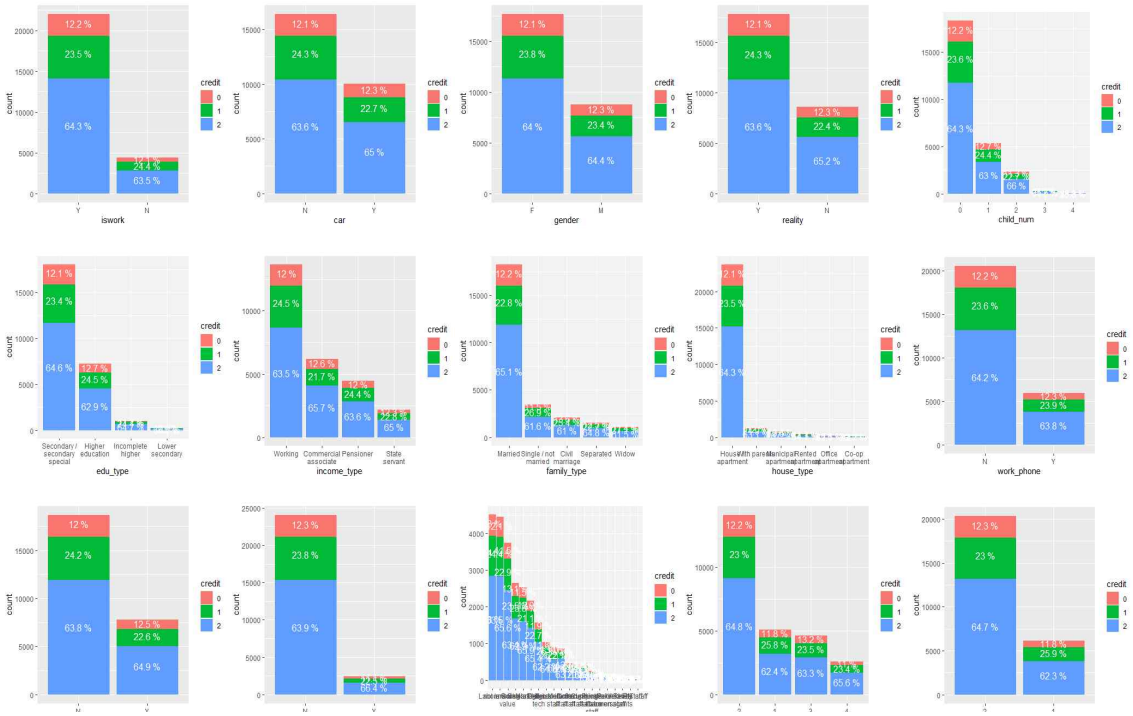
다음의 그래프를 볼 때, 의외의 부분에서 상관관계를 찾을 수 있었으며, 이러한 결과는 변수 선별 시 다중공선을 제거하거나 교호작용을 분석하는 데에 참고하였다.

6. EDA (Credit 비율 확인)



전체 데이터에서 credit 2의 비율이 상대적으로 많기 때문에 데이터 불균형을 해결하기 위해 추후 모델 학습 전 Up Sampling을 고려하였다.

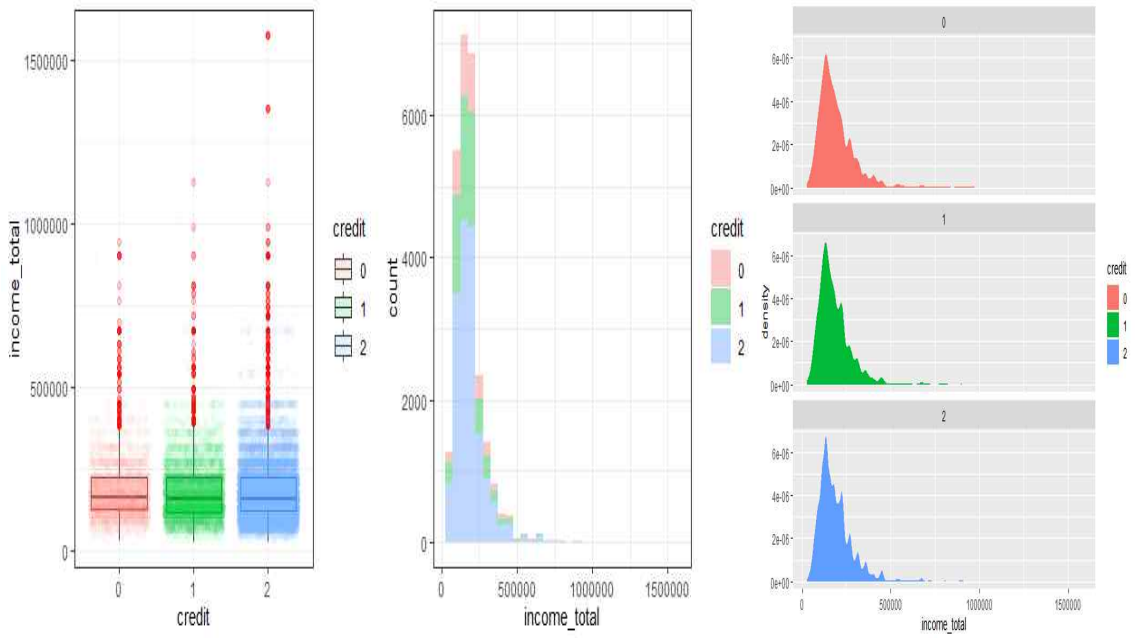
6. EDA (범주형 변수 막대그래프)



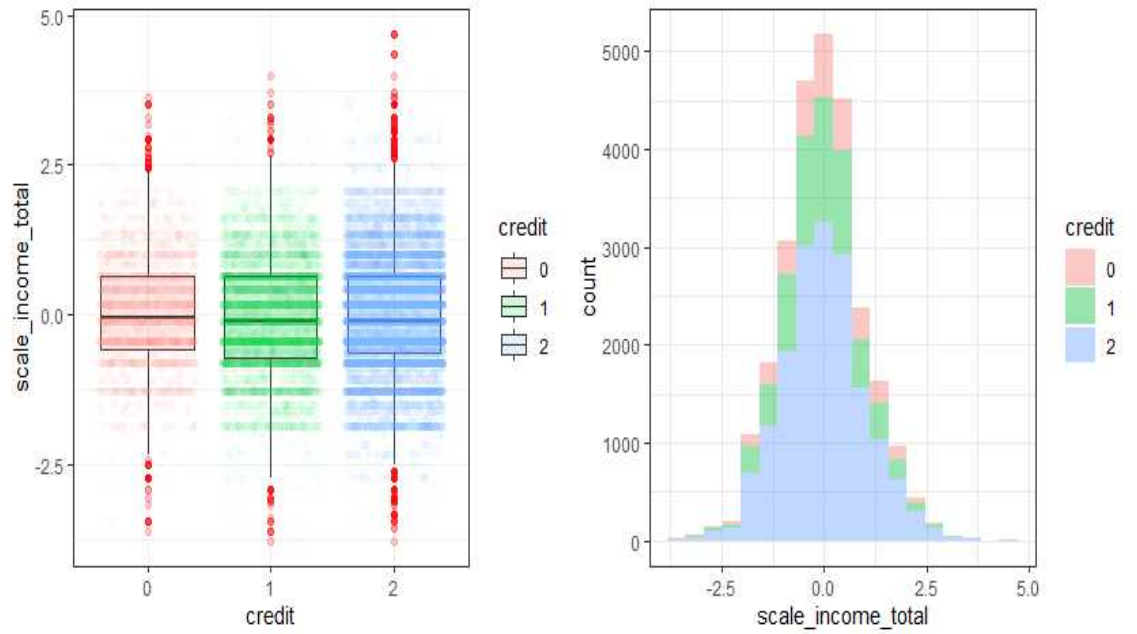
신용카드 고객을 변수별로 그룹을 지어보았을 때 대부분 하나의 그룹에 관측치가 몰려있는 것을 알 수 있다. 또한 그룹별 Credit의 비율이 비슷하기 때문에 의미 있는 변수를 찾기 어렵다.

6. EDA (연속형 변수 박스 플롯, 히스토그램)

총 소득

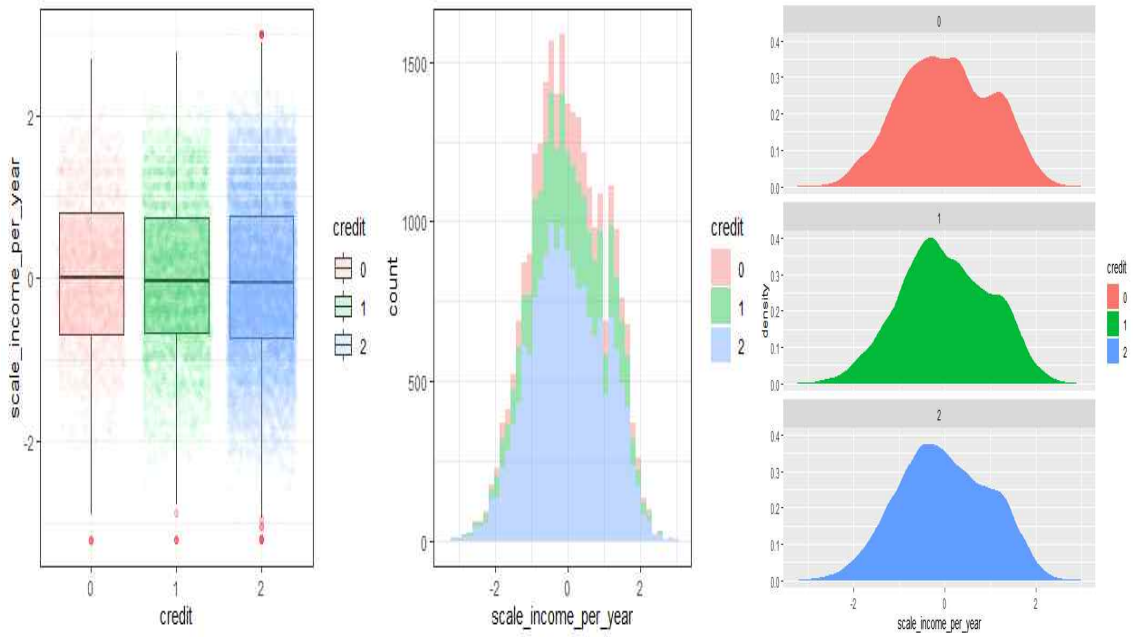


총소득 변수의 이상치가 1529개로 많고 왜도(Skewness)가 2.65912로 로그변환이 필요하다고 판단했다. 총소득 그래프의 정규성을 높이기 위해 자연로그를 취해 극단적인 이상치를 완화시켜 다루었다.



6. EDA (연속형 변수 박스 플롯, 히스토그램)

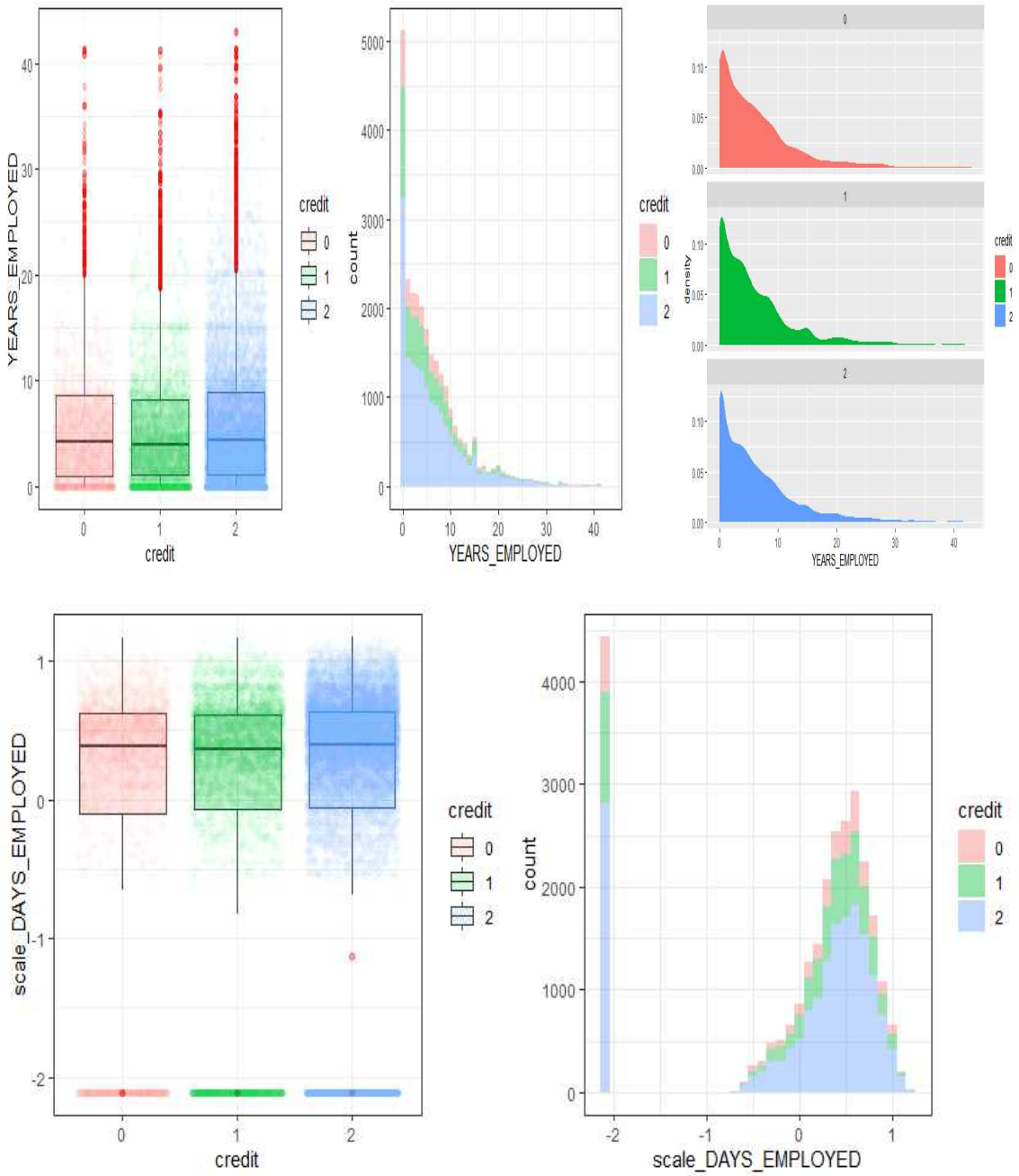
근속 년수 당 총 소득



총소득과 근속 연수가 유의미한 관계를 보이기 때문에 다양한 측면에서 데이터 분석을 위해 파생 변수를 생성했다. DAYS_EMPLOYED의 변수를 단독으로 사용했을 때보다 파생변수의 이상치가 훨씬 적은 것을 확인할 수 있다.

6. EDA (연속형 변수 박스 플롯, 히스토그램)

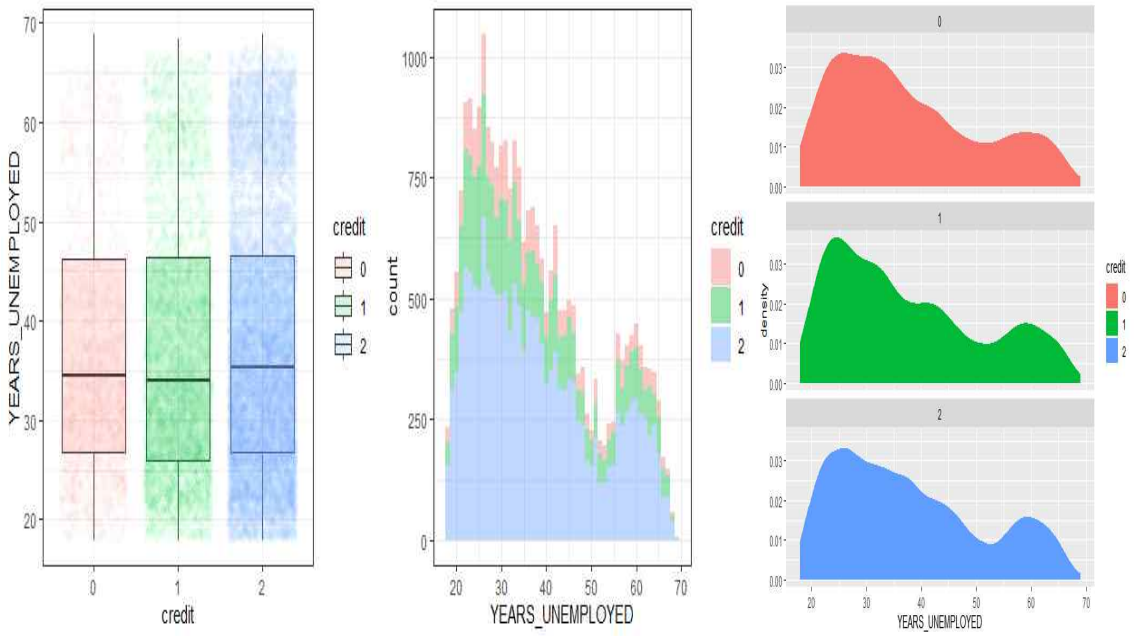
근속 년수



초기 DAYS_EMPLOYED의 이상치(36457개)를 0으로 처리했으나 오히려 이상치(1770개)가 부가적으로 생성되었다. 이후 정규성을 높이기 위한 방안으로 로그변환을 시도했지만 이상치 처리가 불가하여 여러 가지 파생 변수를 생성한 후 이상치를 별도로 다루기로 하였다.

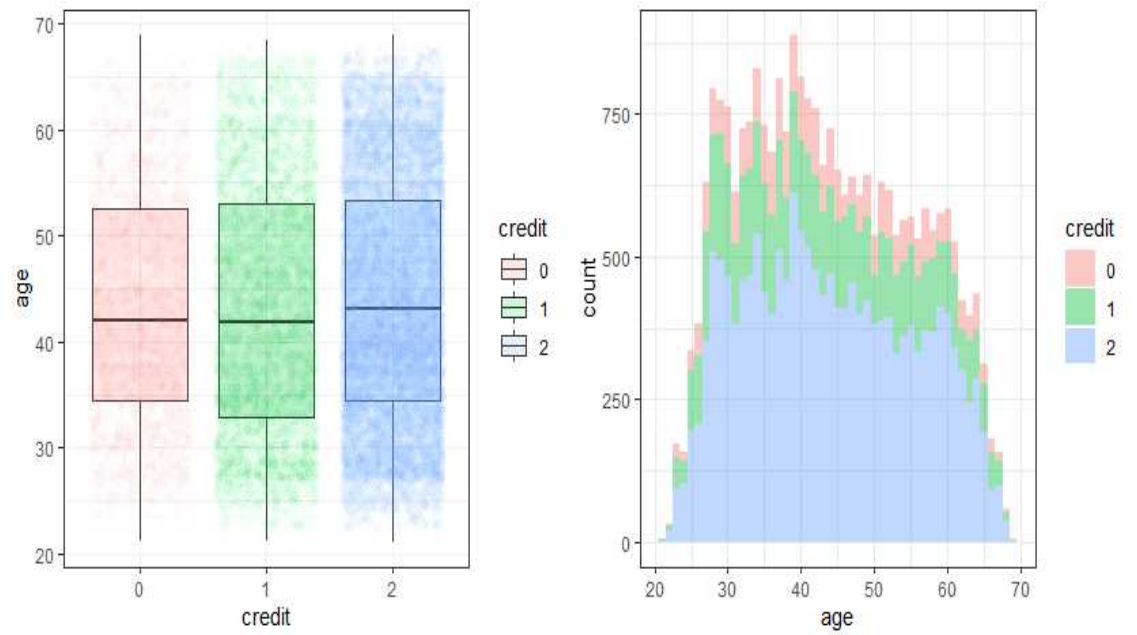
6. EDA (연속형 변수 박스 플롯, 히스토그램)

실업 기간



DAYS_EMPLOYED와 DAYS_BIRTH 변수 사이에서 실업 기간(YEARS_UNEMPLOYED)이라는 파생 변수를 생성했다. 그 후 다중공선성 제거를 위해 DAYS_EMPLOYED와 DAYS_BIRTH 중 한 변수를 제거할 것을 고려했다.

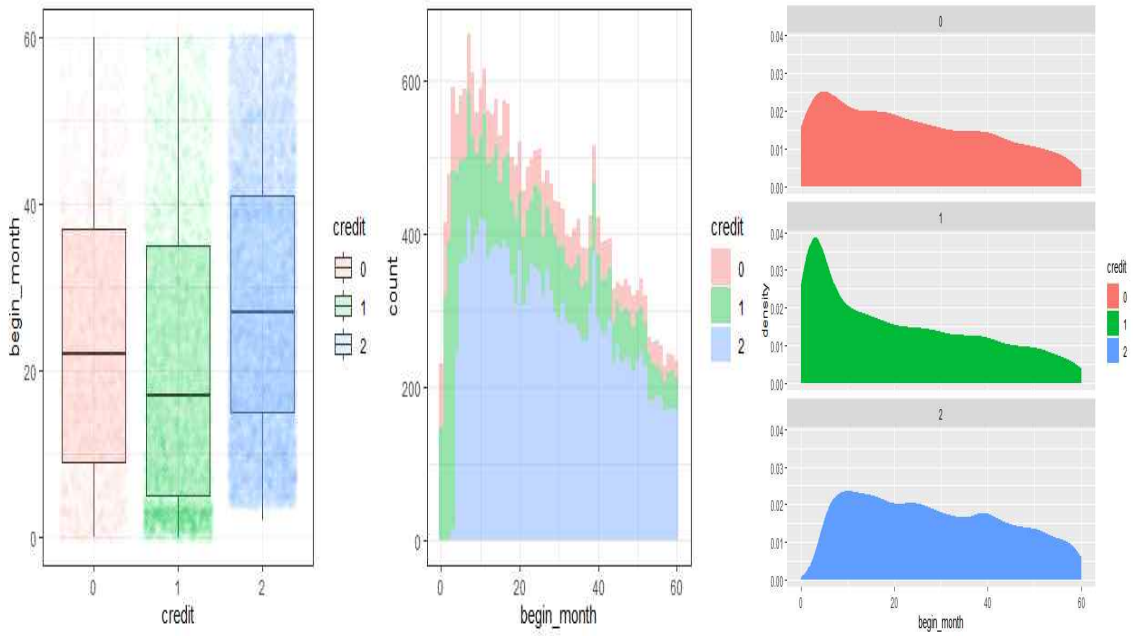
나이



집단 간의 평균과 분산에 큰 차이가 없어서 무의미할 것으로 예상한다.

6. EDA (연속형 변수 박스 플롯, 히스토그램)

카드 이용 기간



[비즈니스적 관점]

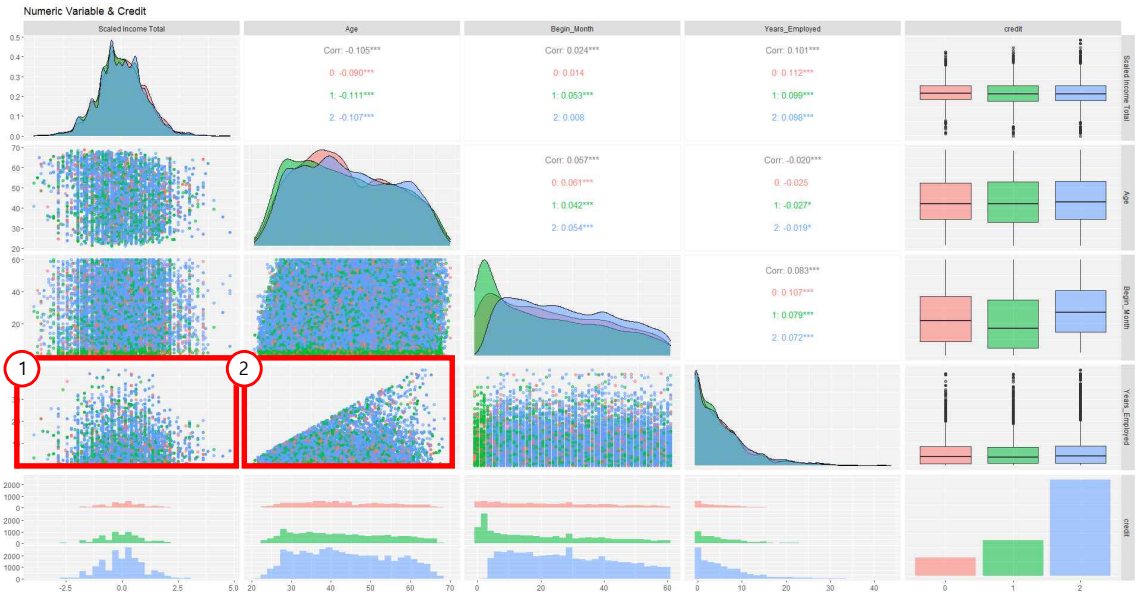
카드 발급일 초기에는 신용도 1의 밀도가 높다. 카드 이용일이 증가할수록 모든 신용도에서 카드 이용자 수가 줄어들고 있다. 이탈 고객이 늘어난다는 비즈니스적 해석이 가능하다.

[통계적 관점]

집단 간의 평균 차이가 확실하게 보이기 때문에 유의미할 것으로 보인다.

6. EDA (상관관계)

[주요 연속형 변수] 산점도

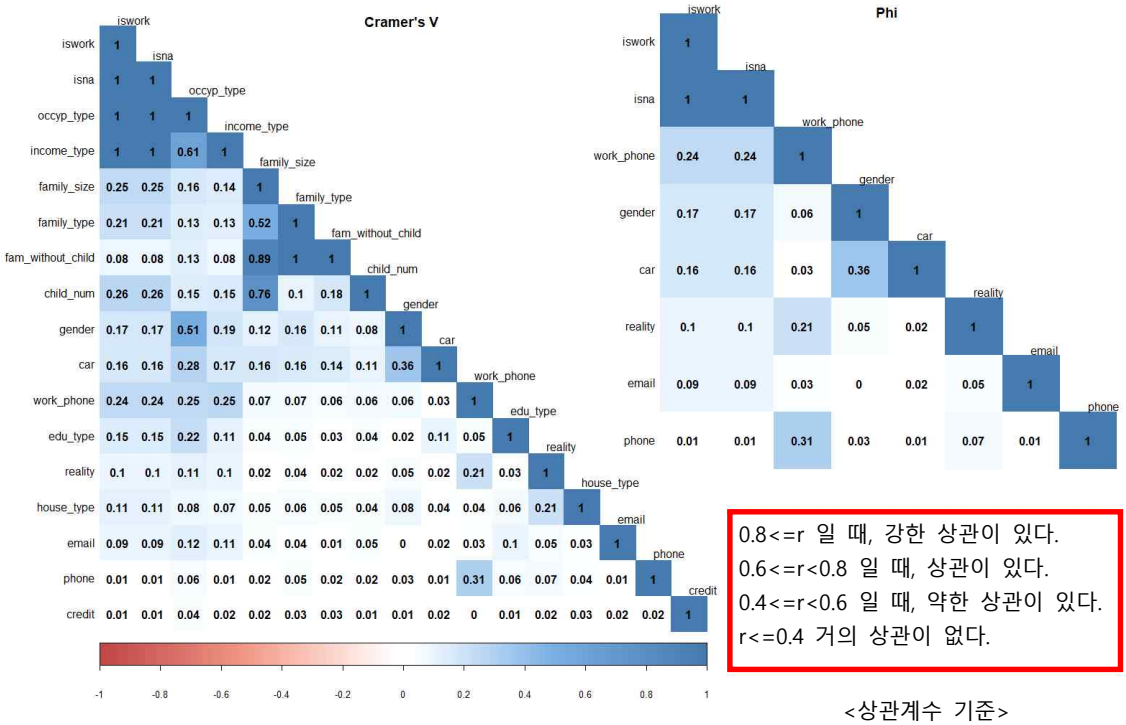


①번 그래프는 근속 년수가 증가할수록 소득이 평균으로 수렴한다. ②번 그래프는 근속 년수와 나이 간의 뚜렷한 한계선이 존재한다. 나이에 따른 소득, 카드 이용 기간에 따른 소득, 나이에 따른 카드 이용기간, 근속 년수에 따른 카드 이용 기간은 뚜렷한 의미를 발견하지 못했다.

6. EDA (상관관계)

[범주형 변수] Cramer's V & Phi

Cramer's V와 Phi 계수는 두 개의 명목척도간의 상관관계를 구하는 분석 방법이다. Phi 계수는 두 변수가 모두 두 개의 범주일 경우 사용하나 Cramer's V 계수는 적어도 한 변수가 세 개 이상의 범주를 가지는 경우 사용한다. 각각의 계수는 0~1의 범위를 갖고 두 변수가 상관이 없을 때 0, 관련이 있을수록 1에 가까워진다.

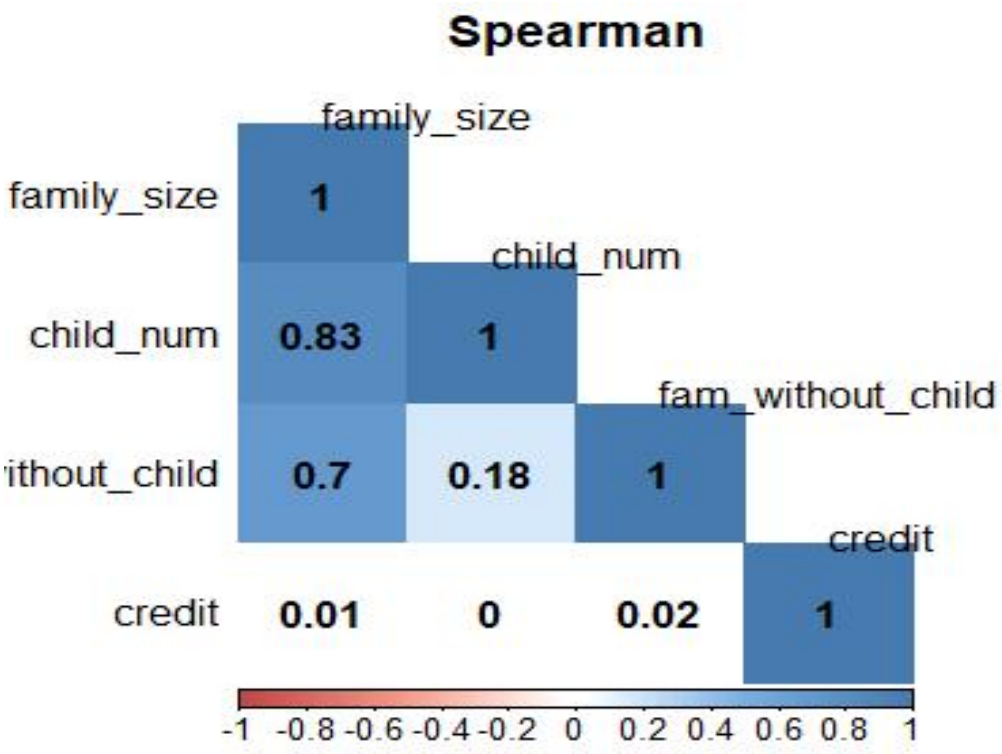


<상관계수 기준>

종속변수인 credit(0, 1, 2값을 가진다.)과의 관계를 확인하기 위해 <상관계수 기준>으로 모든 명목형 변수의 Cramer's V 계수를 확인했다. 모든 독립 변수와 종속 변수 간 선형관계가 없다. 다중공선성을 제거하기 위해 family_size, family_without_child, iswork, isna, income_type, occyp_type 제거를 고려했다. 또한 범주가 두 개인 변수를 별도로 선별하여 Phi 계수를 확인했다.

6. EDA (상관관계)

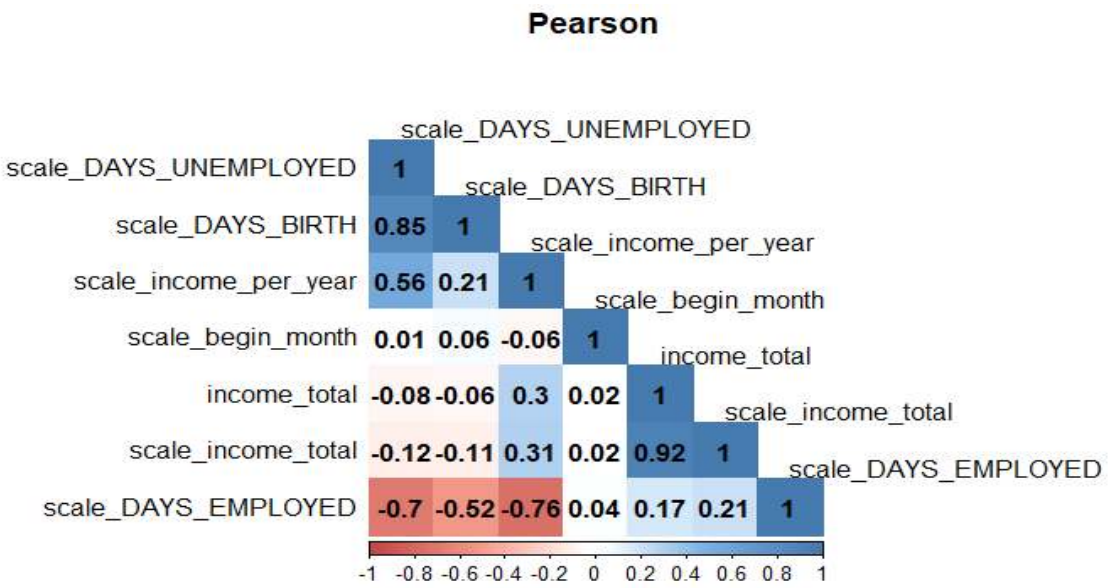
[순서형 변수] 스피어만 상관계수



순서형 변수끼리의 다중공선성을 확인하기 위해서 Cramer's V와 동일한 기준으로 스피어만 상관 계수를 이용했다. 위와 같은 기준을 따를 때 family_size, child_num, family_size와 without_child간의 다중공선성이 존재하므로 이를 해소하기 위해 family_size 제거 고려한다. 각 순서형 변수와 종속변수(credit)간에 선형 관계가 없다.

6. EDA (상관관계)

[연속형 변수] 피어슨 상관계수



피어슨 상관계수 또한 Cramer's V와 동일한 기준으로 scale_DAYS_UNEMPLOYED, scale_DAYS_EMPLOYED, income_total 제거를 고려한다.

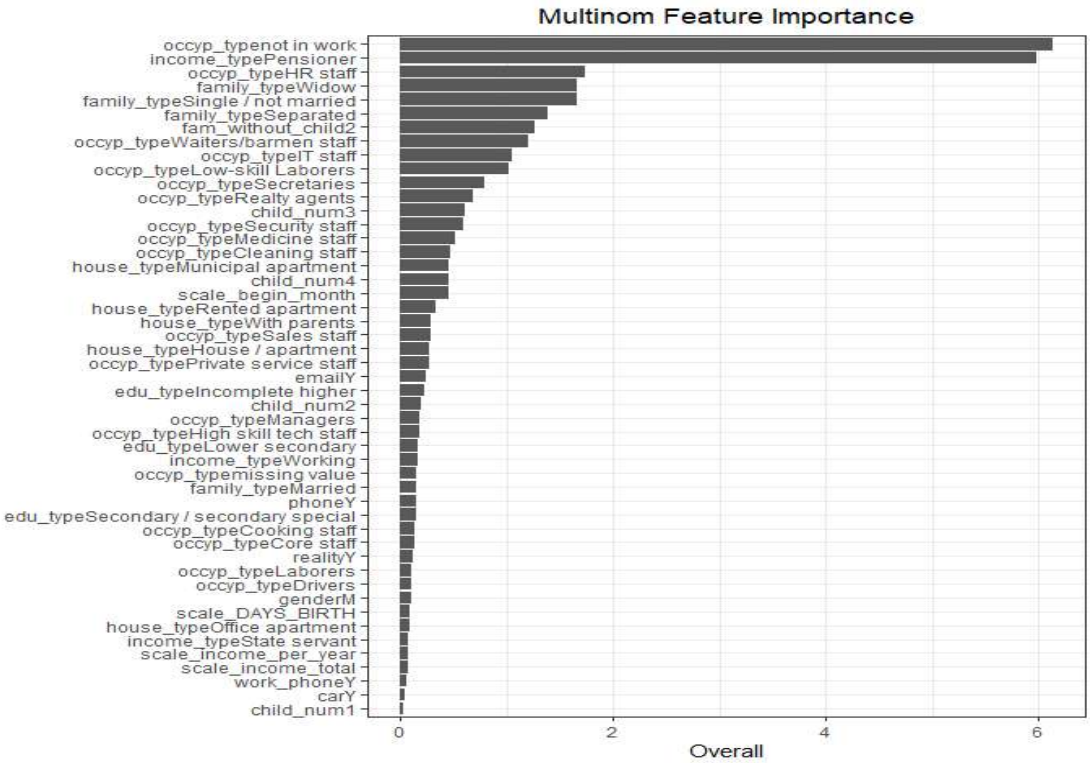
7. 변수 선별

변수 선별 모델

model	설명
Logistic Regression	종속 변수와 독립 변수 간의 선형 관계를 요구하지 않는다. 예측된 교차비에 비선형 로그 변환을 적용하기 때문에 다양한 유형의 관계를 처리할 수 있다. 종속 변수가 다중클래스이므로 다항 로지스틱 회귀
Stepwise	각 변수의 OLS 결과를 보고, 직접 p-value와 F통계량을 비교하며 추가/삭제하는 통계적 변수 선택법
Random Forest	여러개의 의사결정트리를 취합하여 학습 성능을 높이는 앙상블 모형, 각 나무에서 데이터와 변수 모두 랜덤으로 선택하기 때문에, 각 나무가 독립적으로 다양하고 풍부한 표현력을 갖는 강점
XGBoost	feature의 중요도 스코어를 내는 것이 상대적으로 쉽다는 gradient boosting의 장점을 가장 잘 살릴 수 있는 모델

7. 변수 선별

다중 로직스틱 회귀 & 단계적 방법

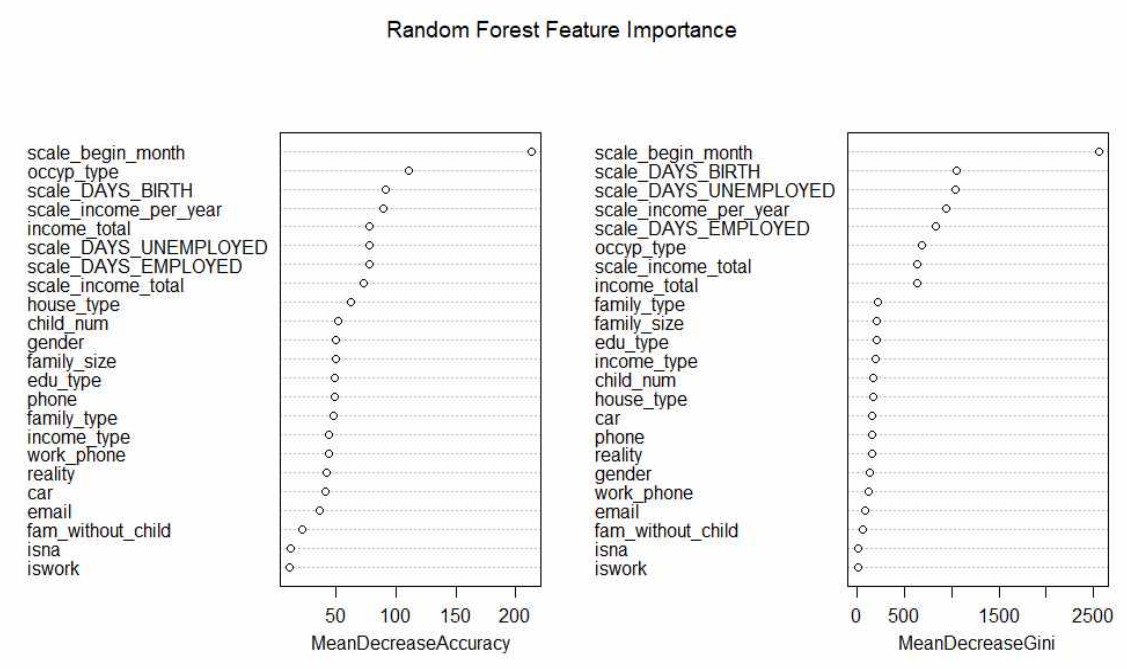


	Df	AIC
<none>	84	45622.82
- fam_without_child	82	45623.11
- phone	82	45623.26
+ +work_phone	86	45624.61
+ +scale_income_per_year	86	45625.09
+ +car	86	45625.23
+ +gender	86	45625.97
- edu_type	78	45627.27
- family_type	76	45628.67
- email	82	45628.74
+ +child_num	92	45629.11
- house_type	74	45629.53
- reality	82	45631.03
- occyp_type	46	45632.80
- scale_income_total	82	45635.30
- income_type	78	45638.27
- scale_DAYS_BIRTH	82	45640.39
- scale_begin_month	82	46548.14

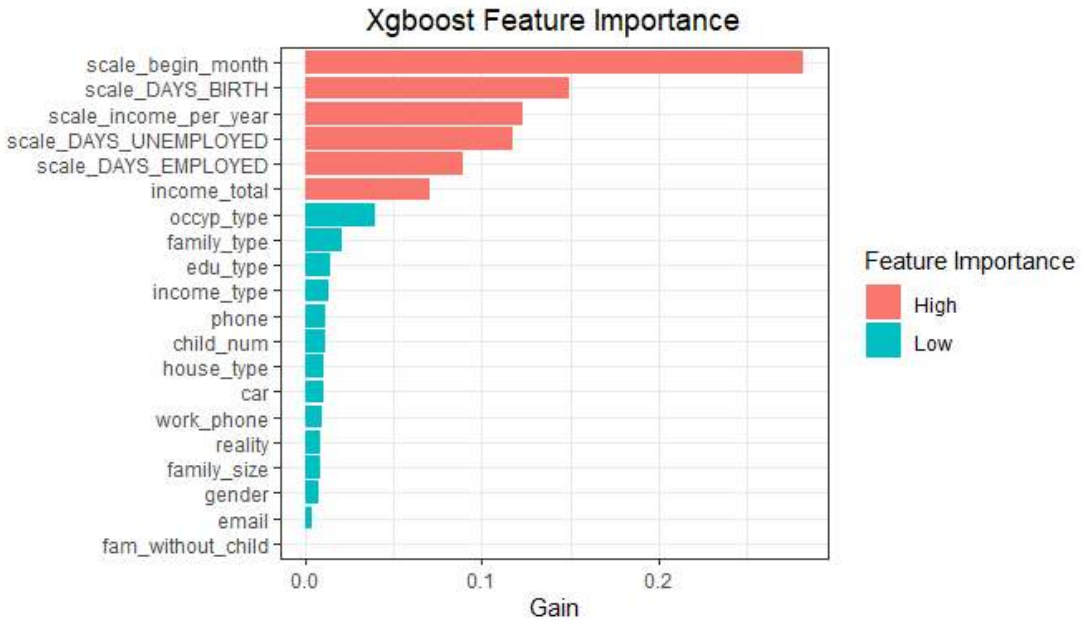
로지스틱 회귀 모형의 정확도가 64.31%로 낮으며 그에 따라 단계적 방법 또한 변수 선별법으로 사용하기 어렵다.

7. 변수 선별

Random Forest



Xgboost



7. 변수 선별

1_ 최종 모델 학습 전 변수 제거
2_ 다중공선성 제거
3_ 변수 선별을 통한 가장 중요한 변수만 추출, 10개

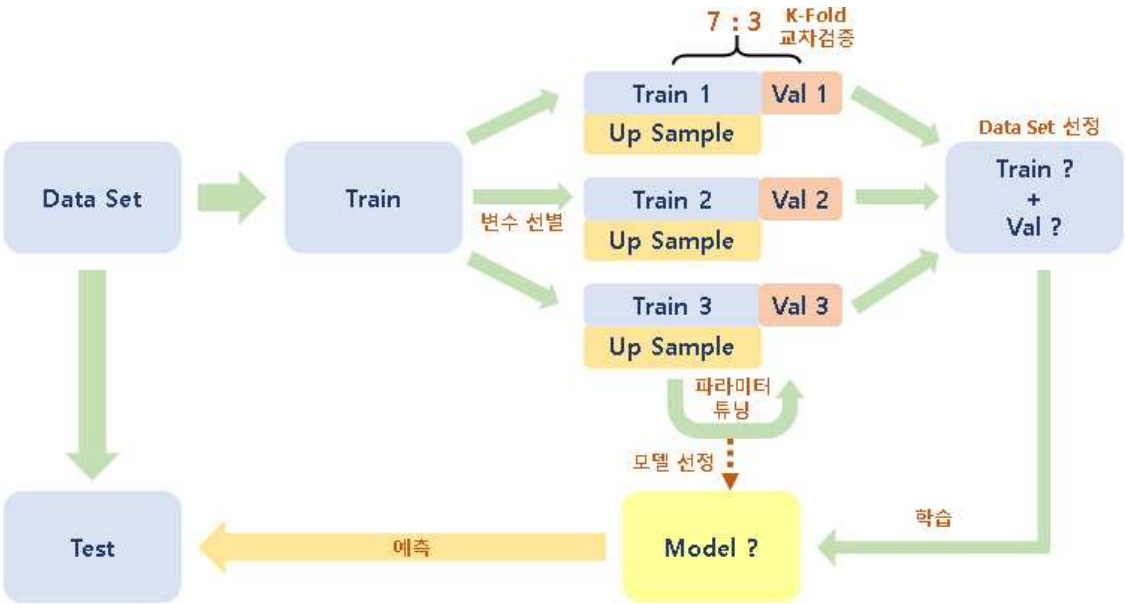
car	child_num	income_type	edu_type	family_type
occyp_type	scale_income_total	scale_DAYS_BIRTH	scale_begin_month	scale_begin_month"

학습 모델

model	설명
SVM	다양한 데이터 분포에서도 잘 작동하는 분류 기법 중 최상의 기법 다중 분류 목적에 부합
Random Forest	여러개의 의사결정트리를 취합한 앙상블 모형, 각 나무가 독립적으로 다양하고 풍부한 표현력
XGBoost	뛰어난 예측 성능과 GBM 대비 빠른 수행 시간, 과적합 규제, 가지치기, 자체 내장된 교차 검증으로 높은 분류 성능 기대
ANN	범주형 데이터에 대해서도 다른 별도의 작업 없이 간단하게 분류
CatBoost	범주형 데이터를 자동으로 전처리하여 모델 튜닝이 간소화

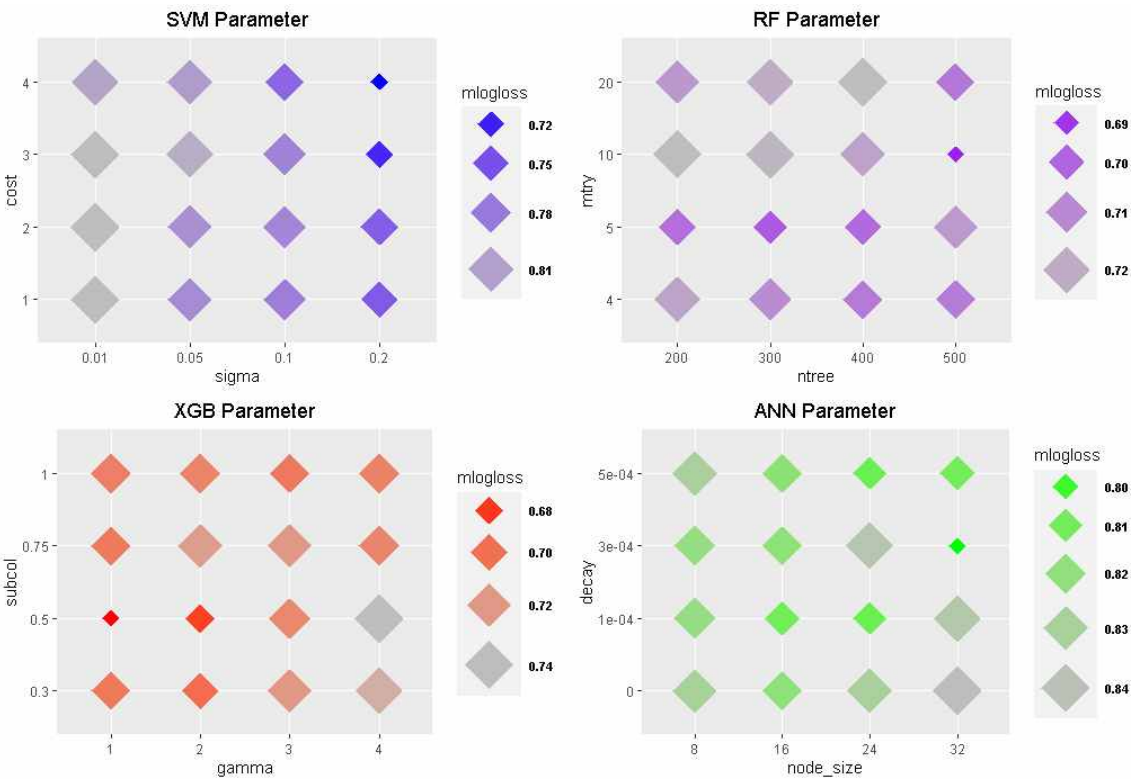
8. 모델 학습

모델 학습 과정



Train 1 : 모든 파생변수를 포함, 총 23개의 변수를 담은 Data Set
Train 2 : 중요하지 않은 변수를 제거, 총 13개의 변수를 담은 Data Set
Train 3 : 다중공선이 없는 중요한 변수를 포함, 총 10개의 변수를 담은 Data Set

모델 파라미터 Grid Search



각 모델별로 가장 낮은 Multi Logloss를 가지는 파라미터를 선택함

8. 모델 학습

SVM

sigma=0.2, cost=4, kernal="Gaussian"

RandomForest

ntree= 500, mtry= 10, nodesize= 10

Xgboost

booster="gbtree", eta=0.08, max_depth=7, gamma=1
subsample=0.75, col_sample=0.5

ANN

size=32, decay=0.0003, maxit=1000

Catboost

learning rate=0.05 (별도의 튜닝을 하지 않음)

9. 모델 평가 및 결과 (Recovery Score)

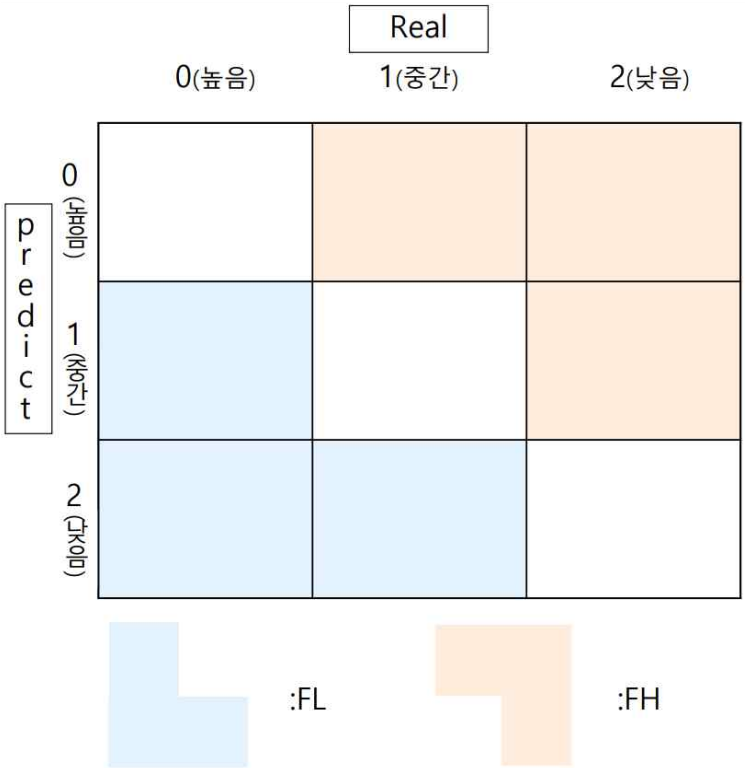
RC(Recovery Score) : FL / (FL+FH)

배경 : 카드사의 입장에서 신용도를 잘못 예측하는 상황 중에서 실제로 낮은 신용도를 높은 신용도로 예측했을 경우에 리스크가 실제로 높은 신용도를 낮은 신용도로 예측하는 것에 비해 상대적으로 더 높다.

생성방법 : 학습 모델의 예측치와 실제값 간의 Cross Table을 생성하여 FL / FH + FL 값을 계산한다.

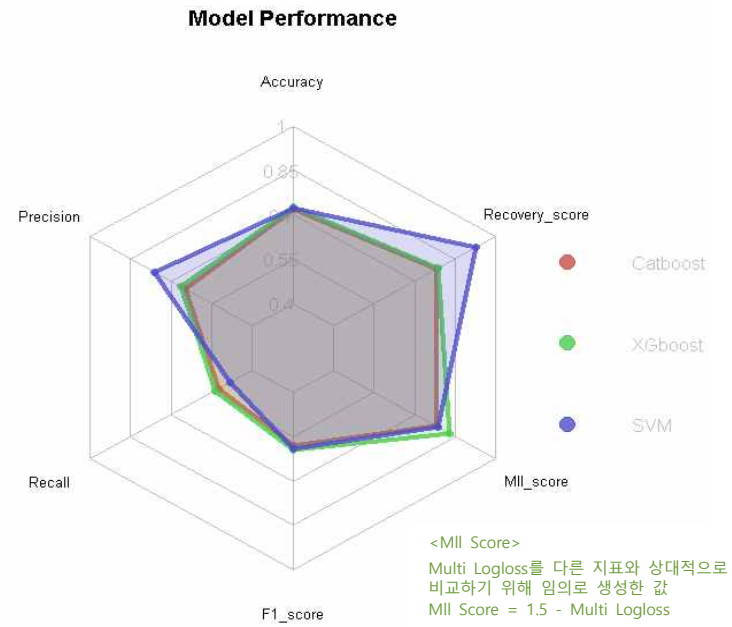
FH (FALSE HIGH : 고객 신용도가 과대평가 된 경우)

FL (FALSE LOW : 고객 신용도가 과대평가 된 경우)



9. 모델 평가 및 결과

Model	Main Performance		Evaluation Indicator	
SVM	Accuracy	0.7241	Precision	0.7613
			Recall	0.4840
	Recovery Score	0.9276	F1	0.5917
			Mll Score	0.7860
			AUC	0.7981
Random Forest	Accuracy	0.7247	Precision	0.6617
			Recall	0.5341
	Recovery Score	0.7824	F1	0.5912
			Mll Score	0.7977
			AUC	0.7759
XGBoost	Accuracy	0.7278	Precision	0.6659
			Recall	0.5395
	Recovery Score	0.7867	F1	0.5961
			Mll Score	0.8274
			AUC	0.7605
CatBoost	Accuracy	0.7187	Precision	0.6472
			Recall	0.5244
	Recovery Score	0.7828	F1	0.5793
			Mll Score	0.7801
			AUC	0.7561
ANN	Accuracy	0.6788	Precision	0.5107
			Recall	0.4231
	Recovery Score	0.8682	F1	0.4628
			Mll Score	0.6754
			AUC	0.6452



<Mll Score>
Multi Logloss를 다른 지표와 상대적으로 비교하기 위해 임의로 생성한 값
Mll Score = 1.5 - Multi Logloss

비즈니스적 관점에서 카드사의 관심사인 리스크 최소화에 부합하기 위한 지표를 생성했다. 카드사의 입장에서 신용도가 낮은 고객에 대해서 높게 측정할 경우 리스크가 가장 크다고 판단했다. 평과 결과는 왼쪽 그림과 같이 비슷한 수치를 보이지만 RC가 가장 높은 SVM을 최종 모델로 선정했다.

SVM을 최종 모델로 선정

10. 결론

프로젝트 의의

기존에 존재하는 신용도의 데이터를 기반으로 한 분석의 한계를 넘어서 미래 가능성을 예측할 수 있었음

데이터 분석 한계

중복되는 데이터 다수 존재, 변수의 다양성 부재

생성 모형 한계

정확도가 높지 않음

활용 방안

금융권의 마이데이터 산업에 적용할 수 있을 것으로 기대

소감

기존의 머신러닝 분석에 더해서 통계적 기법을 활용한 분석을 시도해봄으로써 좀 더 설명력있고 설득력있는 모델 설계가 가능했다. 금융권 기업의 입장에서 마이데이터 산업의 축소판 프로젝트를 경험함으로써 현업에서 동일한 종류의 프로젝트 진행 시 이번 프로젝트를 바탕으로 더 체계적인 진행이 가능할 것으로 예상된다