2022 빅콘테스트

데이터 분석 리그_퓨처스

Honglk

김태용, 권재현, 박세빈, 한혜원, 홍표민

CONTENTS

1. 문제 이해

01. 대출 신청 고객 예측

사용자 신용 정보

유저 연소득, 고용 형태, 기대출수

주거 소유 형태, 대출 목적 , ...

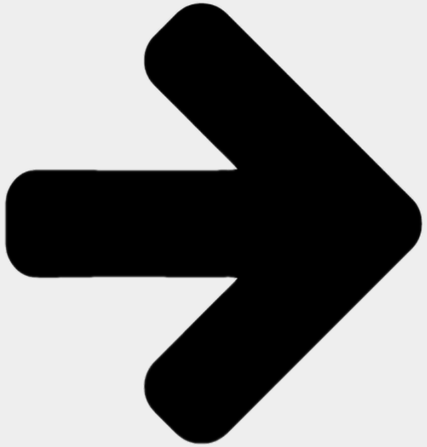
finda App 로그 정보

유저 번호, 행동명,

행동일시, 일 코드

사용자가 신청한 대출별 금융사별 승인결과

신청서 번호, 한도조회 일시, 금융사 번호, 상품 번호, 승인한도, 승인금리

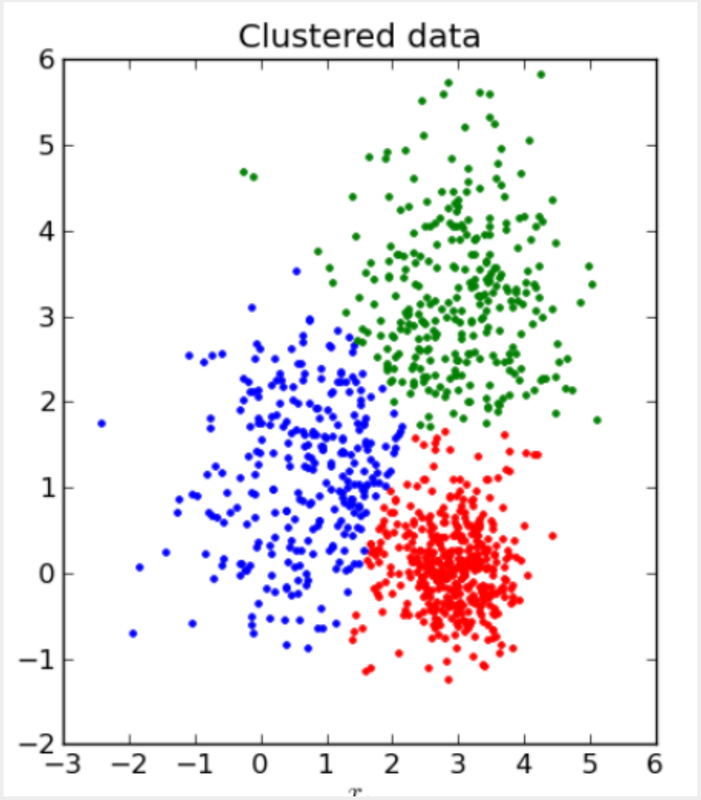


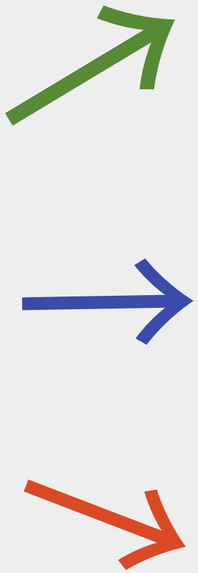
대출 신청 여부

신청 여부(예측 레이블)

02. 모델 기반 군집 분석 및 서비스 메시지 제안

모델 기반으로 사용자를 군집화 한 후 각 군집에 적합한 서비스 메시지 작성





각 군집의 특성을 파악하고

군집 별 특성에 적합한 서비스 메시지 작성

2. 데이터 전처리

01.결측치 처리

user_spec.csv, loan_result.csv의 data가 대출 신청 여부에 더 큰 영향을 끼치고, 제한된 하드웨어 리소스에서 최대 성능을 내기 위해 log_data.csv를 제외한 나머지 두 파일을 user_id 기준으로 병합

· 의미 중복 제거

user_id	birth_year	gender	loan_limit	yearly_income	income_type	employment_type	purpose	loanapply_insert_time	personal_rehabilitation_yn	personal_rehabilitation_complete_yn	existing_loan_cnt	existing_loan_amt	credit_score
341318	1971	1	15000000	30000000	EARNEDINCOME	계약직	생활비	6/21/2022 8:31:13	null	null	4	20000000	710
524359	1976	1	45000000	42000000	EARNEDINCOME	정규직	'living'	8/19/2022 8:23:11	null	null	1	3000000	590
733387	1999	1	43900000	30000000	EARNEDINCOME2	정규직	'switchloan'	1/2/2022 6:45:15	null	null	null	null	580
801057	1993	0	17000000	48000000	EARNEDINCOME	정규직	대환대출	5/13/2022 7:52:48	null	null	0	0	660

'생활비', '대환대출', '사업자금', '기타', '전월세보증금', '주택구입', '투자', '자동차구입', 'living', 'switchloan', 'business', 'etc', 'housedeposit', 'buyhouse', 'invest', 'buycar' 등 중복되는 의미 제거

· 자료형 정렬

user_id	birth_year	gender	loan_limit	yearly_income	income_type	employment_type	purpose	loanapply_insert_time	personal_rehabilitation_yn	personal_rehabilitation_complete_yn	existing_loan_cnt	existing_loan_amt	credit_score
341318	1971	1	15000000	30000000	EARNEDINCOME	계약직	생활비	6/21/2022 8:31:13	null	null	4	20000000	710
524359	1976	1	45000000	42000000	EARNEDINCOME	정규직	'living'	8/19/2022 8:23:11	null	null	1	3000000	590
733387	1999	1	43900000	30000000	EARNEDINCOME2	정규직	'switchloan'	1/2/2022 6:45:15	null	null	null	null	580
801057	1993	0	17000000	48000000	EARNEDINCOME	정규직	대환대출	5/13/2022 7:52:48	null	null	0	0	660

머신 러닝을 위해 문자열을 One-hot-Encoding을 통해 숫자열로 정리 & 시간형태의 데이터를 각각 연, 월, 일, 시로 열 추가

2. 데이터 전처리

01.결측치 제거

· 변수 사이의 관계: 개인회생자 여부와 개인 회생자 납입 완료 여부 사이의 관계 확인

user_id	birth_year	gender	loan_limit	yearly_income	income_type	employment_type	purpose	loanapply_insert_time	personal_rehabilitation_yn	personal_rehabilitation_complete_yn	existing_loan_cnt	existing_loan_amt	credit_score
341318	1971	1	15000000	30000000	EARNEDINCOME	계약직	생활비	6/21/2022 8:31:13	null	null	4	20000000	710
524359	1976	1	45000000	42000000	EARNEDINCOME	정규직	'living'	8/19/2022 8:23:11	null	null	1	3000000	590
733387	1999	1	43900000	30000000	EARNEDINCOME2	정규직	'switchloan'	1/2/2022 6:45:15	null	null	null	null	580
801057	1993	0	17000000	48000000	EARNEDINCOME	정규직	대환대출	5/13/2022 7:52:48	null	null	0	0	660

1. 개인 회생자가 아닐 때, 회생자 납입 완료 여부가 전부 결측치
2. 개인 회생자가 1일 때, 개인 회생자 납입 완료 여부의 결측치 존재 x
   1) 개인 회생자 여부가 0인 경우 회생자 납입 완료 여부 '2'로 대치
   2) 개인 회생자 여부의 결측치는 '2'로 대치 & 개인 회생자 여부가 결측치이고 납입 완료가 결측치인 경우 개인 회생자 납입 완료 여부의 결측치 '3'으로 대치
   3) 문자열로 변환 후 다시 One-hot-encoding 실시

· 결측치 제거

user_id	birth_year	gender	loan_limit	yearly_income	income_type	employment_type	purpose	loanapply_insert_time	personal_rehabilitation_yn	personal_rehabilitation_complete_yn	existing_loan_cnt	existing_loan_amt	credit_score
341318	1971	1	15000000	30000000	EARNEDINCOME	계약직	생활비	6/21/2022 8:31:13	null	null	4	20000000	710
524359	1976	1	45000000	42000000	EARNEDINCOME	정규직	'living'	8/19/2022 8:23:11	null	null	1	3000000	590
733387	1999	1	43900000	30000000	EARNEDINCOME2	정규직	'switchloan'	1/2/2022 6:45:15	null	null	null	null	580
801057	1993	0	17000000	48000000	EARNEDINCOME	정규직	대환대출	5/13/2022 7:52:48	null	null	0	0	660

결측치 수가 작았기 때문에 birth_year,gender, income_type, employment_type, houseown_type, desired_amount, purpose의 결측치 행 삭제

02.결측치 대치

· 평균값 대치

user_id	birth_year	gender	loan_limit	yearly_income	income_type	employment_type	purpose	loanapply_insert_time	personal_rehabilitation_yn	personal_rehabilitation_complete_yn	existing_loan_cnt	existing_loan_amt	credit_score
341318	1971	1	15000000	30000000	EARNEDINCOME	계약직	생활비	6/21/2022 8:31:13	null	null	4	20000000	710
524359	1976	1	45000000	42000000	EARNEDINCOME	정규직	'living'	8/19/2022 8:23:11	null	null	1	3000000	590
733387	1999	1	43900000	30000000	EARNEDINCOME2	정규직	'switchloan'	1/2/2022 6:45:15	null	null	null	null	580
801057	1993	0	17000000	48000000	EARNEDINCOME	정규직	대환대출	5/13/2022 7:52:48	null	null	0	0	660

상대적으로 결측치가 적었지만 중요한 feature라고 판단하여 평균값으로 대치

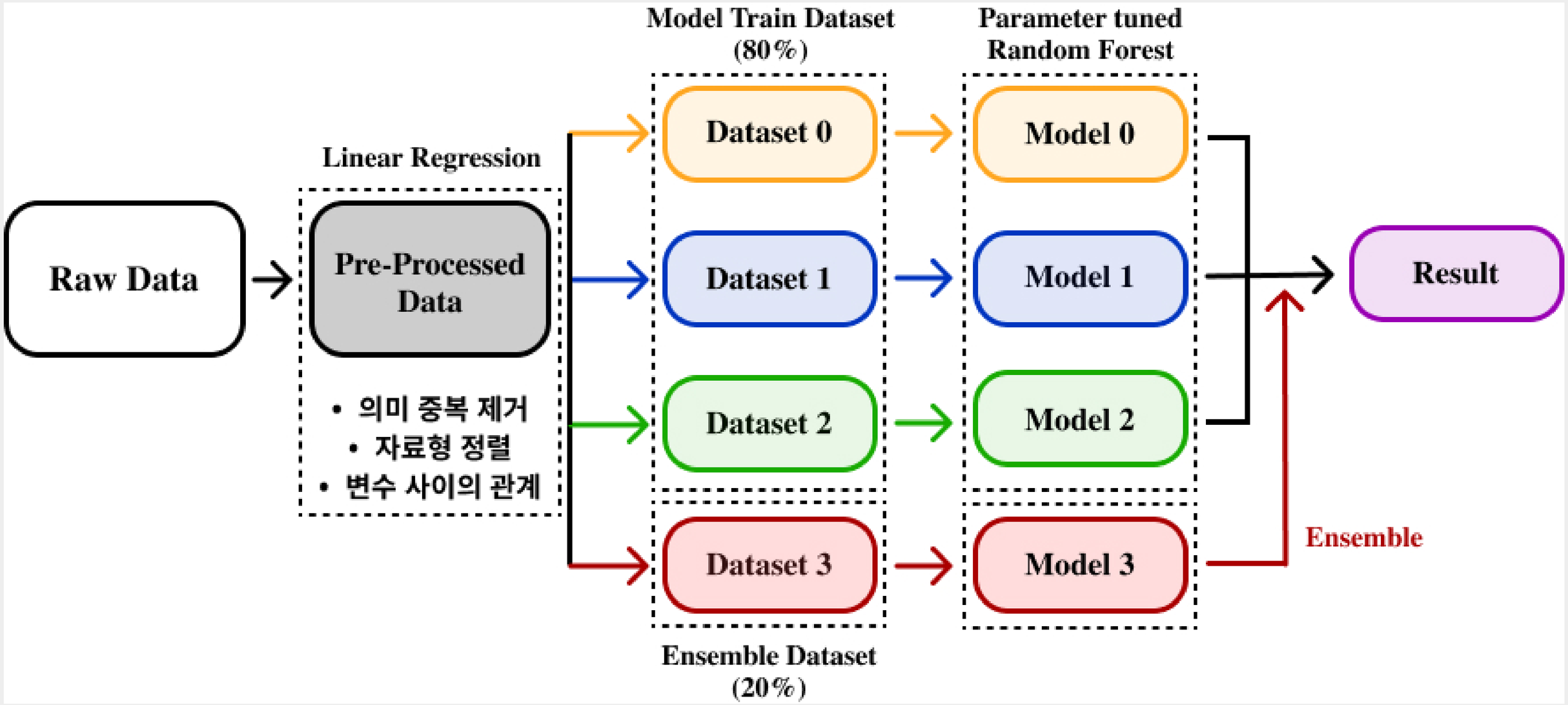
· 선형 회귀

user_id	birth_year	gender	loan_limit	yearly_income	income_type	employment_type	purpose	loanapply_insert_time	personal_rehabilitation_yn	personal_rehabilitation_complete_yn	existing_loan_cnt	existing_loan_amt	credit_score
341318	1971	1	15000000	30000000	EARNEDINCOME	계약직	생활비	6/21/2022 8:31:13	null	null	4	20000000	710
524359	1976	1	45000000	42000000	EARNEDINCOME	정규직	'living'	8/19/2022 8:23:11	null	null	1	3000000	590
733387	1999	1	43900000	30000000	EARNEDINCOME2	정규직	'switchloan'	1/2/2022 6:45:15	null	null	null	null	580
801057	1993	0	17000000	48000000	EARNEDINCOME	정규직	대환대출	5/13/2022 7:52:48	null	null	0	0	660

결측치의 수가 크고 대출 신청 여부와 관련이 있다고 생각한 feature(credit_score, existing_loan_cnt, existing_loan_amt)는 선형회귀 분석하여 값을 채움

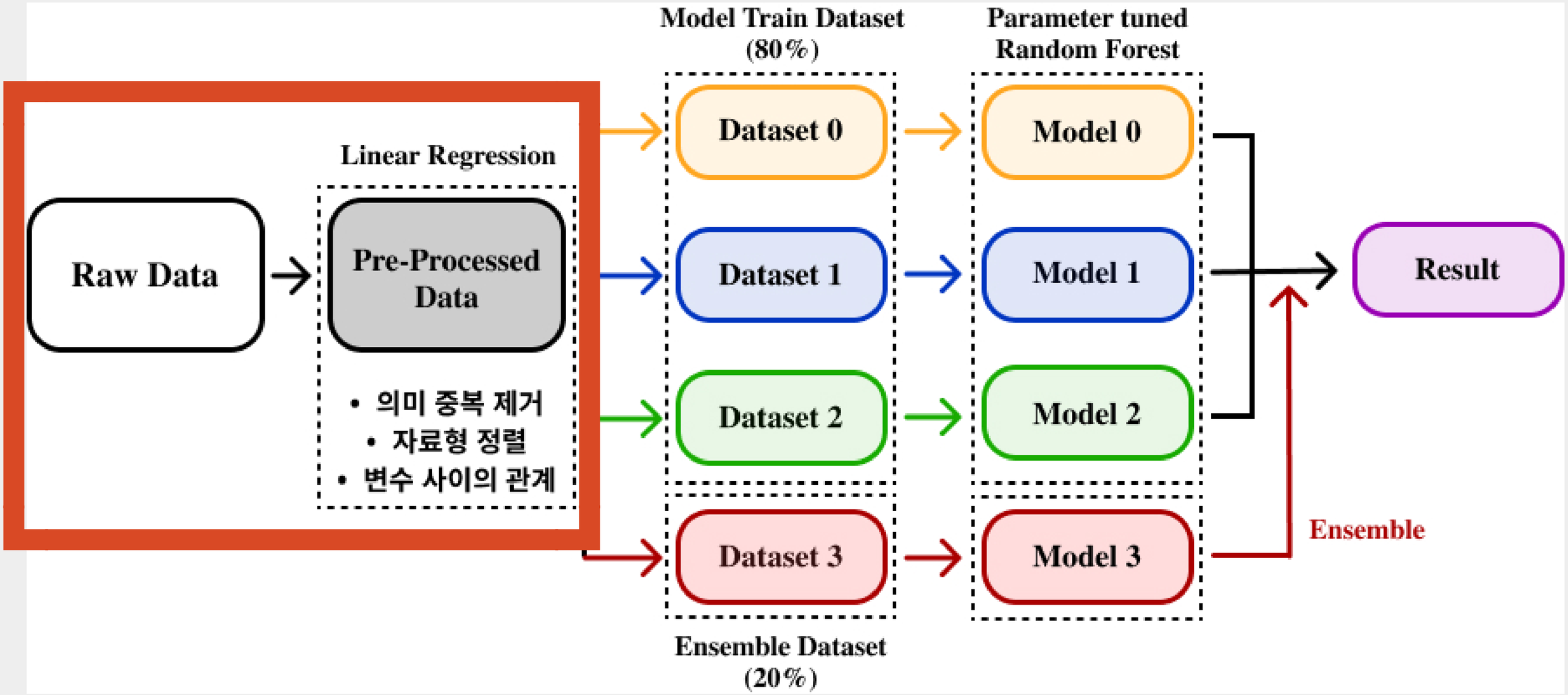
3. 모델링

01. 대출 신청 고객 예측 모델링: Overview



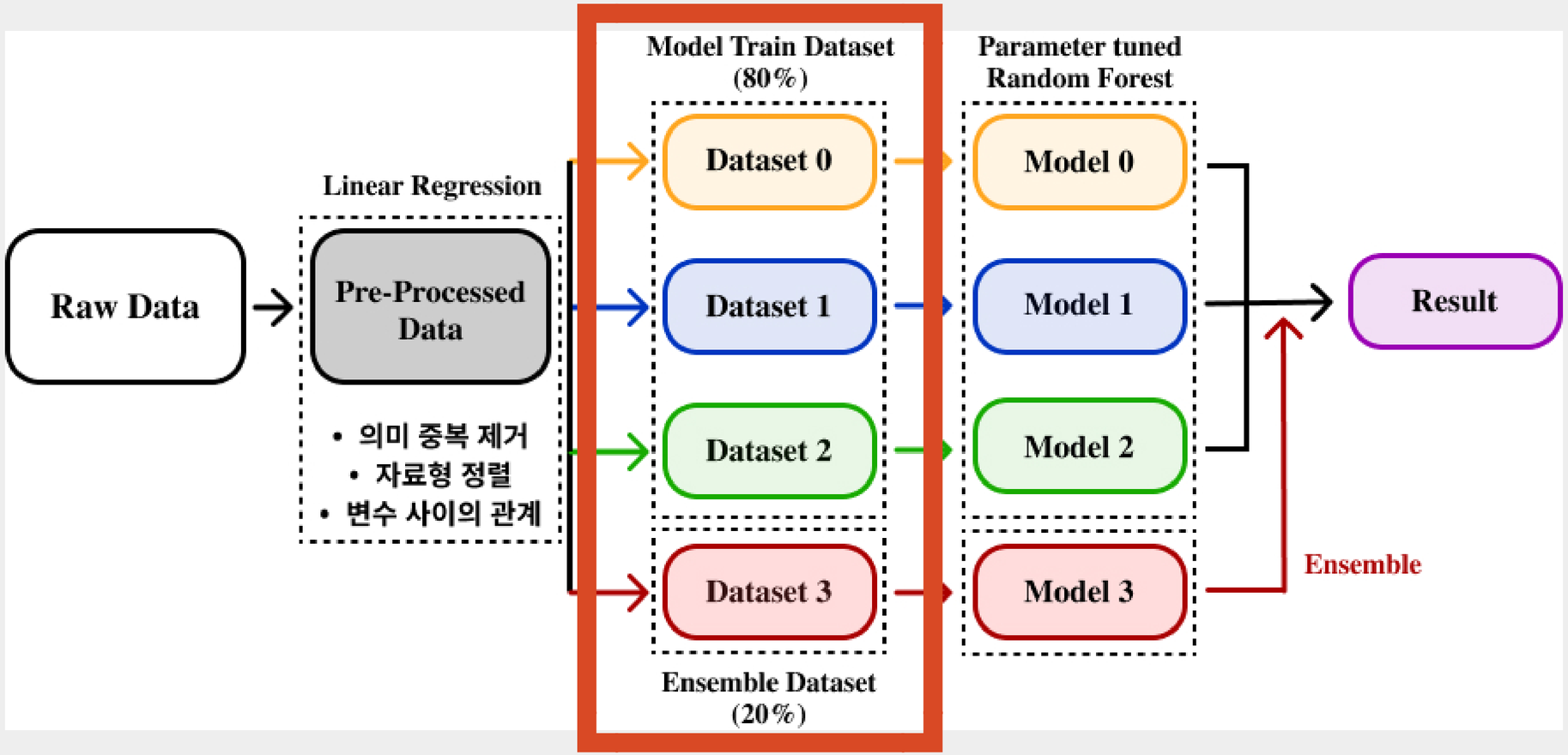
3. 모델링

01. 대출 신청 고객 예측 모델링: Data Pre-Processing



3. 모델링

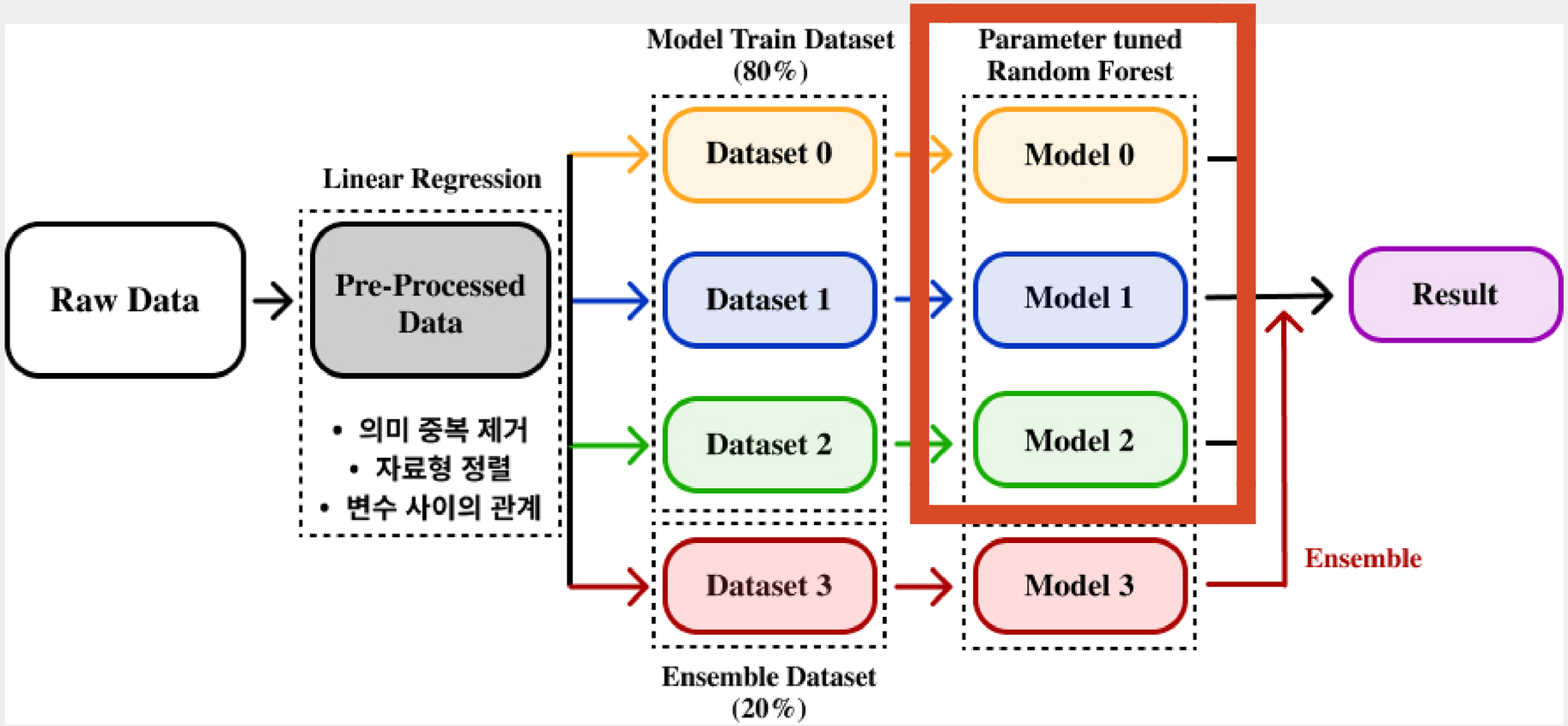
01. 대출 신청 고객 예측 모델링: Dataset Split Method



- 무작위(Random) 추출
- 비복원 (without replacement) 추출
- Training : Ensemble = 80:20

3. 모델링

01. 대출 신청 고객 예측 모델링: Training Method

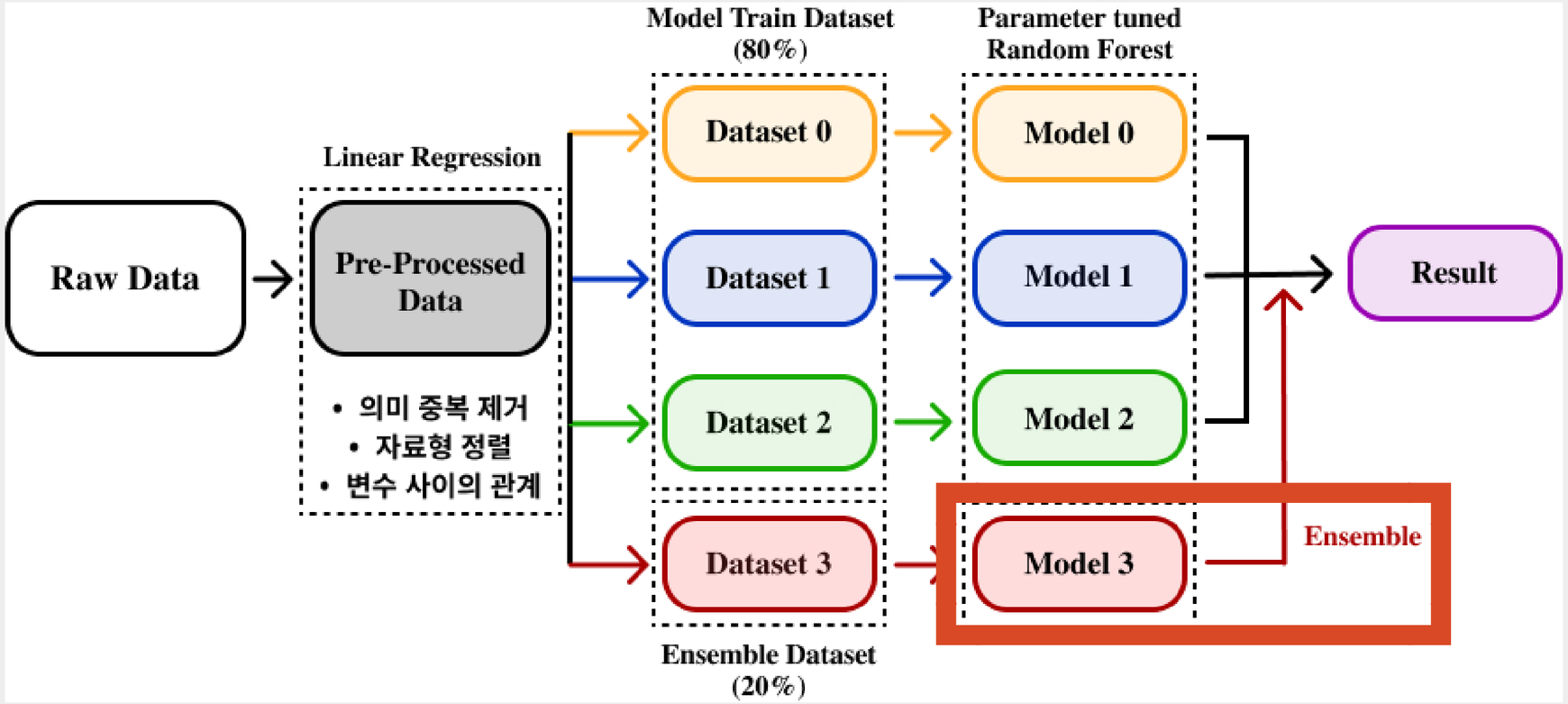


Model	성능 (Accuracy)
XGBoost	94.88
CatBoost	95.02
LightGBM	95.73
Random Forest	**96.77**

- 3개로 분리된 Training Dataset을 Parameter tuning 한 Random Forest 모델에 각각 학습시킴.
- Random Forest의 성능이 다른 알고리즘에 비해 약 1%p~2%p의 정확도가 높음.

3. 모델링

01. 대출 신청 고객 예측 모델링: Ensemble Method



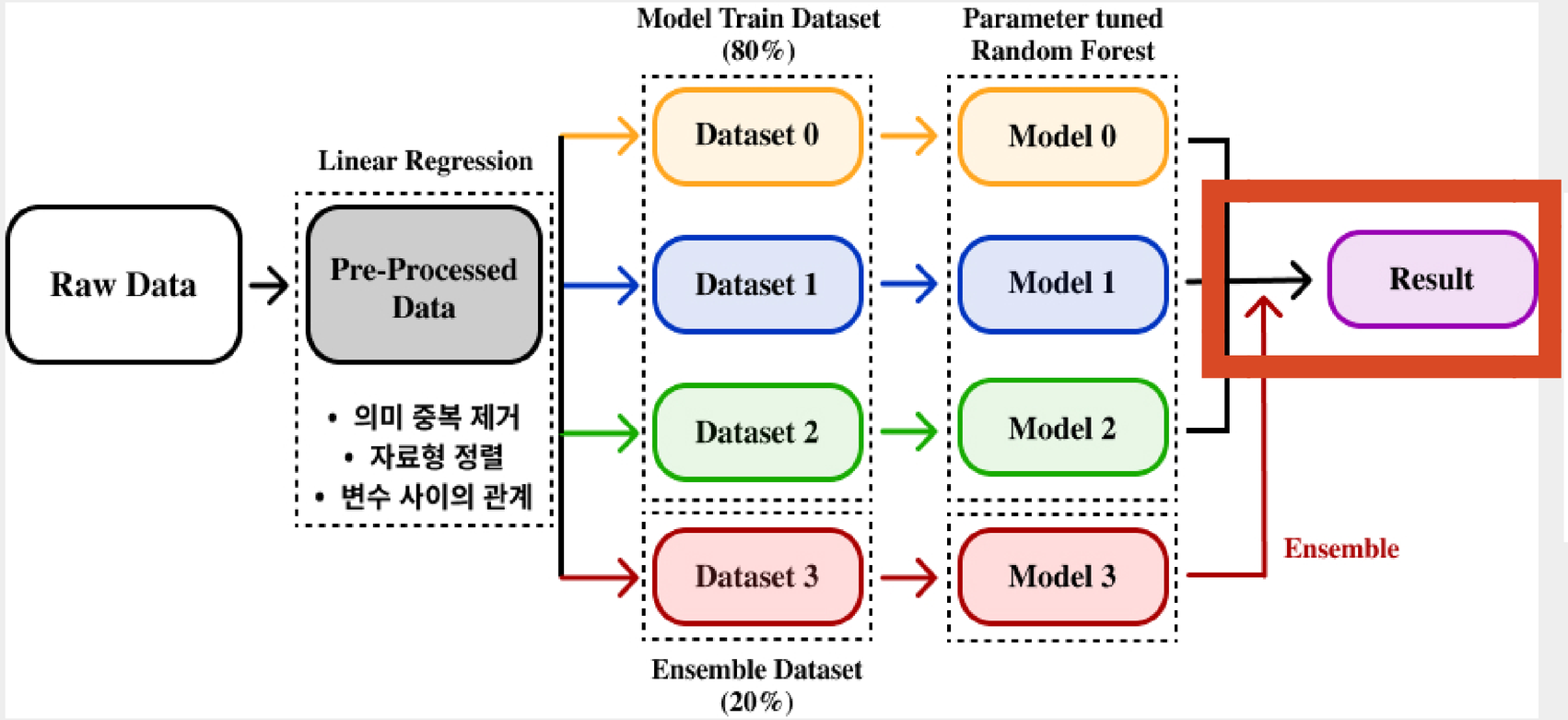
Ensemble Method	Trained Model Algoritm	Training Dataset	성능 (Accuracy)
Voting (soft)	Random Forest, LightGBM, XGBoost	Dataset 0 (trained)	95.14
Voting (soft)	Random Forest, LightGBM, XGBoost	Dataset 3 (non-trained)	94.95
Stacking Ensemble	Catboost + (Random Forest, XGBoost)	Dataset 0 (trained)	96.74
Our method	**Random Forest**	**Dataset 3 (non-trained)**	**98.79**

- Random Forest 모델에 Model 0, Model 1, Model 2의 결과를
Staking Ensemble을 변형한 Ensemble method를 통해 최종 inference를 추출

- Staking Ensemble은 이미 학습시킨 Dataset을 통해 Ensemble을 하지만,
Our method는 학습시키지 않은 Dataset으로 Ensemble하여 Overfitting을 최소화 함.

3. 모델링

01. 대출 신청 고객 예측 모델링: Model Evalution



Model	Model Algoritm	Dataset	성능 (F1-score)
Model 0	Random Forest	Dataset 0	52.25
Model 1	Random Forest	Dataset 1	53.45
Model 2	Random Forest	Dataset 2	25.64
Ensemble (Model 3)	**Random Forest**	**Dataset 3**	**81.04**

- Model 0, Model 1, Model 2과 Our method로 Ensemble한 Model과 F1-score으로 성능 비교

- 단일 모델인 Model 0, Model 1, Model 2와 Ensemble 모델과 확연한 성능 차이(최대 약56%p)가 남.

02. 서비스 메시지 군집 모델링

- 선정한 알고리즘: K-Means Clustering

방대한 데이터를 좋은 성능으로 빠르게(time complexity: O(log k)) 군집화 할 수 있음.

- Method:

1. 여러 ML 모델을 적용하여 Feature Importance로 중요한 feature을 선정
2. 상위 중요도와 고객의 특성을 잘 표현하는 6가지 feature 선정
3. 선정된 feature를 기반으로 K-Means Clustering을 사용하여 최적으로 군집화
4. 시각화를 통해 군집된 고객의 특징의 분포를 파악
5. 맞춤형 서비스 메시지 설계

모든 feature를 고려하여 군집화 하는 것보다 중요한 feature를 정량적으로 선별하여
군집화 하는 것이 더 효과적인 맞춤형 서비스 메시지를 제공할 수 있음.

3. 모델링

02. 서비스 메시지 군집 모델링: Feature Importance

LightGBM(split)	LightGBM(gain)	Random Forest	XGBoost(weight)	XGBoost(gain)	XGBoost(cover)	XGBoost(total_gain)	XGBoost(total_cover)
loan_rate	credit_score	loan_rate	loan_rate	credit_score	income_type_0	credit_score	credit_score
credit_score	loan_rate	credit_score	credit_score	income_type_0	credit_score	loan_rate	loan_rate
bank_id	desired_amount	application_id	desired_amount	desired_amount	insert_hour	desired_amount	income_type_0
product_id	income_type_0	user_id	loan_limit	loan_rate	purpose_6	income_type_0	desired_amount
loan_limit	product_id	loan_limit	product_id	purpose_6	loanapply_insert_hour	loan_limit	company_enter_month
desired_amount	bank_id	birth_year	bank_id	company_enter_month	purpose_4	product_id	product_id
birth_year	loan_limit	company_enter_month	income_type_0	existing_loan_cnt	income_type_2	company_enter_month	bank_id
yearly_income	company_enter_month	yearly_income	company_enter_month	product_id	company_enter_month	existing_loan_cnt	existing_loan_cnt
existing_loan_cnt	existing_loan_cnt	product_id	existing_loan_cnt	insert_hour	desired_amount	bank_id	loan_limit
existing_loan_amt	birth_year	existing_loan_amt	birth_year	income_type_5	loan_rate	birth_year	yearly_income
company_enter_month	yearly_income	desired_amount	employment_type_1	loan_limit	existing_loan_cnt	employment_type_1	birth_year
income_type_0	existing_loan_amt	loanapply_insert_minute	yearly_income	bank_id	purpose_2	purpose_6	purpose_6
loanapply_insert_hour	employment_type_1	insert_minute	existing_loan_amt	employment_type_1	income_type_5	income_type_5	income_type_5
application_id	purpose_6	bank_id	income_type_5	income_type_4	yearly_income	yearly_income	loanapply_insert_hour
loanapply_insert_day	income_type_5	insert_day	loanapply_insert_hour	loanapply_insert_hour	bank_id	insert_hour	employment_type_1
user_id	loanapply_insert_hour	loanapply_insert_day	purpose_2	income_type_2	product_id	loanapply_insert_hour	insert_hour
income_type_5	loanapply_insert_day	insert_hour	purpose_6	yearly_income	loan_limit	existing_loan_amt	income_type_2
loanapply_insert_minute	insert_hour	existing_loan_cnt	income_type_2	purpose_4	employment_type_1	purpose_2	purpose_2
loanapply_insert_month	purpose_2	loanapply_insert_hour	insert_hour	purpose_2	birth_year	income_type_2	purpose_4
purpose_6	loanapply_insert_month	insert_month	purpose_4	income_type_1	income_type_4	purpose_4	existing_loan_amt

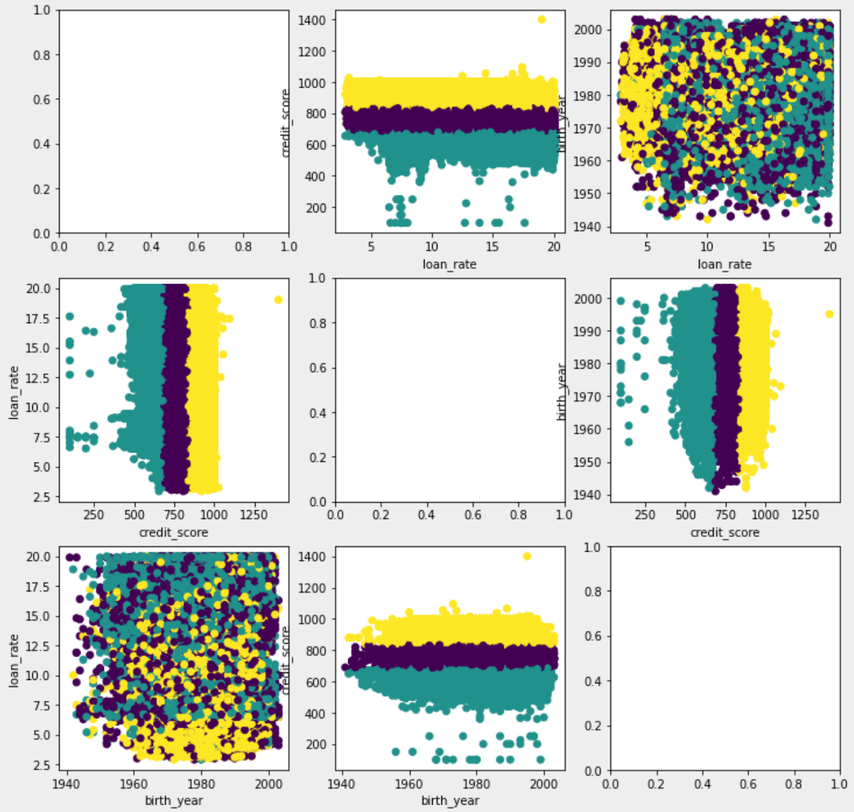
- LightGBM, XGBoost, RandomForest 등의 트리 기반 알고리즘을 통해 feature importance 추출
- 상위 20개의 feature를 수집한 후, 고객의 특징을 담고 있는 feature 6가지 선정

=> birth_year, credit_score, desired_amount, loan_limit, loan_rate, yearly_income

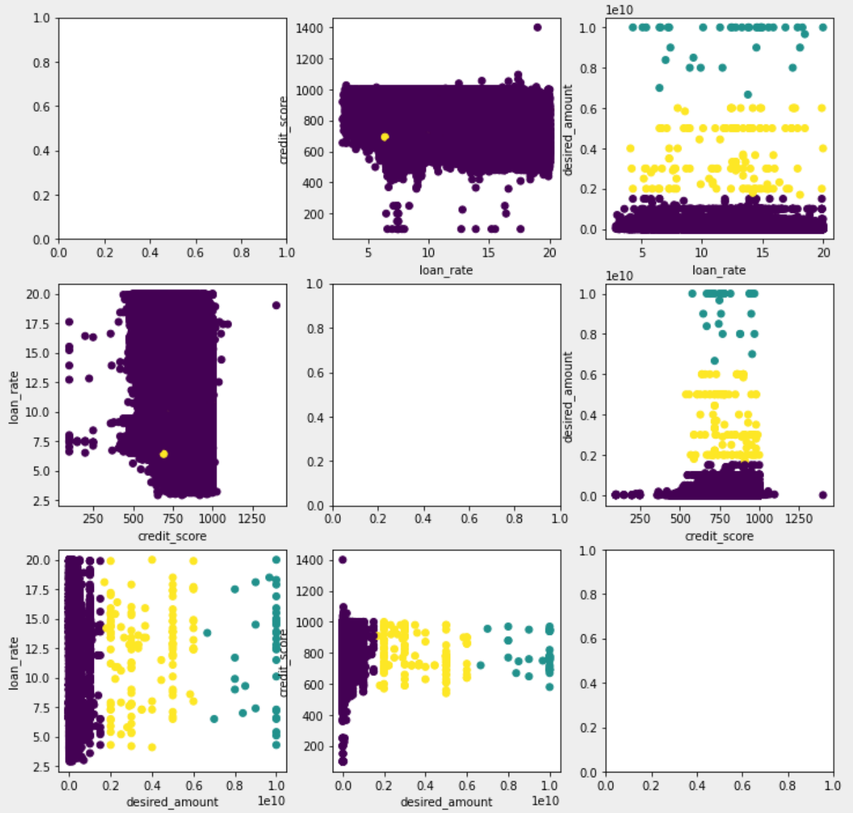
3. 모델링

02. 서비스 메시지 군집 모델링: K-Means

birth_year, credit_score, loan_rate



loan_rate, credit_score, desired_amount



- 선정한 6가지 feature를 3가지의 feature씩 군집화 시각화
- 육안으로 군집이 잘 구별되는 결과 취합

4. 결론

01. 대출 신청 고객 예측

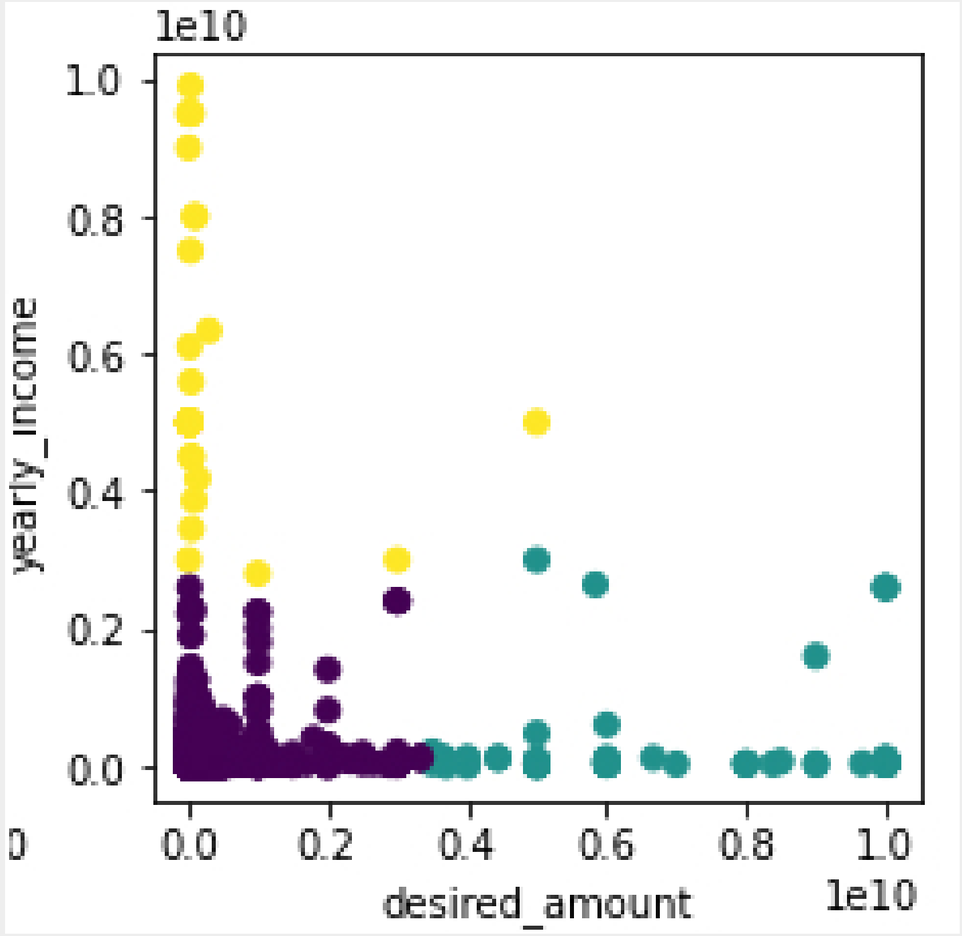
[Problem]
제한된 하드웨어 리소스에서 방대한 양의 Data를 모델에 학습시켜야 하는 문제가 존재했음.

[Our Solution]
연산량이 많고 시간이 오래걸리는 단점을 해결하기 위해,
Dataset을 Split하여 각각 모델에 학습시킨 뒤,
Staking Ensemble을 응용하여 고안한 Ensemble Method로 Overfitting 문제 최소화함.

[Conclusion]
F1-Score가 약 81%인 모델을 생성해 냄.

4. 결론

02. 군집 모델링 서비스 메시지: 군집화 분석



[Result]

1. 연봉이 **적고**(상위 25% 이하) 대출 희망 금액이 **적은**(상위 25% 이하) 사람 (보라)

2. 연봉이 **적고**(상위 25% 이하) 대출 희망 금액이 **높은**(상위 25% 이상) 사람 (초록)

3. 연봉이 **높은**(상위 25% 이상) 사람 (노랑)

4. 결론

02. 군집 모델링 서비스 메시지: 서비스 메시지 제안

1. 연봉이 적고(상위 25% 이하) 대출 희망 금액이 적은(상위 25% 이하) 사람 (보라)

⇒ 대출 한도와 상관없이 **가장 빠르게 대출**되는 상품 비교 메시지

2. 연봉이 **적고(상위 25% 이하)** 대출 희망 금액이 **높은(상위 25% 이상)** 사람 (초록)

⇒ **이자가 낮고 대출 한도가 높은** 대출 상품 비교 메시지

3. 연봉이 **높은(상위 25% 이상)** 사람 (노랑)

⇒ **금리가 낮은** 대출 상품 비교 메시지

THANK YOU

끝까지 봐주셔서 감사합니다.

Honglk

김태용, 권재현, 박세빈, 한혜원, 홍표민