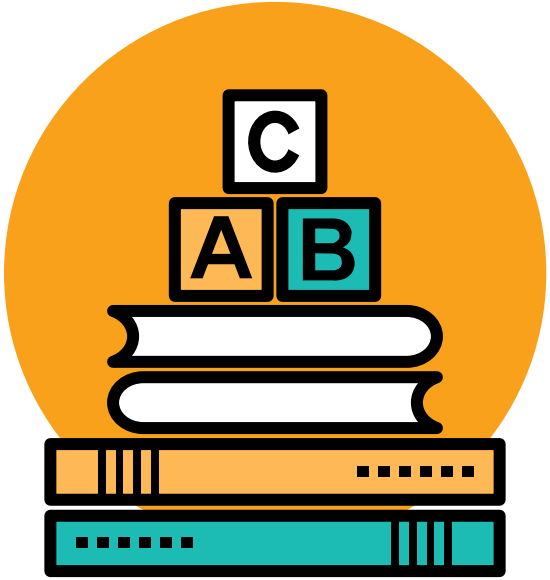


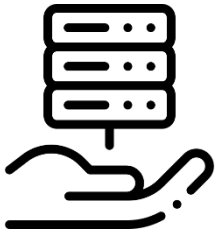
정액제 구매 여부 예측

강주연 · 김학빈 · 박수민 · 유용준

Kdata 한국데이터산업진흥원 ONE store books KGU 경기대학교

목차

1. 주제 선정

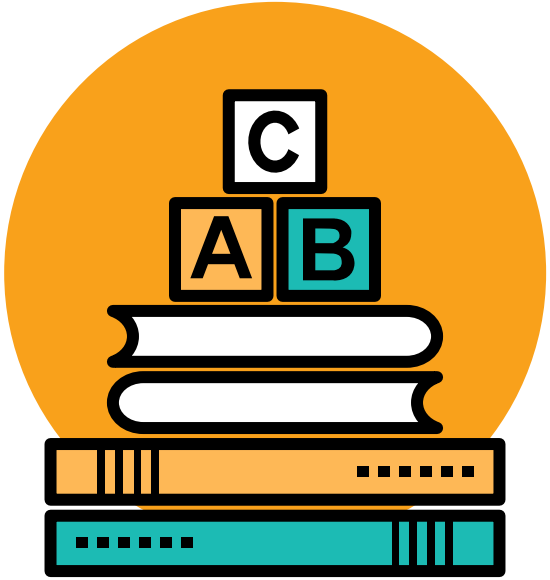


2. EDA

3. 데이터 전처리

4. 모델링
및 결과

5. 기대 효과



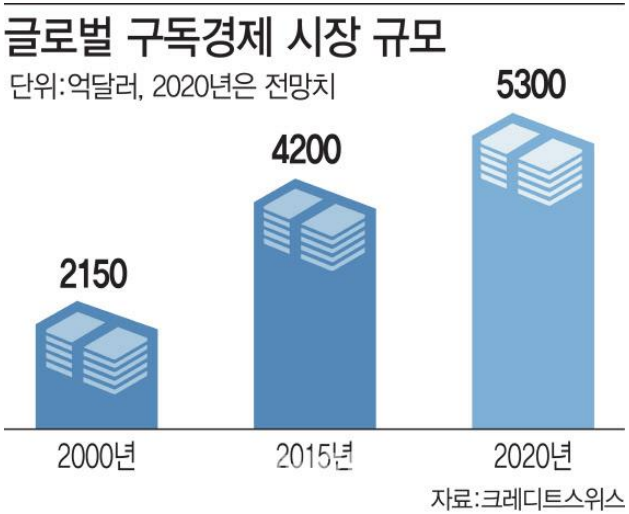
1. 주제 선정

주제 선정배경

구독경제?

소비자가 가입 후 월/분기/연 단위의
정기결제를 선택하면 약정한 기간동안
유무형의 제품/서비스를 제공받는 것을 의미

주제 선정배경

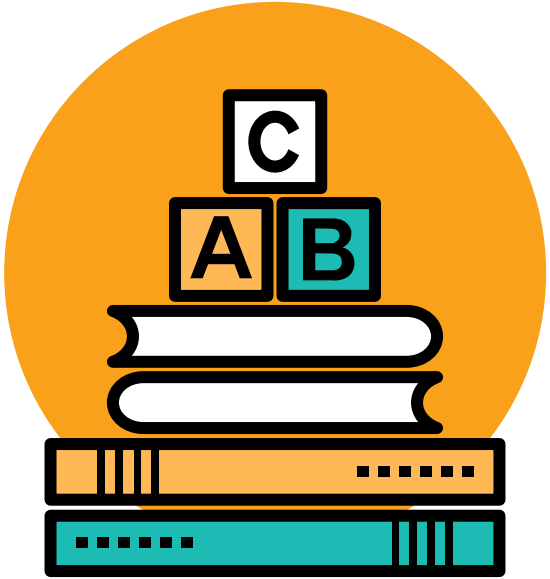


- ✓ 해마다 늘어가는 시장규모
- ✓ 향후 전자책 시장에서 구독경제가 활발해질 것으로 전망
- ✓ 다양한 형태의 독서 경험을 기대하는 소비자

원스토어 북스 정액제 종류

정액제

<소장>	<대여>	<기타>
eBook 전권 소장	eBook 전권 대여	eBook 정액제
웹소설 전권 소장	웹소설 전권 대여	오디오북 정액제
웹툰 전권 소장	웹툰 전권 대여	코믹 정액제
코믹 전권 소장	코믹 전권 대여	TV방송시리즈



2. EDA

제공받은 데이터

	partition_dt	prod_id	prchs_id	cust_payment_amt	sett_target_cpn_amt	prod_amt	dtl_category_nm
15	20200201	H038835484	20020113412620311081	0	800	2700	로맨스
16	20200201	H038835484	20020113412620311081	1630	0	2700	로맨스
17	20200201	H038835484	20020113412620311081	0	270	2700	로맨스

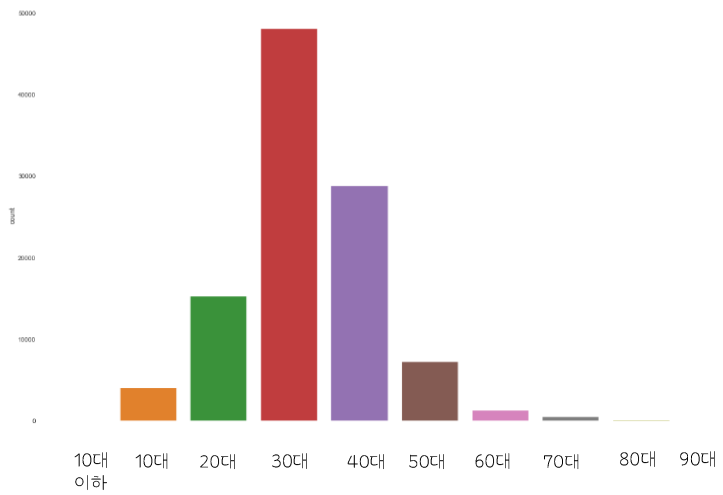
변수명	의미	변수명	의미	변수명	의미
prchs_id	구매 ID	age_cd	나이	prchs_tm_clsf_nm	구매 시간
prod_id	상품 ID	category_nm	카테고리	mno_cd	통신사
partition_dt	구매 날짜	dtl_category_nm	상세 카테고리	cust_payment_amt	총 결제
insd_usermbr_no	고객 ID	sex_clsf_cd	성별	sett_target_cpn_amt	쿠폰 사용금액
prod_amt	상품금액				

2020.02월 ~07월 (6개월) 까지의 고객 거래 데이터

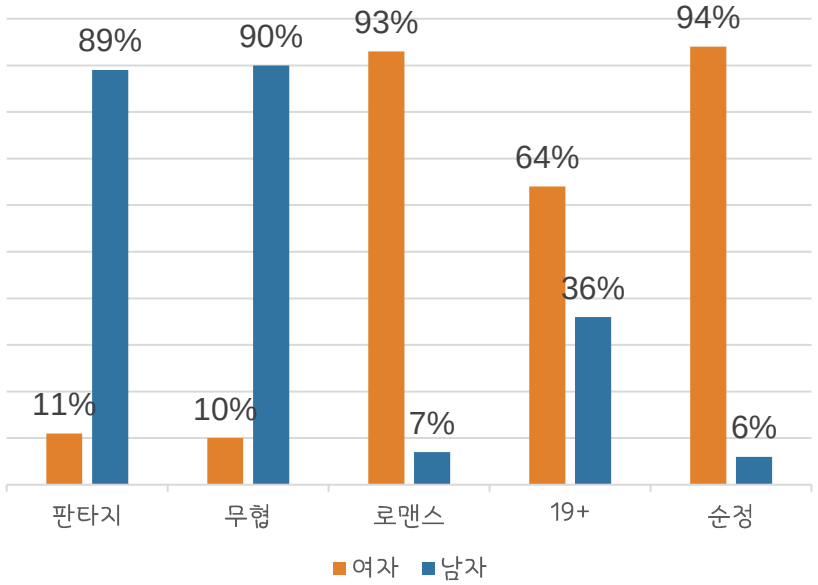
EDA

전체 데이터 분석

< 전체 구매자 연령대 분포 >

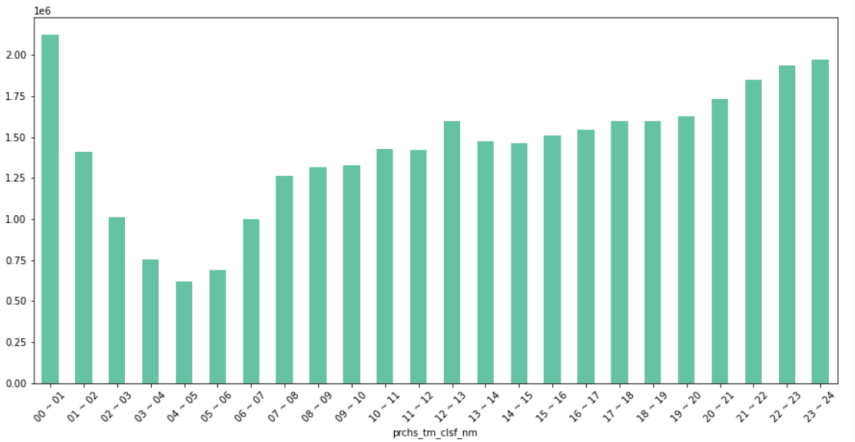


< 상위 5개 세부 카테고리 별 성별 분포>

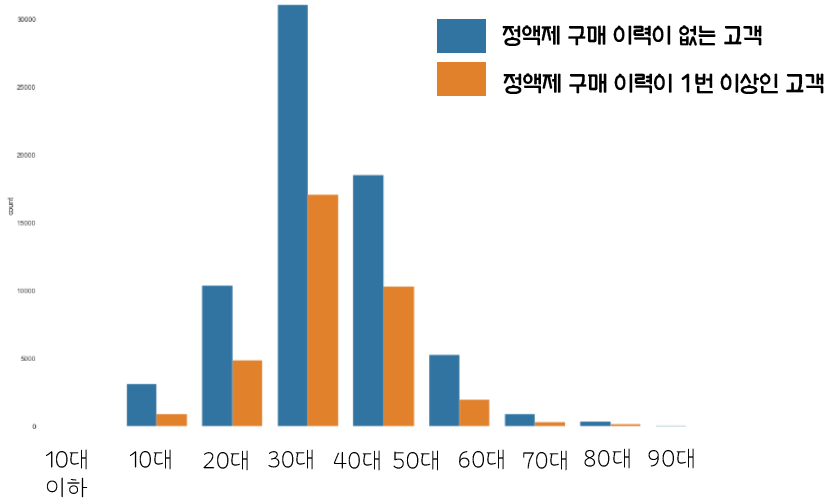


전체 데이터 분석

< 구매에 따른 시간대 분포 >

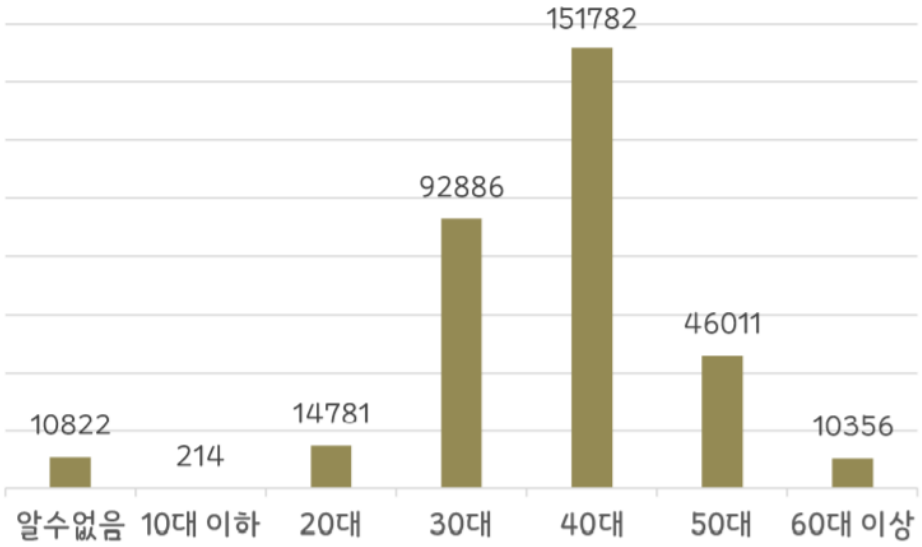


< 나이에 따른 정액제 결제 분포>

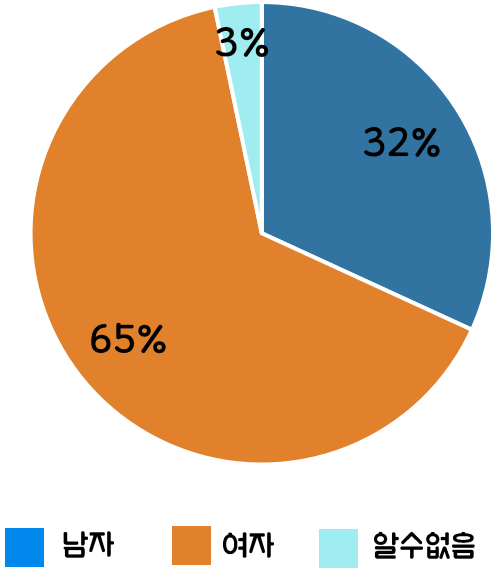


전체 데이터 분석

< 정액제 구매자 연령 분포 >

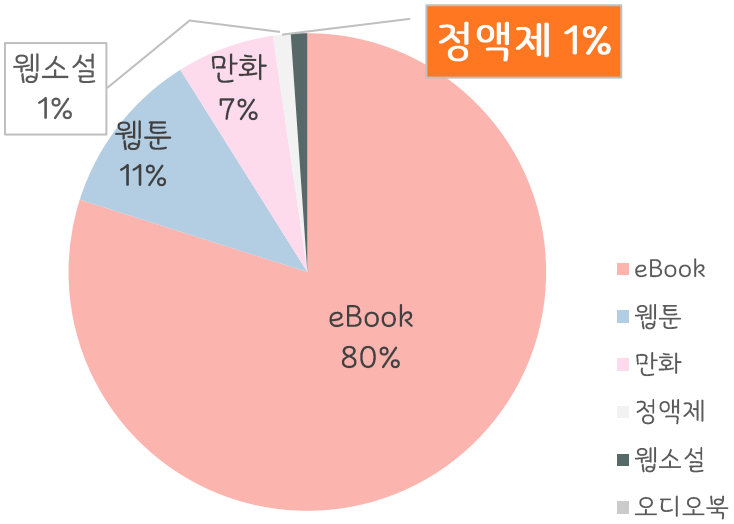


< 정액제 구매자 성별 분포 >

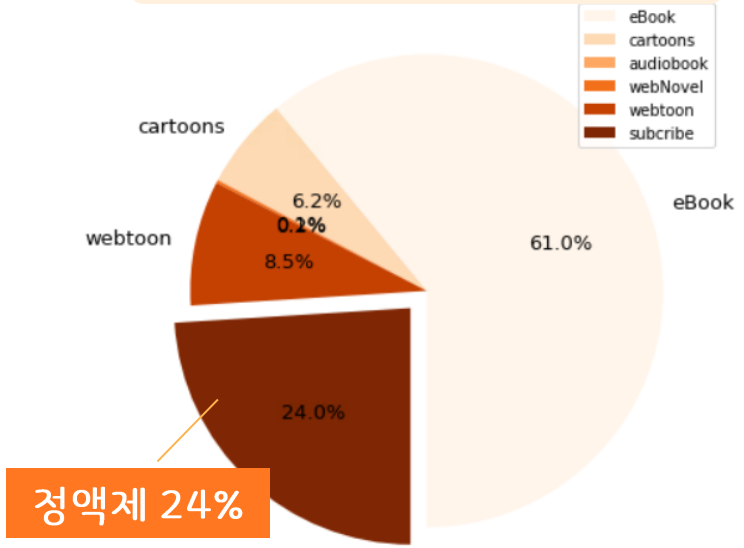


전체 데이터 분석

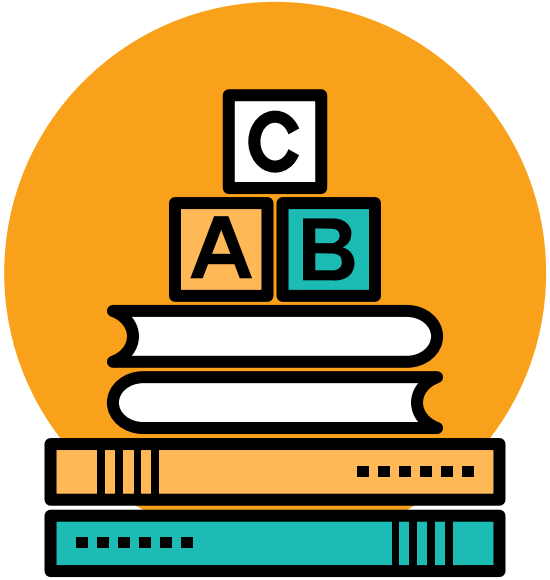
< 카테고리 별 거래 건 수>



< 카테고리 별 전체 매출>



전체 거래건 수에서 정액제가 차지하는 비율은 **1%**
But, 전체 매출에서 정액제는 **24%**를 차지



3. 데이터 전처리

Raw Data Preprocessing

	partition_dt	prod_id	prchs_id	cust_payment_amt	sett_target_cpn_amt	prod_amt	dtl_category_nm
15	20200201	H038835484	20020113412620311081	0	800	2700	로맨스
16	20200201	H038835484	20020113412620311081	1630	0	2700	로맨스
17	20200201	H038835484	20020113412620311081	0	270	2700	로맨스

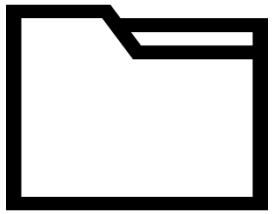
* 약 4천만건의 구매

	partition_dt	prod_id	prchs_id	cust_payment_amt	sett_target_cpn_amt	prod_amt	dtl_category_nm
99309	20200201	H038835484	20020113412620311081	1630	1070	2700	로맨스

* 약 3천만건의 구매

✓ 고객이 상품을 구매할 때 여러가지 지불방식을 사용하면 각 지불방식마다 레코드가 하나씩 생겨서 이를 합치는 작업을 수행

고객 테이블 생성



구매 데이터

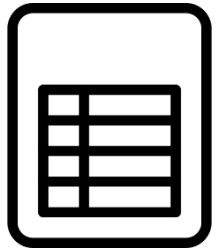
- ✓ 날짜를 Datetime으로 변환하고 [month / week]
- ✓ 주말 여부 [is _weekend]
- ✓ Frequency, Recency, T
- ✓ 구매 날짜 간격 [dt_step_min/max]
- ✓ 주말 구매횟수 [weekend_prchs]
- ✓ 월, 주간 구매 금액 평균 [wly / mly_prchs_amt_mean]

.

등등

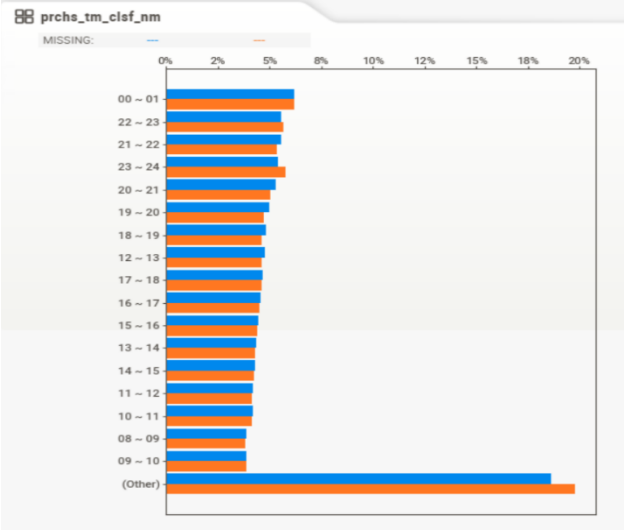
총 40개의 column으로 구성

*** 총 15만명의 고객테이블 생성**

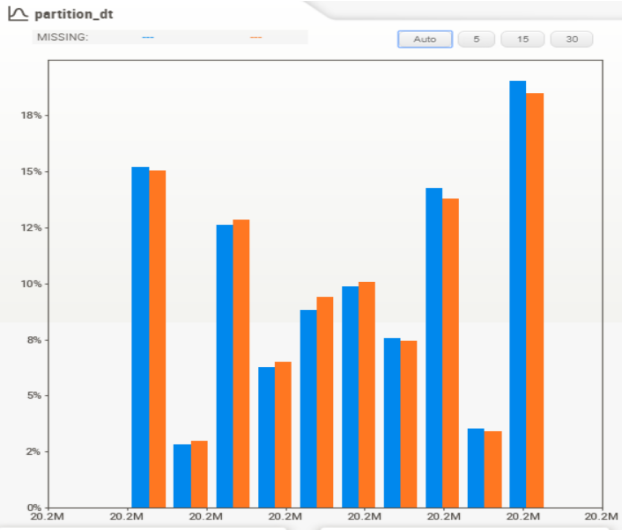


고객 테이블

개인정보제공 비동의 구매자 처리



-> 구매 시간



-> 구매 날짜

개인정보제공 비 동의 고객

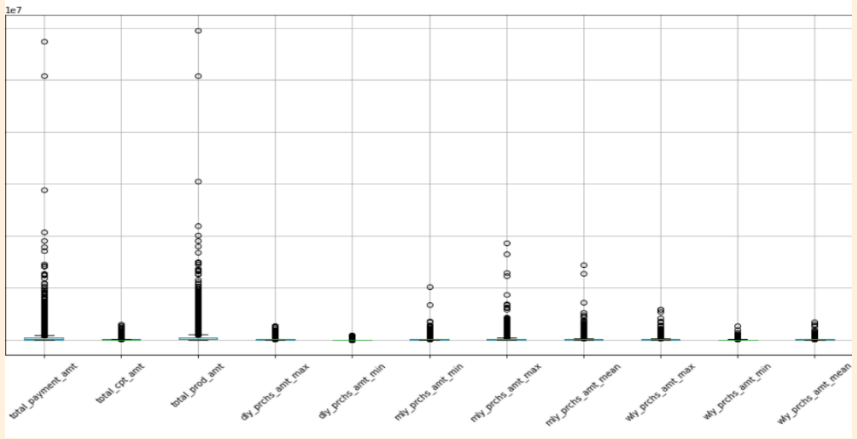
개인정보제공 동의 고객

✓ 비교한 결과 유의미한 차이가 없고, 데이터의 양이 충분하다고 판단

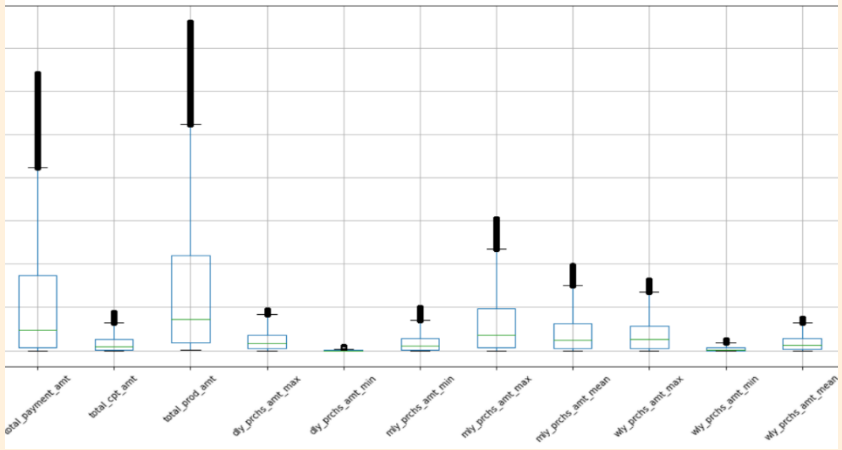
정보제공 비동의 구매자 **제거 결정**

이상치 처리

1.5 x IQR 이상치 제거



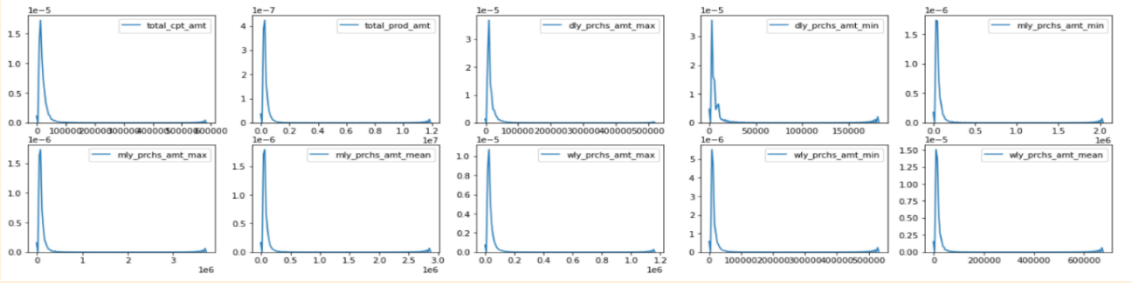
큰 값들이 많아서 박스가 보이지 않음.



박스가 보이긴 하지만, 제거하기엔 손실데이터량이 커, **제거하지 않기로 함**.

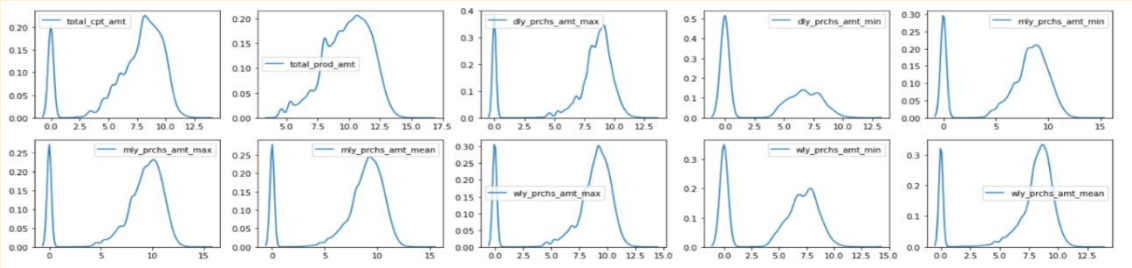
로그 변환

변환 전



수치형 feature 전부 longtail 형태를 띄고 있어, ***log 변환*** 을 하기로 결정

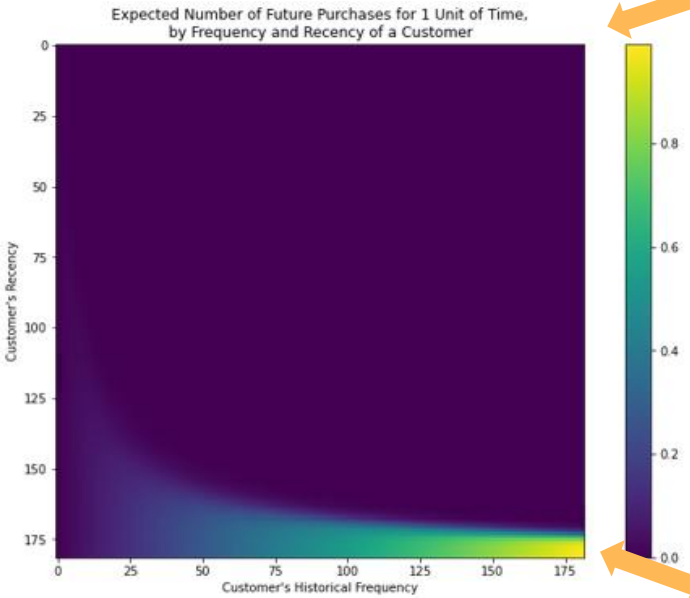
변환 후



로그 변환 후 그래프의 왜도를 낮춤.

왜도	변환 전	8.70	15.11	8.44	13.28	15.29	14.45	16.19	9.62	22.57	12.49
	변환 후	-1.28	-0.45	-1.57	0.14	-0.99	-1.43	-1.42	-1.52	-0.47	-1.50

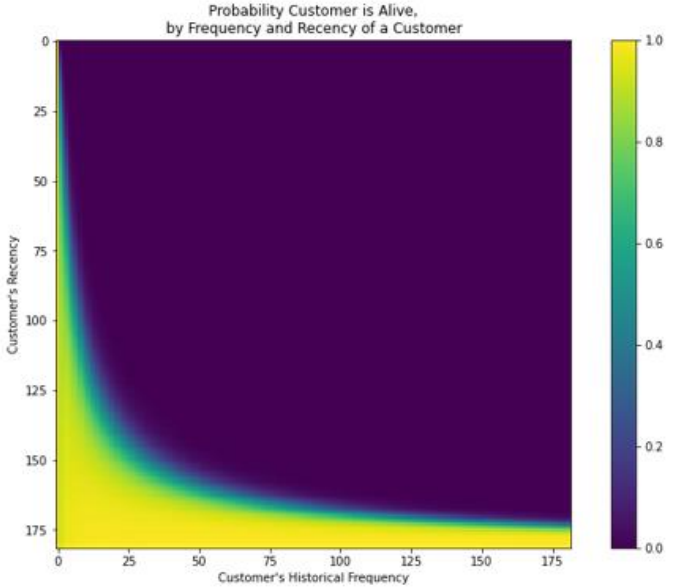
Consumer Lifetime visualization



차가운 고객

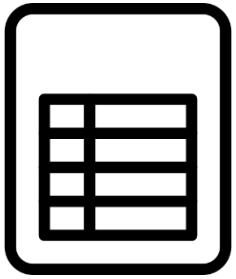
Best 고객

< Frequency / Recency matrix >



< Probability of customer alive >

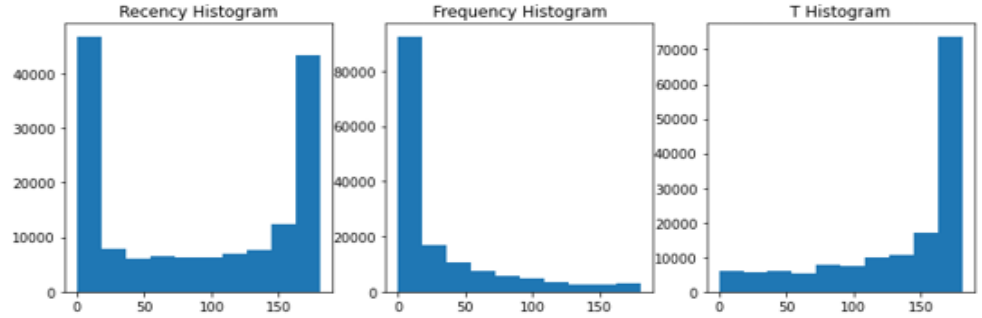
R,F,T 변수를 통한 군집 생성



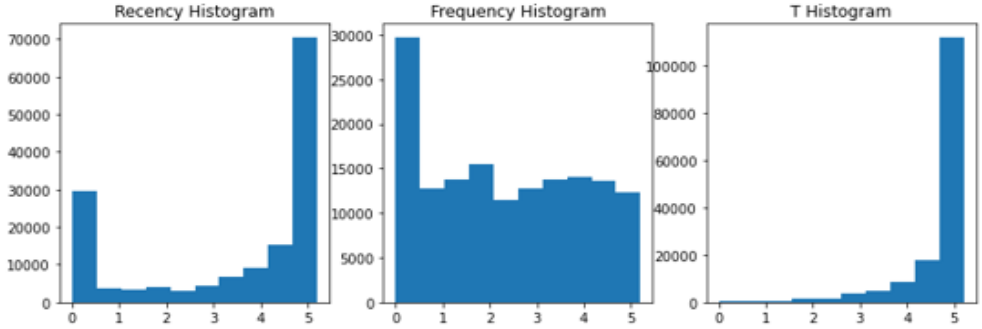
전체 거래데이터

'frequency'
'recency'
'T'

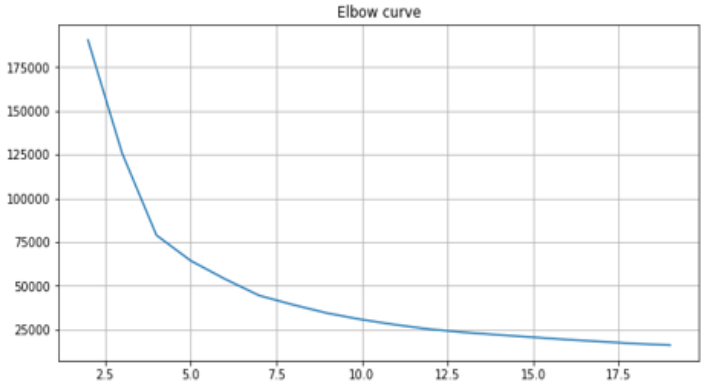
변수추출

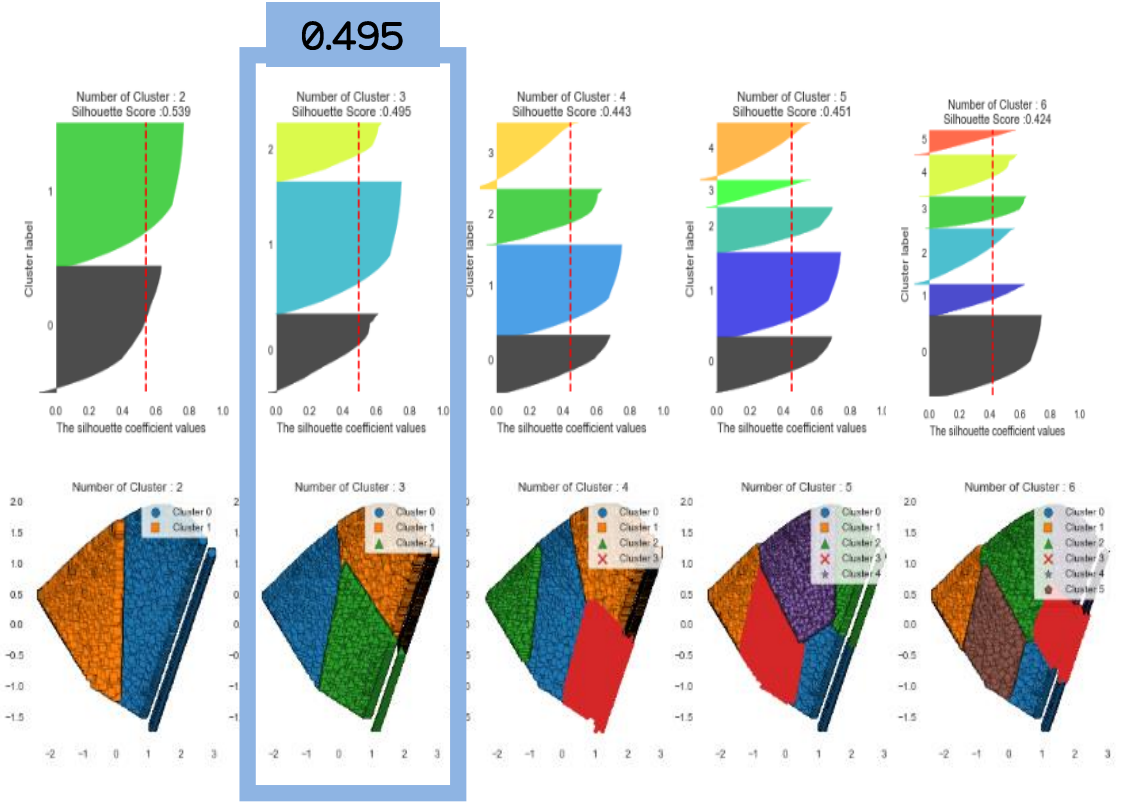


Frequency만 로그변환



K-means Clustering





✓ Number of cluster = 3 일때 Silhouette Score가 최대이고, 개별 군집의 평균값의 편차가 크지 않으므로 3개의 군집으로 나누어 **Feature로 추가**

Encoding

Label

시간대(pref_tm)
성별(sex_clsf_cd)

시간대	Label
02~06	1
06~10	2
10~14	3
14~18	4
18~22	5
22~02	6

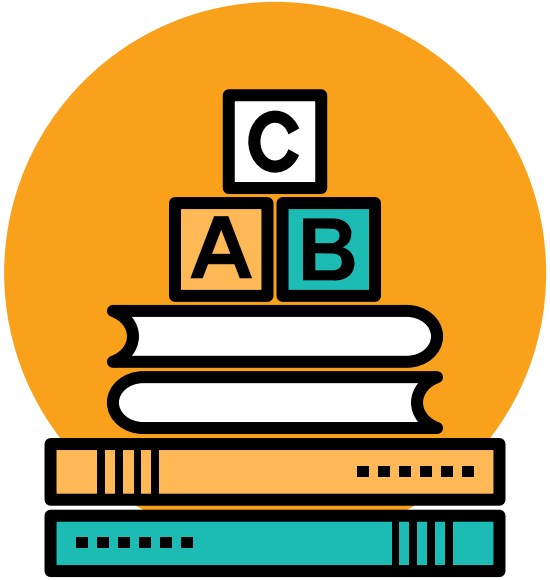
One-Hot

카테고리(pref_category)
통신사(mno_cd)
고객 클러스터 (cluster_label_)

카테고리
Ebook
웹툰
•
•

Ebook	웹툰
1	0
0	1
•	•
•	•

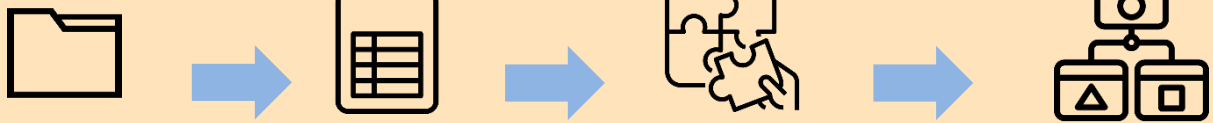
• •



4. 모델링

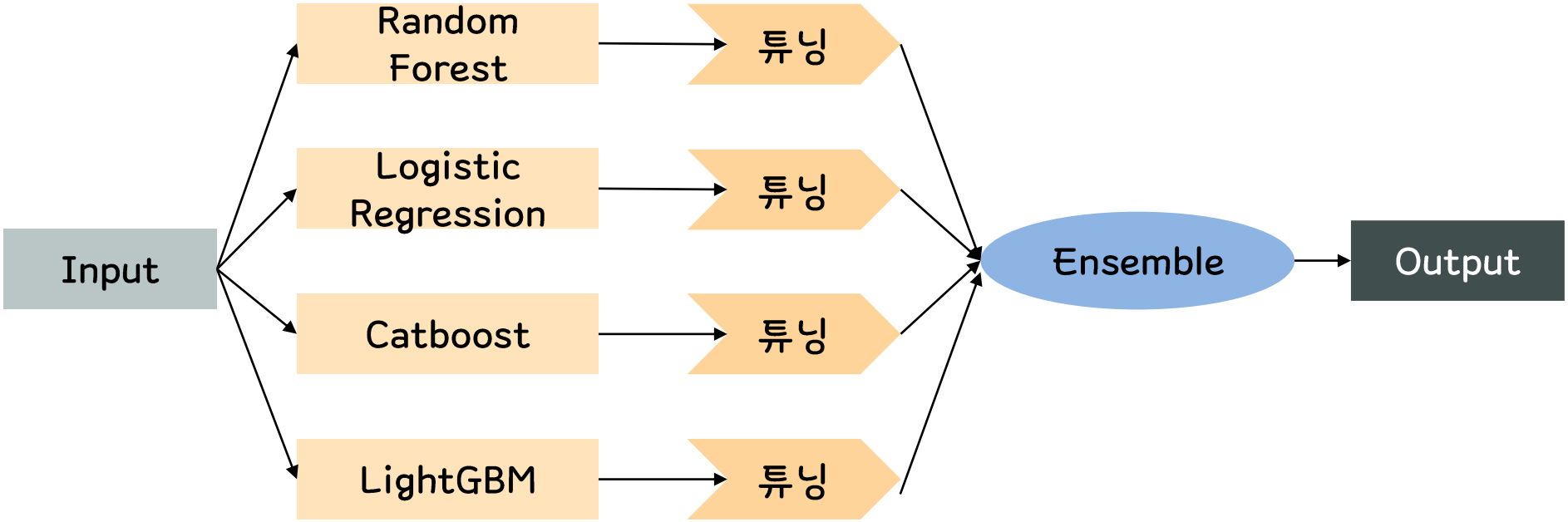
활용 데이터 정의

Train set



Test set

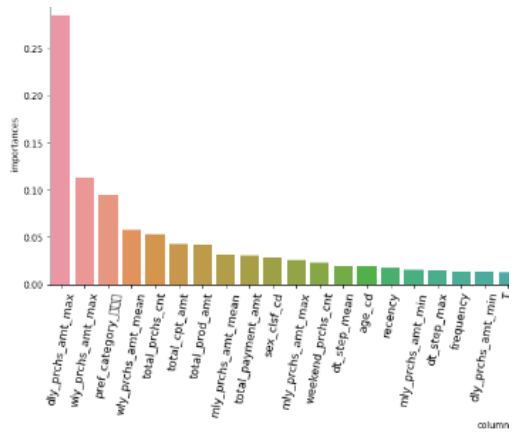
모델링 개요

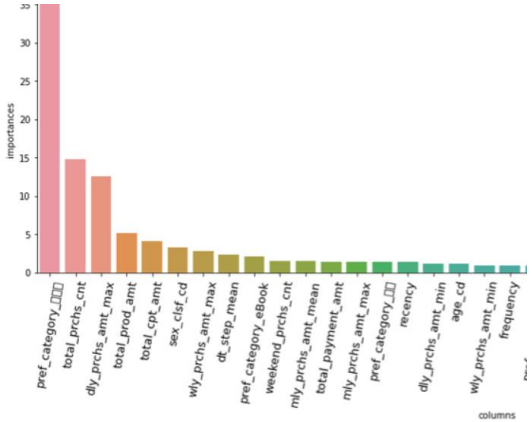

Input
Random Forest
Logistic Regression
Catboost
LightGBM
튜닝
튜닝
튜닝
튜닝
Ensemble
Output

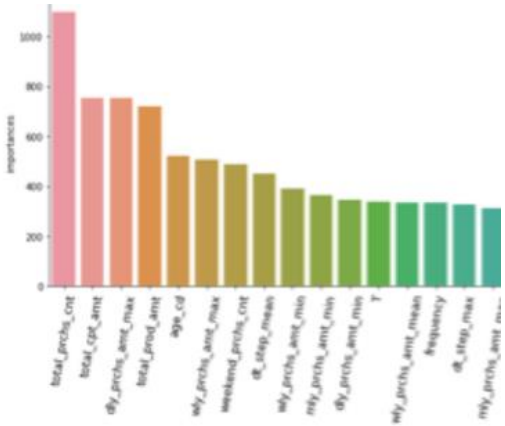
결과

	알고리즘	정확도	정밀도	재현율	AUC	F1 score
튜닝 전	Random Forest	0.8568	0.7765	0.7482	0.9256	0.7621
	Logistic Regression	0.8261	0.7140	0.7216	0.8928	0.7178
	Catboost	0.8659	0.7740	0.7942	0.9338	0.7840
	LightGBM	0.8613	0.7692	0.7821	0.9314	0.7683
튜닝 후	Random Forest	0.8635	0.7775	0.7772	0.9315	0.7773
	Logistic Regression	0.8261	0.7359	0.6751	0.8924	0.7042
	Catboost	0.8737	0.7878	0.7727	0.9380	0.7802
	LightGBM	0.8622	0.7706	0.7840	0.9326	0.7772

Feature Importance

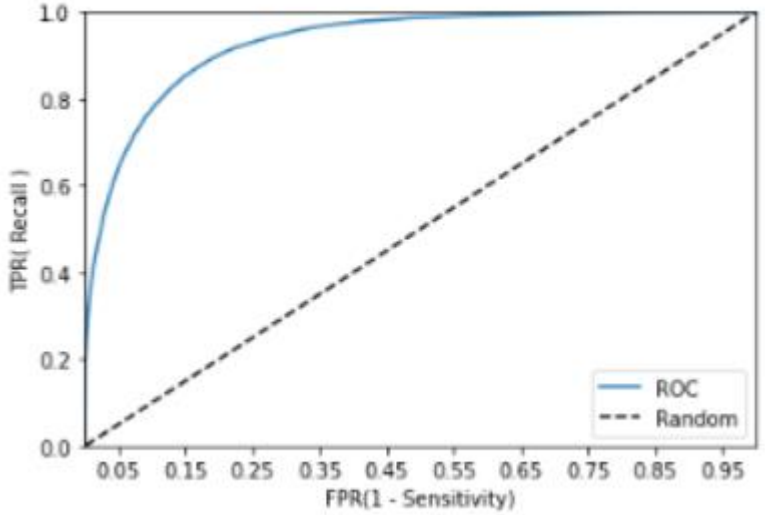






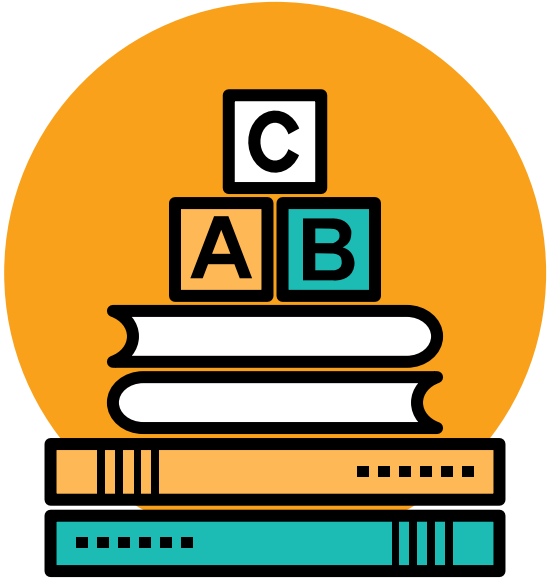
	Random Forest	Catboost	LightGBM
1	일일 구매금액 최대	선호 카테고리 <정액제>	총 구매 횟수
2	주간 구매금액 최대	총 구매 횟수	총 구매 금액

앙상블 결과



	Positive	Negative
True	56642	5914
False	6465	21185

	정확도	정밀도	재현율	AUC	F1 score
Soft voting	0.8628	0.7818	0.7662	0.9311	0.7739



5. 기대효과

기대효과 및 활용방안

회사

: 정액제 구매 여부를 예측해 CRM에 활용한다.

- 정액제 구매 확률 예측으로 이탈 가능성 높은 고객에게 쿠폰 발행
- 정액제를 구매할 가능성이 높은 고객들을 유도

: 정액제 고객들의 선호도 분석 등을 고려하여 북패스 내 책 추가

소비자

: 다양한 형태의 더 나은 독서 서비스로 만족도 향상

한계점 및 개선점

- ✓ 보안문제로 제공된 가상환경에서의 작업에 제약이 많았음
- ✓ 고객의 동향을 반영할 feature 부족
- ✓ Feature Engineering 에서 다양한 시도를 할 시간 부족
- ✓ 노이즈 및 결측치 데이터에 대한 처리방안을 다르게 적용
- ✓ 다양한 요인과 시계열 데이터를 활용할 수 있는 LSTM 기법을 적용

감사합니다

Q&A

감사합니다

Appendix

변수명	의미	변수명	의미	변수명	의미
insd_usermbr_no	유저 id	pref_dtl_category	선호세부카테고리	mly_prchs_amt_mean	월간구매금액 평균
sex_clsf_cd	성별	pref_tm	선호구매시간	wly_prchs_amt_max	주간구매금액 최대
age_cd	나이	mno_cd	고객 통신사	wly_prchs_amt_min	주간구매금액 최소
total_prchs_cnt	총구매횟수	dly_prchs_amt_max	일일구매금액 최대	wly_prchs_amt_mean	주간구매금액평균
weekend_prchs_cnt	주말 구매 횟수	dly_prchs_amt_min	일일구매금액 최소	total_buys	정액권 구매 횟수
total_payment_amt	총구매금액	mly_prchs_amt_max	월간구매금액 최대	dt_step_min	구매 간격 최소일수
total_prod_amt	총구매원가	mly_prchs_amt_min	월간구매금액 최소	dt_step_max	구매 간격 최대일수
recency	활동기간	mly_prchs_amt_mean	월간구매금액 평균	dt_step_mean	구매간격 평균
cluster_label_0 ~2	고객 군집 클러스터링	bin_count	예측 target 값	Frequency	반복구매횟수
mno_cd_US001201 ~US001210	통신사 One-Hot 인코딩	pref_category _eBook ~ 정액제	선호 카테고리 One-Hot 인코딩	T	최근 구매 안한 기간

참고문헌

- **Lifetime 라이브러리**https://lifetimes.readthedocs.io/en/master/Quickstart.html#the-gamma-gamma-model-and-the-independence-assumption
- **IQR** https://en.wikipedia.org/wiki/Interquartile_range
- https://brunch.co.kr/@gimmesilver/53
- https://www.kaggle.com/pavanraj159/telecom-customer-churn-prediction
- https://www.kaggle.com/c/kkbox-churn-prediction-challenge/notebooks
- https://www.kaggle.com/barelydedicated/bank-customer-churn-modeling