

A hand holding a credit card over a workbench with tools. The background is a dark, out-of-focus workbench with various tools like a wrench, screwdriver, and pliers. A hand is holding a credit card, with the word 'Capital' and 'new' visible. A semi-transparent grey rounded rectangle is overlaid on the image, containing the title and authors.

Home Credit Default Risk

이중한 김주희 박춘수 송주혁 이수영

CONTENTS

Home Credit Default Risk

01

프로젝트 배경 및 목적

02

분석 데이터 정의

03

데이터 처리 및 분석 과정

04

서비스 활용 방안 및 기대 효과

The background of the slide is a dark, grayscale photograph. It features a small model of a house with a gabled roof and several windows, resting on a dark wooden surface. In the foreground, a set of keys with a circular ring is visible, slightly out of focus. On the left side, there is a bright yellow rectangular block.

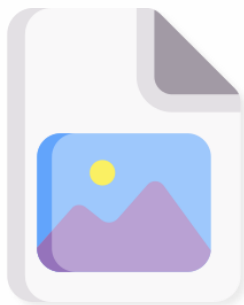
01

Home Credit Default Risk
프로젝트 배경 및 목적

01. 프로젝트 배경

정형 데이터

비정형 데이터 X



정형 데이터

kaggle™



DACON

DATA TO VALUE

01. 프로젝트 배경

주제 선정 이유

Home Credit

- 체코에 설립된 국제적인 비은행 금융기관
- 신용기록이 거의 없는 사람을 대상으로 대출
- 고객의 상환 능력을 예측하기 위해 거래 정보 등 다양한 대체 데이터 사용
- 고객의 더 나은 상환 능력 예측을 위해 Kaggle Contest 진행(~ 2018.08.29)

사회적 배경

- 코로나19로 당장 생계가 급한 소상공인·자영업자를 위한 긴급 대출 심사에 많은 시간 소요
- 간편 금융 서비스 확장으로 대출 상환 예측에 대한 수요 증가 예상
- 비대면 대출 서비스에 대한 니즈

01. 프로젝트 목적 및 목표











프로젝트 목적: 대출 상환 능력 여부 평가

프로젝트 목표:

- Kaggle Score **0.792**
- 1735등** / 7190등, 24% 이내

*"0.792를 baseline으로 정하고
그 이상의 점수가 나타나는 결과는
충분히 가치가 있다."*

- Kaggle Home Credit Competition

<div><div>In the money</div><div>Gold</div><div>Silver</div><div>Bronze</div></div>								
#	△pub	Team Name	Notebook	Team Members	Score 🏆	Entries	Last	
1	▲10	Home Aloan		 +3	0.80570	499	2y	
2	—	ikiri_DS		 +9	0.80561	477	2y	
3	▲1	alijs & Evgeny			0.80511	143	2y	
4	▲6	Quad Machine			0.80474	178	2y	
5	▼4	Kraków, Lublin i Zhabinka			0.80449	329	2y	
6	▲8	silver		 +4	0.80419	372	2y	
7	▲10	A.Assklou _ Aguiar			0.80396	200	2y	
8	▼2	七上八下		 +4	0.80376	476	2y	
9	▲42	International Fit Club		 +5	0.80374	513	2y	
10	▲19	Best Friend Forever: CV			0.80354	476	2y	

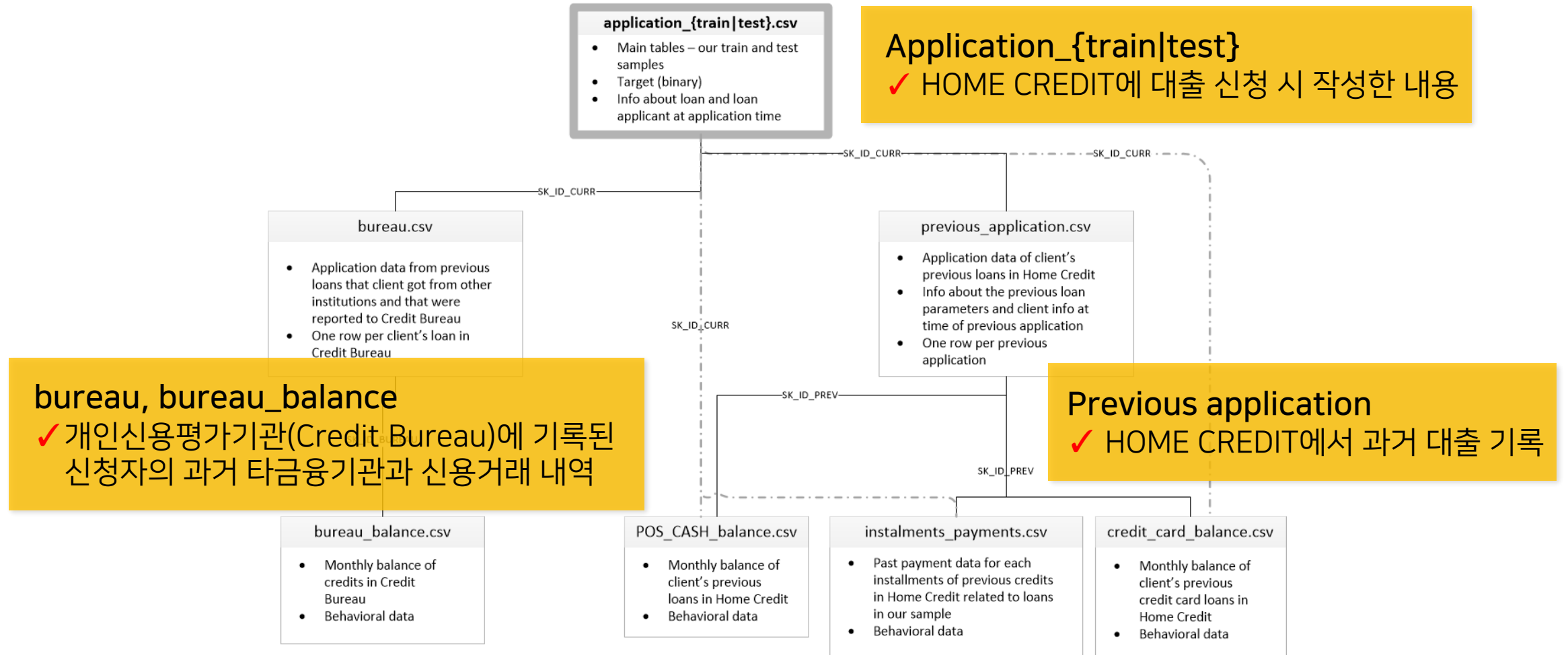
The background of the slide is a dark, grayscale photograph. It features a house with a gabled roof and several windows, positioned in the upper right. In the lower left, a set of keys with a large, dark keychain is visible. The overall tone is somber and professional.

02

Home Credit Default Risk 분석 데이터 정의

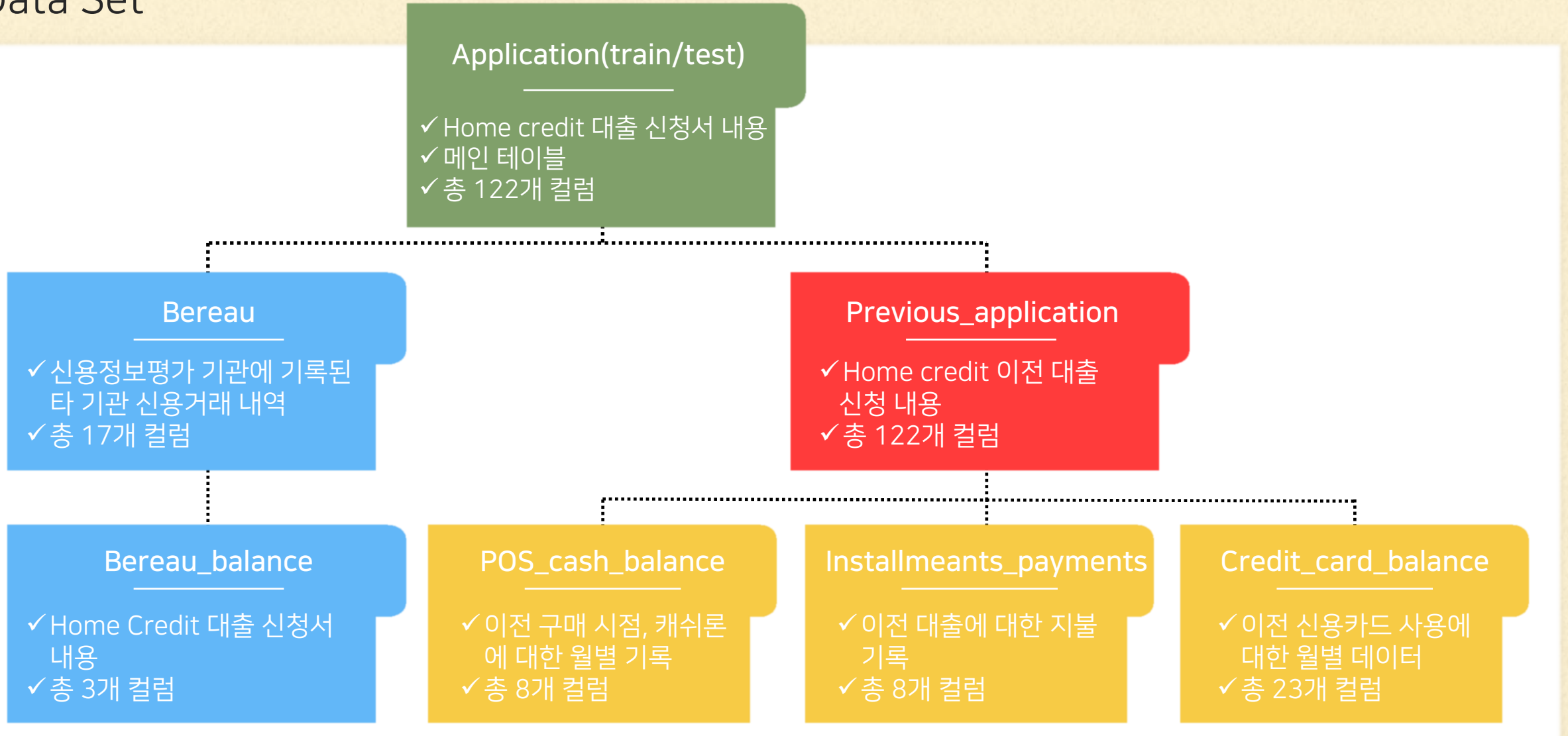
02. 분석 데이터 정의

Data Set



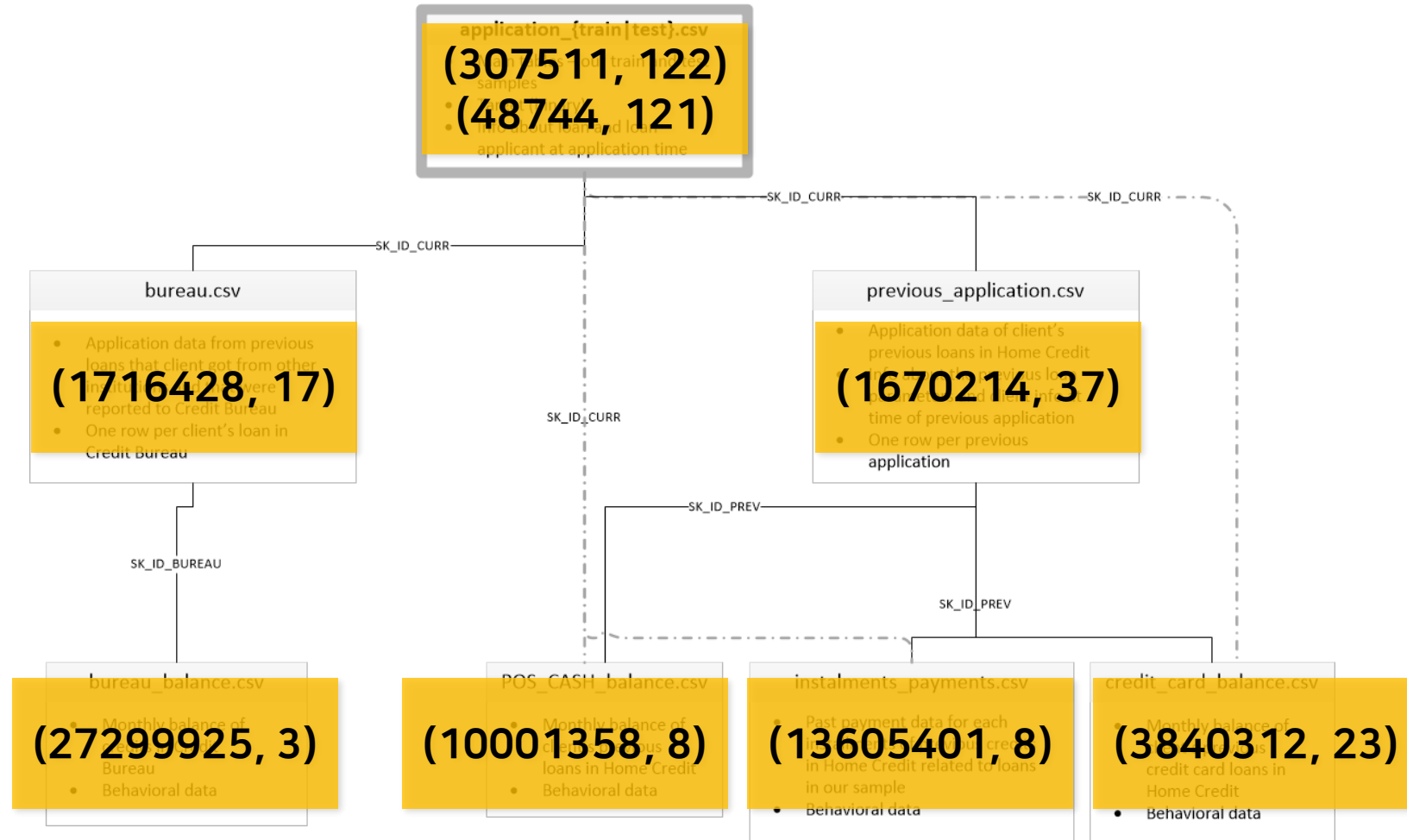
02. 분석 데이터 정의

Data Set



02. 분석 데이터 정의

Data Set



The background of the slide is a dark, grayscale photograph. It features a small model of a house with a gabled roof and several windows, resting on a dark wooden surface. In the foreground, a set of keys with a large, dark, circular ring is visible. The overall mood is professional and related to real estate or finance.

03

Home Credit Default Risk

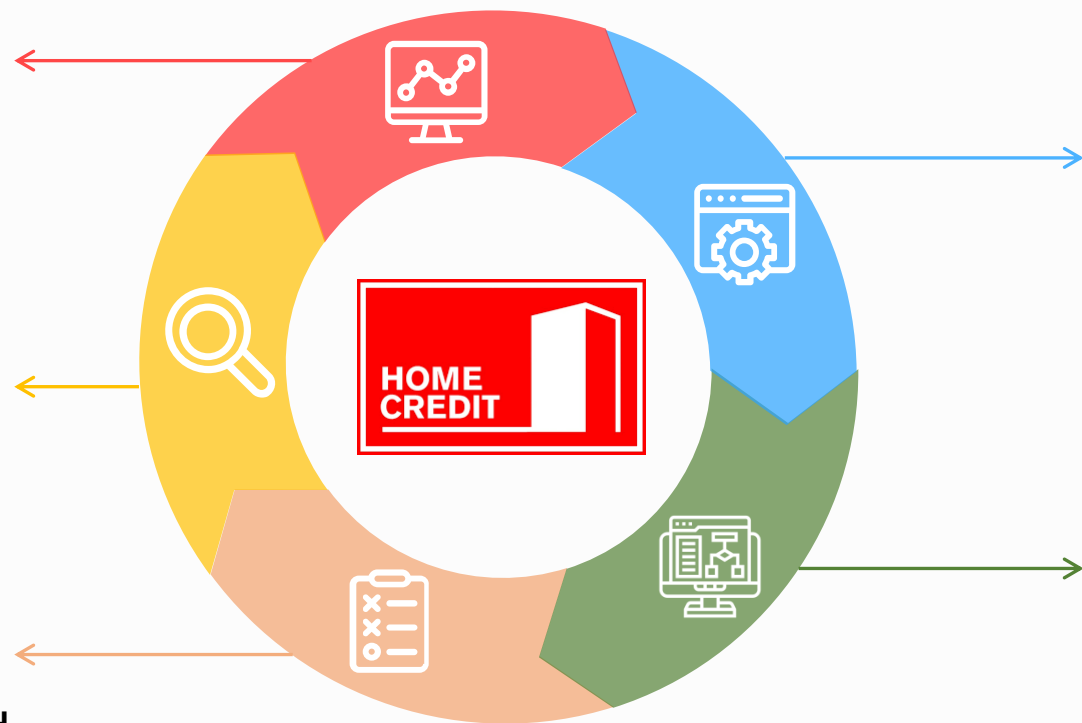
데이터 처리 및 분석 과정

03. 데이터 처리 및 분석 과정

01
EDA & Data
Preprocessing

05
개선방안 탐구

04
모델 평가 및 채택

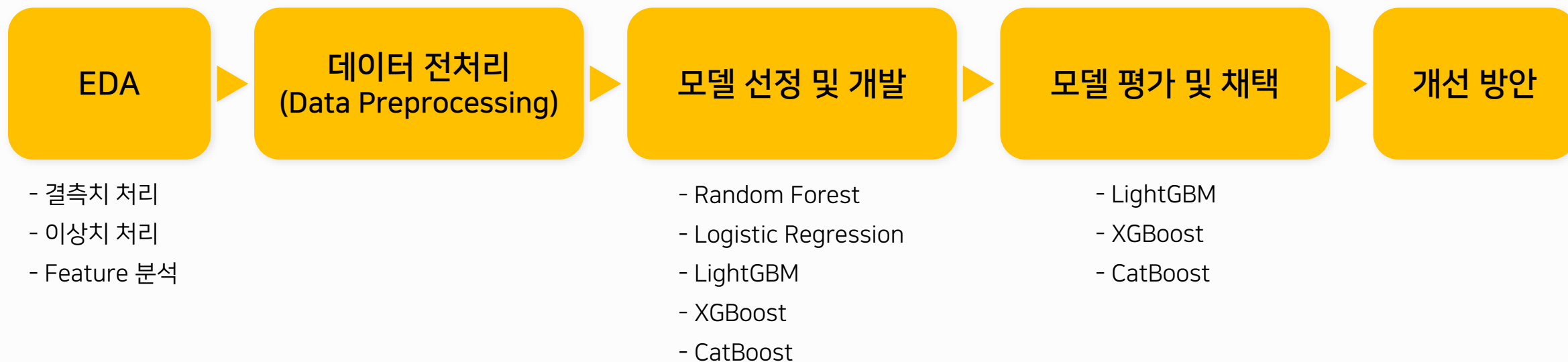


02
Feature
Engineering

03
모델 선정 및 개발

03. 데이터 처리 및 분석 과정

1차 데이터 처리 및 분석 과정



03. 데이터 처리 및 분석 과정

1차 데이터 처리 및 분석 과정 – EDA & Data Preprocessing

Feature 분석으로 컬럼에 대해 이해

결측치 확인 및 처리

이상치 처리

Groupby할 때 개인 판단으로 aggregation 값 지정

범주형 데이터 인코딩

- 2개: label encoding
- 3개 이상: one-hot encoding

03. 데이터 처리 및 분석 과정

1차 데이터 처리 및 분석 과정 - EDA & Data Preprocessing

결측치 확인

	Total	Percent
COMMONAREA_MEDI	248360	69.714109
COMMONAREA_AVG	248360	69.714109
COMMONAREA_MODE	248360	69.714109
NONLIVINGAPARTMENTS_MODE	246861	69.293343
NONLIVINGAPARTMENTS_MEDI	246861	69.293343
NONLIVINGAPARTMENTS_AVG	246861	69.293343
FONDKAPREMONT_MODE	243092	68.235393
LIVINGAPARTMENTS_MEDI	242979	68.203674
LIVINGAPARTMENTS_MODE	242979	68.203674
LIVINGAPARTMENTS_AVG	242979	68.203674
FLOORSMIN_AVG	241108	67.678489
FLOORSMIN_MEDI	241108	67.678489
FLOORSMIN_MODE	241108	67.678489

·
·
·

AMT_REQ_CREDIT_BUREAU_WEEK	47568	13.352234
AMT_REQ_CREDIT_BUREAU_DAY	47568	13.352234
AMT_REQ_CREDIT_BUREAU_HOUR	47568	13.352234
AMT_REQ_CREDIT_BUREAU_MON	47568	13.352234
NAME_TYPE_SUITE	2203	0.618377
OBS_30_CNT_SOCIAL_CIRCLE	1050	0.294733
OBS_60_CNT_SOCIAL_CIRCLE	1050	0.294733
DEF_60_CNT_SOCIAL_CIRCLE	1050	0.294733
DEF_30_CNT_SOCIAL_CIRCLE	1050	0.294733
EXT_SOURCE_2	668	0.187506
AMT_GOODS_PRICE	278	0.078034
AMT_ANNUITY	36	0.010105
CNT_FAM_MEMBERS	2	0.000561
DAYS_LAST_PHONE_CHANGE	1	0.000281

·
·
·

➡ application_{train|test}

✓ 결측치가 있는 컬럼 67개

03. 데이터 처리 및 분석 과정

1차 데이터 처리 및 분석 과정 - EDA & Data Preprocessing

결측치 확인

	Total	Percent
AMT_ANNUITY	1226791	71.473490
AMT_CREDIT_MAX_OVERDUE	1124488	65.513264
DAYS_ENDDATE_FACT	633653	36.916958
AMT_CREDIT_SUM_LIMIT	591780	34.477415
AMT_CREDIT_SUM_DEBT	257669	15.011932
DAYS_CREDIT_ENDDATE	105553	6.149573
AMT_CREDIT_SUM	13	0.000757

➡ bureau

- ✓ 결측치가 있는 컬럼 7개
- ✓ 결측치 51% 이상인 컬럼
['AMT_ANNUITY'] 제외

	Total	Percent
CNT_INSTALLMENT_FUTURE	1153	0.180066
CNT_INSTALLMENT	1152	0.179909
SK_DPD_DEF	1	0.000156
SK_DPD	1	0.000156
NAME_CONTRACT_STATUS	1	0.000156

➡ POS_CASH_balance

- ✓ 결측치가 있는 컬럼 5개

	Total	Percent
AMT_PAYMENT	2905	0.021352
DAYS_ENTRY_PAYMENT	2905	0.021352

➡ instalments_payments

- ✓ 결측치가 있는 컬럼 2개

03. 데이터 처리 및 분석 과정

1차 데이터 처리 및 분석 과정 – EDA & Data Preprocessing

결측치 확인

	Total	Percent
AMT_PAYMENT_CURRENT	767988	19.998063
AMT_DRAWINGS_OTHER_CURRENT	749816	19.524872
CNT_DRAWINGS_POS_CURRENT	749816	19.524872
CNT_DRAWINGS_OTHER_CURRENT	749816	19.524872
CNT_DRAWINGS_ATM_CURRENT	749816	19.524872
AMT_DRAWINGS_ATM_CURRENT	749816	19.524872
AMT_DRAWINGS_POS_CURRENT	749816	19.524872
CNT_INSTALMENT_MATURE_CUM	305236	7.948208
AMT_INST_MIN_REGULARITY	305236	7.948208

➡ credit_card_balance

- ✓ 결측치가 있는 컬럼 9개

	Total	Percent
RATE_INTEREST_PRIVILEGED	1664263	99.643698
RATE_INTEREST_PRIMARY	1664263	99.643698
RATE_DOWN_PAYMENT	895844	53.636480
AMT_DOWN_PAYMENT	895844	53.636480
NAME_TYPE_SUITE	820405	49.119754
DAYS_TERMINATION	673065	40.298129
NFLAG_INSURED_ON_APPROVAL	673065	40.298129
DAYS_FIRST_DRAWING	673065	40.298129
DAYS_FIRST_DUE	673065	40.298129
DAYS_LAST_DUE_1ST_VERSION	673065	40.298129
DAYS_LAST_DUE	673065	40.298129
AMT_GOODS_PRICE	385515	23.081773
AMT_ANNUITY	372235	22.286665
CNT_PAYMENT	372230	22.286366
PRODUCT_COMBINATION	346	0.020716
AMT_CREDIT	1	0.000060

➡ previous_application

- ✓ 결측치가 있는 컬럼 16개
- ✓ 결측치가 51% 이상인 컬럼
['RATE_INTEREST_PRIVILEGED'],
['RATE_INTEREST_PRIMARY'],
['RATE_DOWN_PAYMENT'],
['AMT_DOWN_PAYMENT'] 제외

03. 데이터 처리 및 분석 과정

1차 데이터 처리 및 분석 과정 - EDA & Data Preprocessing

결측치 처리

- ✓ 결측치를 처리할 근거가 있는 경우 합리적 접근법으로 결측치 처리

(예)

	Total	Percent
AMT_PAYMENT_CURRENT	767988	19.998063
AMT_DRAWINGS_OTHER_CURRENT	749816	19.524872
CNT_DRAWINGS_POS_CURRENT	749816	19.524872
CNT_DRAWINGS_OTHER_CURRENT	749816	19.524872
CNT_DRAWINGS_ATM_CURRENT	749816	19.524872
AMT_DRAWINGS_ATM_CURRENT	749816	19.524872
AMT_DRAWINGS_POS_CURRENT	749816	19.524872
CNT_INSTALMENT_MATURE_CUM	305236	7.948208
AMT_INST_MIN_REGULARITY	305236	7.948208

['AMT_DRAWINGS_CURRENT'](총 인출량),
['CNT_DRAWINGS_CURRENT'](총 인출횟수)가 0인
경우에 대해 세부항목 인출량, 인출횟수가 NA로 확인됨

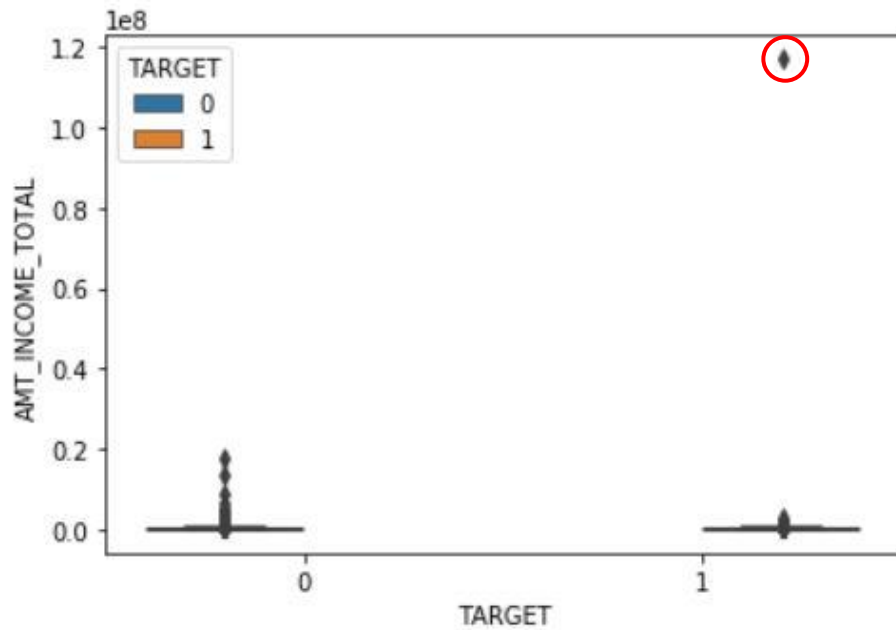
➡ 이 경우 NA를 0으로 대체함

03. 데이터 처리 및 분석 과정

1차 데이터 처리 및 분석 과정 - EDA & Data Preprocessing

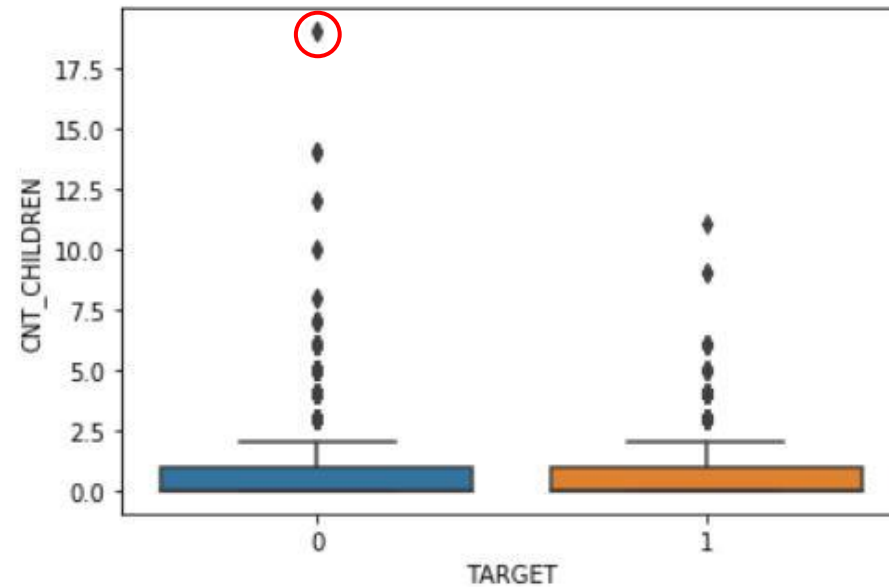
이상치 처리

AMT_INCOME_TOTAL



✓ 이상치 "117.000,000" 제거

CNT_CHILDREN



✓ 이상치 "19" 제거

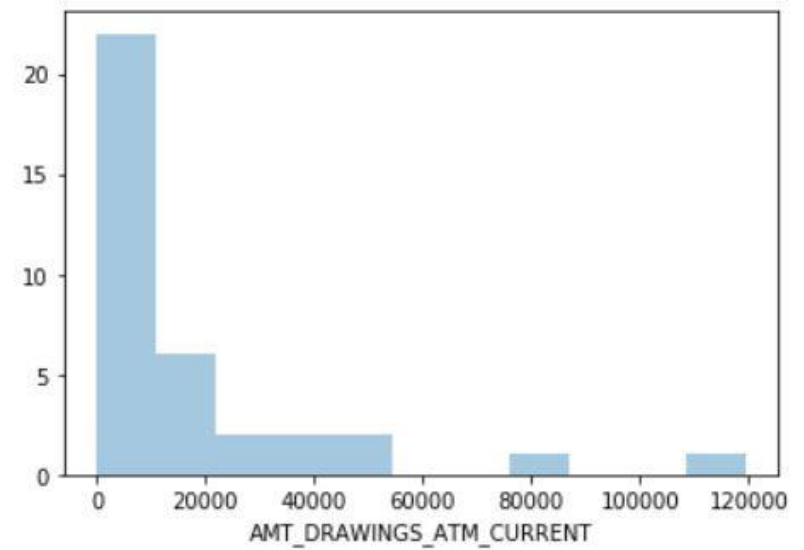
03. 데이터 처리 및 분석 과정

1차 데이터 처리 및 분석 과정 - EDA & Data Preprocessing

Groupby할 때 개인 판단으로 aggregation 값 지정

(예)

SK_ID_CURR	MONTHS_BALANCE
328243	-11
328243	-10
328243	-9
328243	-8
328243	-7
328243	-6
328243	-5
328243	-4
328243	-3
328243	-2



- ['MONTHS_BALANCE']는 max값으로 남은 대출기간을 알 수 있고, size로 지난 대출기간을 알 수 있음
- max와 size로 aggregation 결정

- ['AMT_DRAWINGS_ATM_CURRENT']는 0이 50% 이상을 차지함
- 중앙값 0, 평균 14062.5
- 평균으로 aggregation 결정

03. 데이터 처리 및 분석 과정

1차 데이터 처리 및 분석 과정 - 모델 선정 및 개발

Random Forest

- 트리 기반 Bagging 앙상블
- 사용성이 쉽고, 성능이 우수함

Logistic Regression

- 결과값이 범주형일 때 사용되는 회귀 분석 알고리즘
- 데이터의 결과가 특정 분류로 나뉘지는 회귀 분석으로 분류 기법으로 사용 가능

Light GBM

- Light Gradient Boosting Machine
- 속도가 매우 빠르고, 성능이 우수함

XGBoost

- eXtreme Gradient Boosting

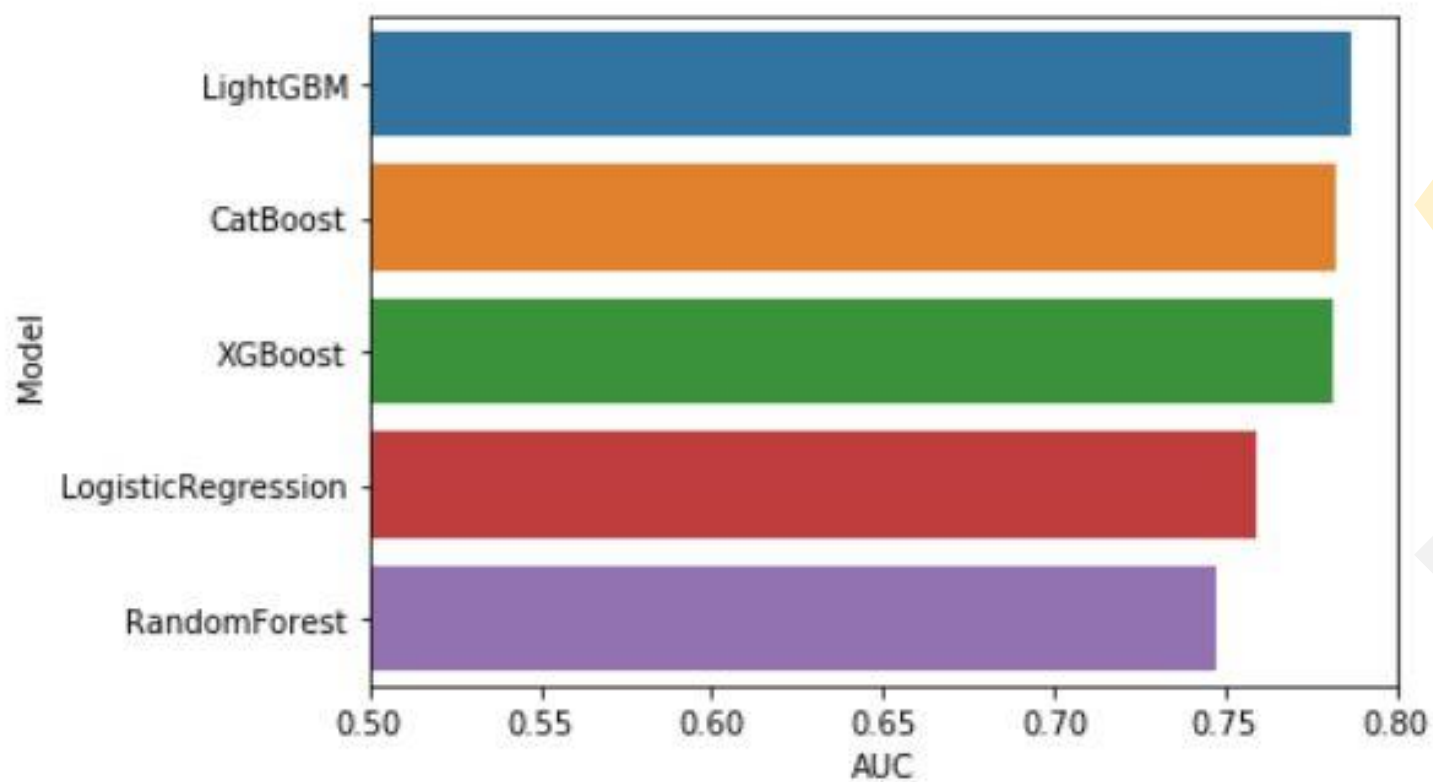
CatBoost

- Categorical Boosting

- ✓ 데이터 테이블을 하나씩 결합하면서 각 모델에 적합한 Data Set을 선정
- ✓ Light GBM은 Bayesian Optimization 방법, 그 외 모델은 Grid Search 방법을 통해 Hyperparameter를 조정해 성능 향상시킴

03. 데이터 처리 및 분석 과정

1차 데이터 처리 및 분석 과정 - 모델 평가 및 채택



LightGBM, CatBoost,
XGBoost 채택

성능이 좋지 않은 하위 모델
RandomForest,
LogisticRegression 제외

03. 데이터 처리 및 분석 과정

1차 데이터 처리 및 분석 과정 - 개선 방안

1 Grid Search 방법 단점

- ✓ 탐색 대상 hyperparameter 개수를 한 번에 많이 사용할수록, 탐색 시간이 기하급수적으로 증가
- ✓ 다음에 시도할 hyperparameter 값을 선정하는 과정에서, 이전 조사에서 얻어진 hyperparameter 값의 성능 결과에 대한 '사전 지식' 미반영



- ✓ Bayesian Optimization을 다른 모델에도 적용
- ✓ 매 회 조사 대상 선정을 자동화하고, 확률적 추정을 통해 '사전 지식'을 충분히 반영

2 개인 판단에 따른 aggregation 값 지정



- ✓ 통합적으로 aggregation 값 지정

03. 데이터 처리 및 분석 과정

1차 데이터 처리 및 분석 과정 - 개선 방안

3 instalments_payments 테이블에만 새로운 컬럼 추가

✓ 컬럼 추가 시 성능이 향상되었음

(예)

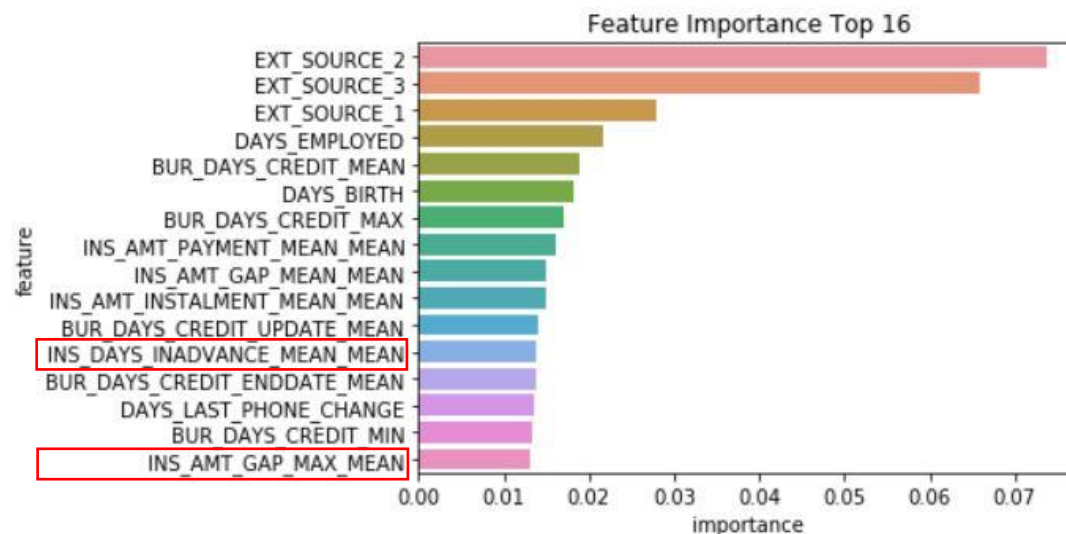
납입일과 실제 납입일 차이

```
ins['DAYS_INADVANCE'] = (ins['DAYS_INSTALLMENT'] - ins['DAYS_ENTRY_PAYMENT'])
```

할부금과 실제 납입금 차이

```
ins['AMT_GAP'] = (ins['AMT_INSTALLMENT'] - ins['AMT_PAYMENT'])
```

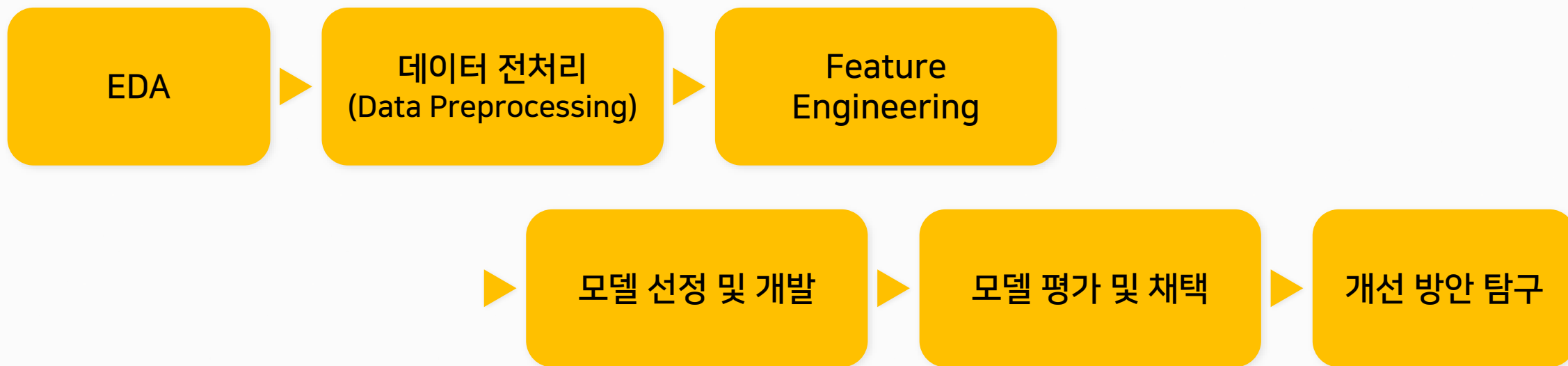
✓ 컬럼 추가 시 성능이 향상되었음



✓ 모든 테이블에 새로운 컬럼 추가

03. 데이터 처리 및 분석 과정

2차 데이터 처리 및 분석 과정



03. 데이터 처리 및 분석 과정

2차 데이터 처리 및 분석 과정 – EDA & Data Preprocessing

Feature 심화 분석 : 컬럼에 대한 보다 깊은 이해

결측치 확인: 결측치 비율이 51% 초과하는 컬럼을 제외하지 않은 Data Set도 만들어 비교

이상치 처리, 범주형 데이터 인코딩: 1차와 동일

Groupby할 때 통합적으로 aggregation 값 지정

- 수치형 데이터: min, max, mean, median, sum, size
- 범주형 데이터: mean, sum

03. 데이터 처리 및 분석 과정

2차 데이터 처리 및 분석 과정 – Feature Engineering

Feature 심화 분석을 통해 새로운 Feature 추가

Polynomial Feature 추가

Feature Selection

- Feature들 간의 상관관계(Correlation)가 기준(0.9)보다 높으면 둘 중 하나를 제외
- Light GBM에서 Feature importance가 0인 Feature 제외
- PCA(Principal Component Analysis, 주성분분석)

03. 데이터 처리 및 분석 과정

2차 데이터 처리 및 분석 과정 - Feature Engineering

Feature 심화 분석을 통해 새로운 Feature 추가

(예) 한도에 대한 카드대금 비율

```
df['AMT_BALANCE_RATIO'] = df['AMT_BALANCE'] / df['AMT_CREDIT_LIMIT_ACTUAL']
```

1회당 인출량

```
df['ONCE_DRAWINGS_ATM_CURRENT'] = df['AMT_DRAWINGS_ATM_CURRENT'] / df['CNT_DRAWINGS_ATM_CURRENT']  
df['ONCE_DRAWINGS_CURRENT'] = df['AMT_DRAWINGS_CURRENT'] / df['CNT_DRAWINGS_CURRENT']  
df['ONCE_DRAWINGS_OTHER_CURRENT'] = df['AMT_DRAWINGS_OTHER_CURRENT'] / df['CNT_DRAWINGS_OTHER_CURRENT']  
df['ONCE_DRAWINGS_POS_CURRENT'] = df['AMT_DRAWINGS_POS_CURRENT'] / df['CNT_DRAWINGS_POS_CURRENT']
```

credit_card_balance
테이블

(예) 연체일 = 최초 상환 예정일 - 실제 상환일

```
df['DAYS_FIRST_OVERDUE'] = df['DAYS_FIRST_DUE'] - df['DAYS_FIRST_DRAWING']
```

지불기간 = 최종 지불일 - 최초 상환 예정일

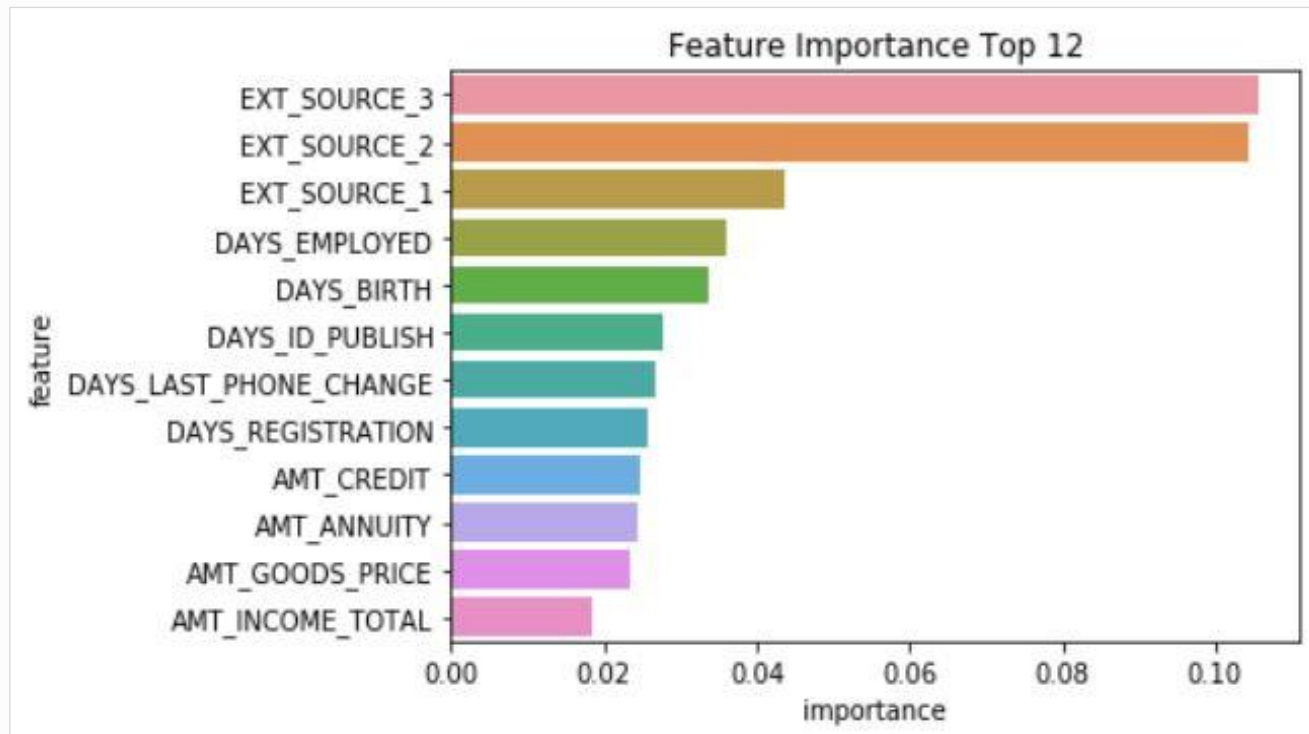
```
df['DAYS_PAYMENT_PERIOD'] = df['DAYS_LAST_DUE'] - df['DAYS_FIRST_DUE']
```

previous_application
테이블

03. 데이터 처리 및 분석 과정

2차 데이터 처리 및 분석 과정 - Feature Engineering

Polynomial Feature 추가



Random Forest와 Light GBM에서
Feature Importance 상위 12개의
Feature 선정

degree=2인
Polynomial Feature를 추가

03. 데이터 처리 및 분석 과정

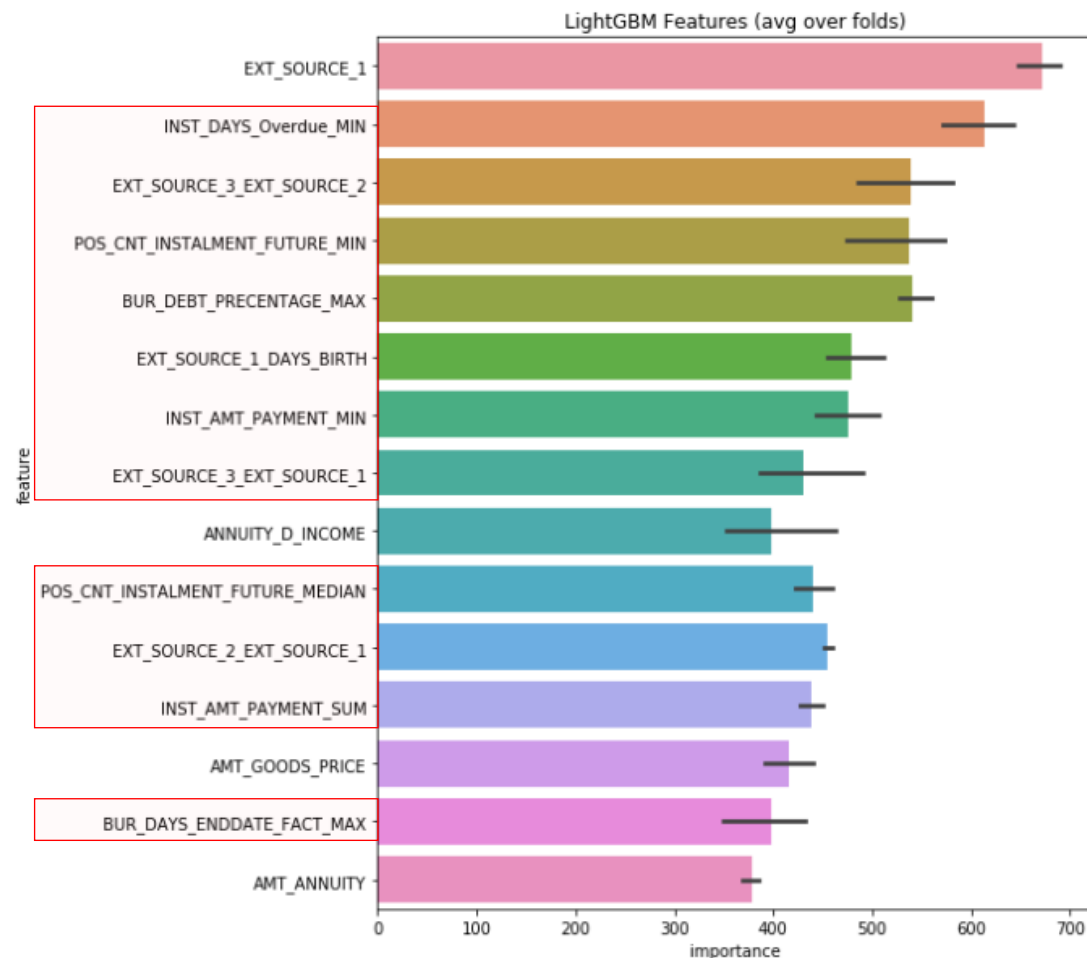
2차 데이터 처리 및 분석 과정 - Feature Engineering

Feature 추가 결과

Light GBM 모델에서 K-fold 교차 검증으로 구한
Feature importance 상위 15위에서

- ✓ **통합적인 aggregation**
- ✓ **Feature 심화 분석**
- ✓ **polynomial Feature**로 11개 컬럼 새로 추가

→ 다른 모델에서도 유의미한 영향을 주면서 1차에
비해 성능 향상



03. 데이터 처리 및 분석 과정

2차 데이터 처리 및 분석 과정 - Feature Engineering

Feature Selection

- Feature Engineering으로 Feature를 추가해 총 1293개 컬럼을 갖는 Data Set 생성
- 불필요한 Feature 제거 필요



Feature Selection

- Feature들 간의 상관관계(Correlation)가 기준(0.9)보다 높으면 둘 중 하나를 제외
- Light GBM에서 Feature importance가 0인 Feature 제외
- PCA(Principal Components Analysis)

→ Feature Selection 방법 3가지를 모두 적용한 Data Set을 만들 계획이었으나 한 단계마다 컬럼 수가 급격히 감소하여 한번에 3가지를 모두 적용하지 않고 1~2가지만 적용하여 6가지 종류의 Data Set을 구축

03. 데이터 처리 및 분석 과정

2차 데이터 처리 및 분석 과정 - Feature Engineering

Feature Engineering을 통해 만들어진 Data Set

Data Set	설명	Feature 개수
df-fi_d	<ul style="list-style-type: none">- Bureau, previous_application에서 결측치 비율이 51% 초과하는 컬럼 제외- Feature importance가 0인 컬럼 제외	643
df-fi_n	<ul style="list-style-type: none">- Feature importance가 0인 컬럼 제외	639
df-corr-fi_d	<ul style="list-style-type: none">- Bureau, previous_application에서 결측치 비율이 51% 초과하는 컬럼 제외- Feature사이 상관관계가 0.9이상인 경우 둘 중 하나 제외- Feature importance가 0인 컬럼 제외	431
df-corr-fi_n	<ul style="list-style-type: none">- Feature사이 상관관계가 0.9이상인 경우 둘 중 하나 제외- Feature importance가 0인 컬럼 제외	454
df-fi_pca_d	<ul style="list-style-type: none">- Bureau, previous_application에서 결측치 비율이 51% 초과하는 컬럼 제외- Feature importance가 0인 컬럼 제외- PCA를 이용하여 분산의 비율이 0.999로 유지하는데 필요한 최소한의 차원으로 차원 축소	326
df-fi_pca_n	<ul style="list-style-type: none">- Feature importance가 0인 컬럼 제외- PCA를 이용하여 분산의 비율이 0.999로 유지하는데 필요한 최소한의 차원으로 차원 축소	313

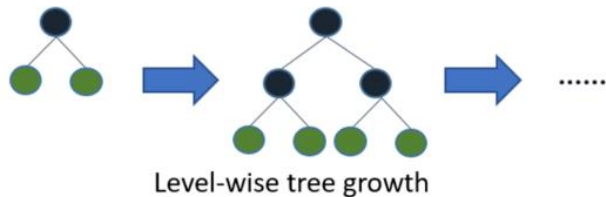
03. 데이터 처리 및 분석 과정

2차 데이터 처리 및 분석 과정 - 모델 선정 및 개발

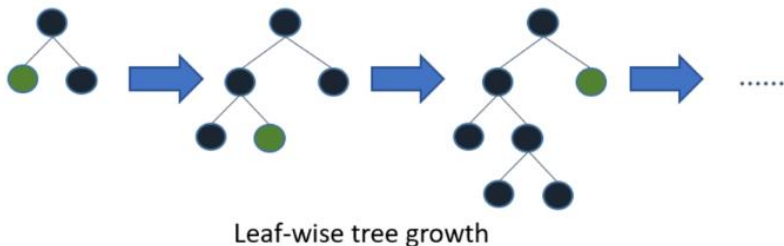
Light GBM

모델 소개

Microsoft
LightGBM



- 기존의 알고리즘 Level-wise tree growth 방식:
나무를 수평으로 확장하여 포화 트리를 만드는 방향으로 학습



- Light GBM은 Leaf-wise tree growth 방식:
 - 나무를 수직적으로 확장
 - 확장하기 위해서 max delta loss를 가진 leaf를 선택해 level-wise 알고리즘보다 더 많은 손실을 줄일 수 있음
- 모델 특징: 데이터 크기가 작은 경우 과적합되기 쉬움
→ 프로젝트 데이터 개수가 많아서 LGBM 사용

03. 데이터 처리 및 분석 과정

2차 데이터 처리 및 분석 과정 - 모델 선정 및 개발

Light GBM

Data Set 선정

Microsoft
LightGBM

Data Set	AUC
df-fi-d	0.78313
df-fi-n	0.78334
df-corr-fi-d	0.78249
df-corr-fi-n	0.78272
df-fi-pca-d	0.76167
df-fi-pca-n	0.76341

- ✓ PCA를 이용하여 규모를 축소한 Data Set은 AUC 점수가 낮아 제외



Data Set	AUC	Feature 개수	시간(초)
df-fi-d	0.78961	643	2536
df-fi-n	0.78958	639	2423
df-corr-fi-d	0.78901	431	1623
df-corr-fi-n	0.78952	454	1666

- ✓ 상관관계가 높은 컬럼을 제거한 Data Set을 이용한 경우 시간은 적게 소모되고 AUC 점수는 비슷함
- ✓ 4개의 Data Set을 계속 이용

03. 데이터 처리 및 분석 과정

2차 데이터 처리 및 분석 과정 - 모델 선정 및 개발

Light GBM

Hyperparameter 조정

Microsoft
LightGBM

- Bayesian Optimization을 이용

```
time_start = time.time()

bo.maximize(init_points=3, n_iter=5)

time_end = time.time()
```

iter	target	colsam...	learn...	max_depth	min_ch...	min_sp...	num_le...	reg_alpha	reg_la...	subsample
1	0.7873	0.9483	0.01476	8.946	38.89	0.0128	33.06	0.04397	0.06841	0.9635
2	0.7876	0.8387	0.01542	7.469	39.87	0.0199	34.98	0.04643	0.07184	0.8713
3	0.7876	0.855	0.01031	7.527	38.82	0.01352	34.2	0.03085	0.07051	0.8869
4	0.7876	0.8738	0.01216	8.999	38.04	0.01499	33.91	0.03477	0.0749	0.9075
5	0.7877	0.8378	0.01477	7.219	39.19	0.02134	34.97	0.04785	0.06086	0.9174
6	0.7874	0.9011	0.01744	8.881	39.7	0.01991	33.11	0.04461	0.0697	0.9957
7	0.7878	0.8235	0.01028	7.15	38.23	0.01832	34.77	0.0478	0.0771	0.8715
8	0.7873	0.9709	0.01304	7.03	38.08	0.02453	33.51	0.03487	0.06158	0.8001
9	0.7874	0.9029	0.01379	7.272	39.94	0.0116	34.94	0.04635	0.07478	0.8103
10	0.7874	0.892	0.01775	8.91	38.07	0.01088	33.47	0.03399	0.0776	0.8886

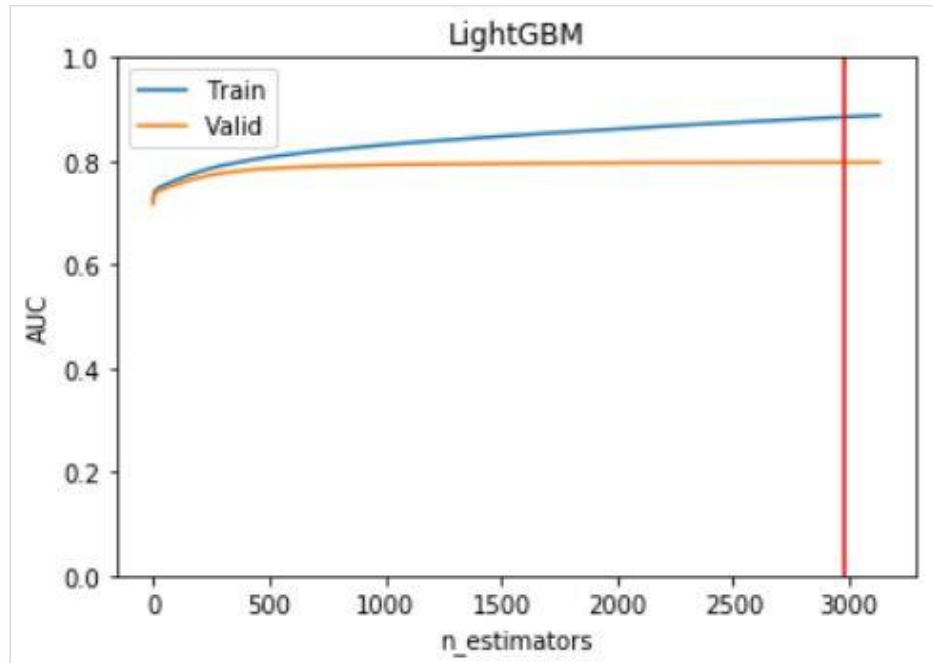
03. 데이터 처리 및 분석 과정

2차 데이터 처리 및 분석 과정 - 모델 선정 및 개발

Light GBM

Hyperparameter 조정

Microsoft
LightGBM



- Early stopping 기능을 통해 Validation Set의 AUC가 최대가 되는 최소한의 n_estimators 설정
- K-fold 교차 검증을 통해 예측 확률을 구한 후, 평균으로 최종 예측 확률을 구함

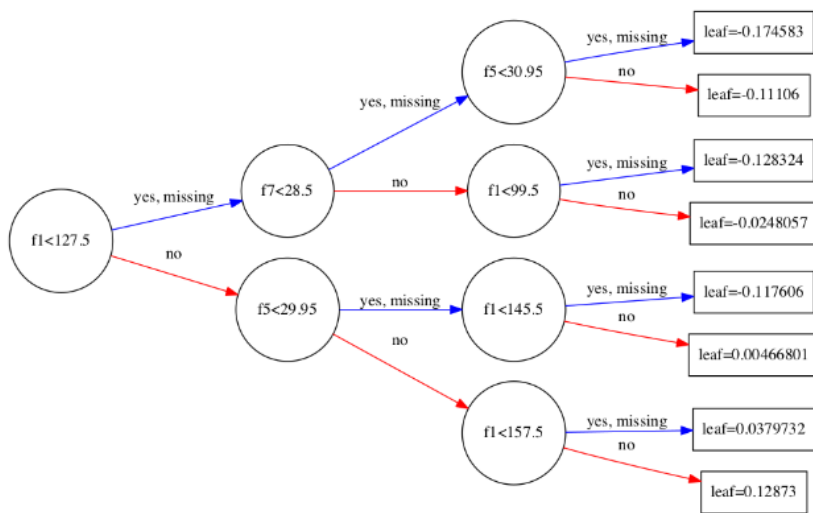
03. 데이터 처리 및 분석 과정

2차 데이터 처리 및 분석 과정 - 모델 선정 및 개발

XGBoost

모델 소개

dmlc
XGBoost



모델 특징

- eXtreme Gradient Boosting Library
- 병렬 처리를 사용해 학습과 분류가 빠름
- 유연성이 좋음. 평가 함수를 포함해 다양한 커스텀 최적화 옵션 제공
- Greedy-algorithm을 사용한 자동 가지치기 기능으로 과적합이 적게 발생
- 다른 알고리즘과 연계 활용성이 좋음

03. 데이터 처리 및 분석 과정

2차 데이터 처리 및 분석 과정 - 모델 선정 및 개발

XGBoost

Data Set 선정

dmlc
XGBoost

Data Set	시간(초)	AUC
df-fi-d	1957	0.78725
df-fi-n	1961	0.78583
df_corr-fi-d	1257	0.78499
df_corr-fi-n	1284	0.78539
df-fi-pca-n	1761	0.75946
df-fi-pca-d	1772	0.75867

- ✓ PCA를 이용하여 규모를 축소한 Data Set은 AUC 점수가 낮아 제외



Data Set	AUC
df-fi-d	0.78725
df-fi-n	0.78583
df_corr-fi-d	0.78505
df_corr-fi-n	0.78539

- ✓ 상관관계 높은 컬럼을 제거한 Data Set이 feature importance만 한 Data Set보다 지속적으로 점수가 낮아 feature importance 처리한 2개 Data Set 이용

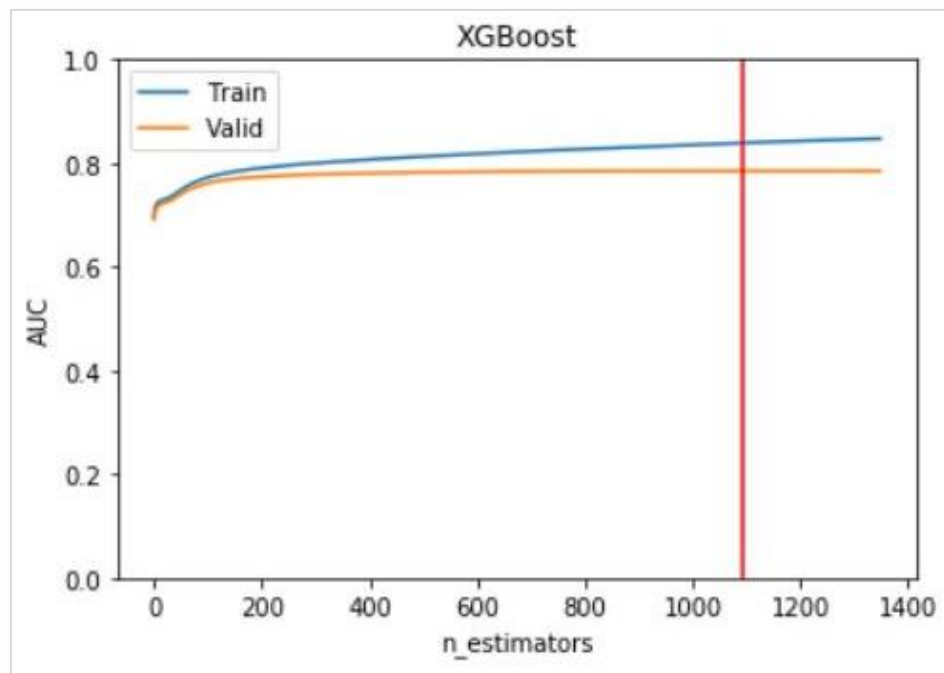
03. 데이터 처리 및 분석 과정

2차 데이터 처리 및 분석 과정 - 모델 선정 및 개발

XGBoost

Hyperparameter 조정

dmlc
XGBoost



- Bayesian Optimization 이용
- Early stopping 기능을 통해 Validation Set의 AUC가 최대가 되는 최소한의 n_estimators 설정
- K-fold 교차 검증을 통해 예측 확률을 구한 후, 평균으로 최종 예측 확률을 구함

03. 데이터 처리 및 분석 과정

2차 데이터 처리 및 분석 과정 - 모델 선정 및 개발

CatBoost

모델 소개



모델 특징

- 트리 기반 그래디언트 부스팅 모델
- 범주형 데이터 자동 one-hot encoding 기능
- Hyperparameter 튜닝 없이도 높은 성능
- 다른 모델에 비해 빠른 속도

03. 데이터 처리 및 분석 과정

2차 데이터 처리 및 분석 과정 - 모델 선정 및 개발

CatBoost

Data Set 선정



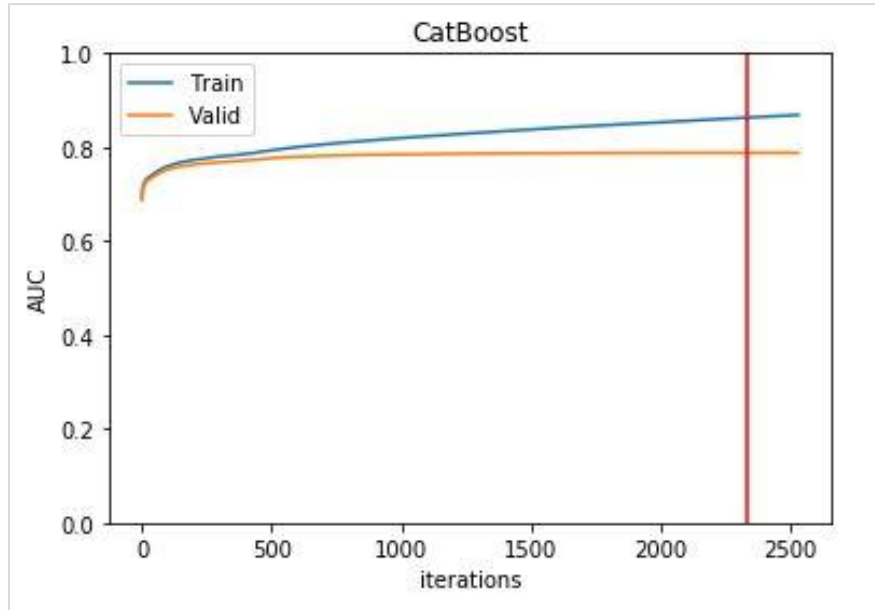
Data Set	컬럼 수	모델	파라미터	시간(초)	AUC	선택여부
df-fi_d	643			534.36	0.78647	✗
df-fi_n	639			531.52	0.78874	✓
df_corr-fi_d	431			369.59	0.78662	✓
df_corr-fi_n	454	CatBoost	Default 교차검증 3회	381.36	0.78511	✗
df_pca-fi_d	326			369.04	0.78303	✗
df_pca-fi_n	313			345.24	0.78406	✗

03. 데이터 처리 및 분석 과정

2차 데이터 처리 및 분석 과정 - 모델 선정 및 개발

CatBoost

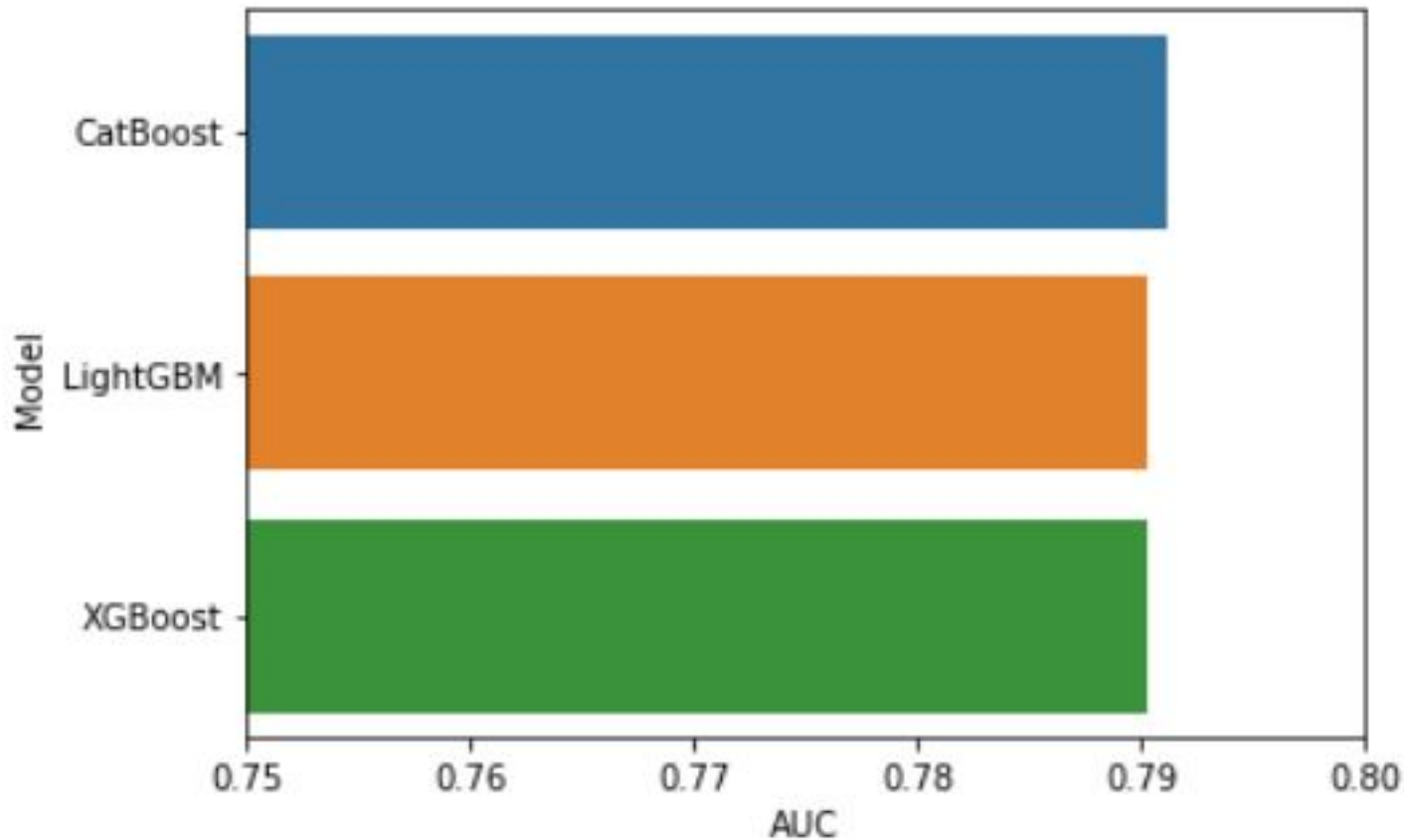
Hyperparameter 조정



- Bayesian Optimization 이용
- Early stopping 기능을 통해 Validation Set의 AUC가 최대가 되는 최소한의 iterations 설정
- K-fold 교차 검증을 통해 예측 확률을 구한 후, 평균으로 최종 예측 확률을 구함

03. 데이터 처리 및 분석 과정

2차 데이터 처리 및 분석 과정 - 모델 평가 및 채택



▶ CatBoost 0.79121

▶ LightGBM 0.79029

▶ XGBoost 0.79029

03. 데이터 처리 및 분석 과정

2차 데이터 처리 및 분석 과정 - 개선방안

- 1 SVM 등 더 다양한 분류 모델을 적용
- 2 Featuretools 라이브러리를 사용해서 자동화된 Feature Engineering 적용
- 3 Hyperparameter 조정 시 Grid Search, Random Search 등 다양한 방법 시도
- 4 Feature Engineering부터 단계적으로 '데이터 처리 및 분석과정' 수행

03. 데이터 처리 및 분석 과정

3차 데이터 처리 및 분석 과정 - Preview

CREDIT_ACTIVE	
Closed	1079273
Active	630607
Sold	6527
Bad debt	21

99% 차지 →

bureau 테이블

['CREDIT_ACTIVE'] 컬럼: CB에 보고된 신용거래 상태

- Closed: 비활성화된 신용거래
- Active: 활성화된 신용거래

→ 행 추출 후 aggregation으로 새로운 컬럼 생성

NAME_CONTRACT_STATUS	
Approved	1036781
Canceled	316319
Refused	290678
Unused offer	26436

79% 차지 →

previous_application 테이블

['NAME_CONTRACT_STATUS'] 컬럼: 이전 신청의 계약 상태

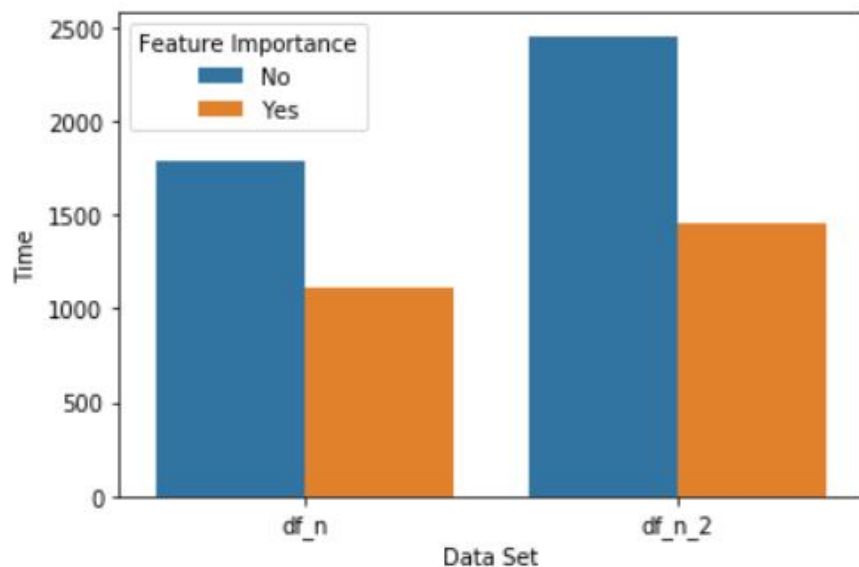
- Approved: 신용거래 계약 승인
- Refused: 신용거래 계약 거절

→ 행 추출 후 aggregation으로 새로운 컬럼 생성

03. 데이터 처리 및 분석 과정

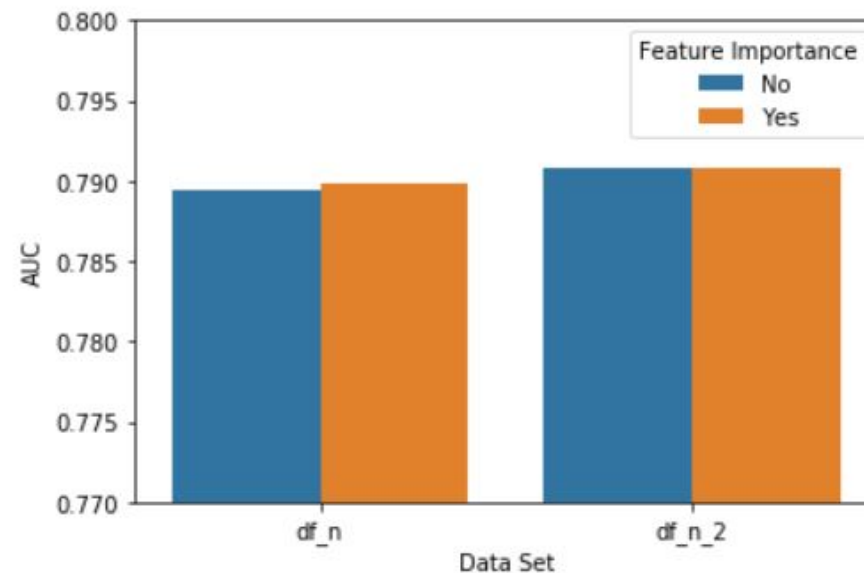
3차 데이터 처리 및 분석 과정 - Preview

LightGBM에서 Feature importance가 0인 Feature를 제외한 효과



<기대한 효과>

불필요한 Feature를 제거하여 시간 단축

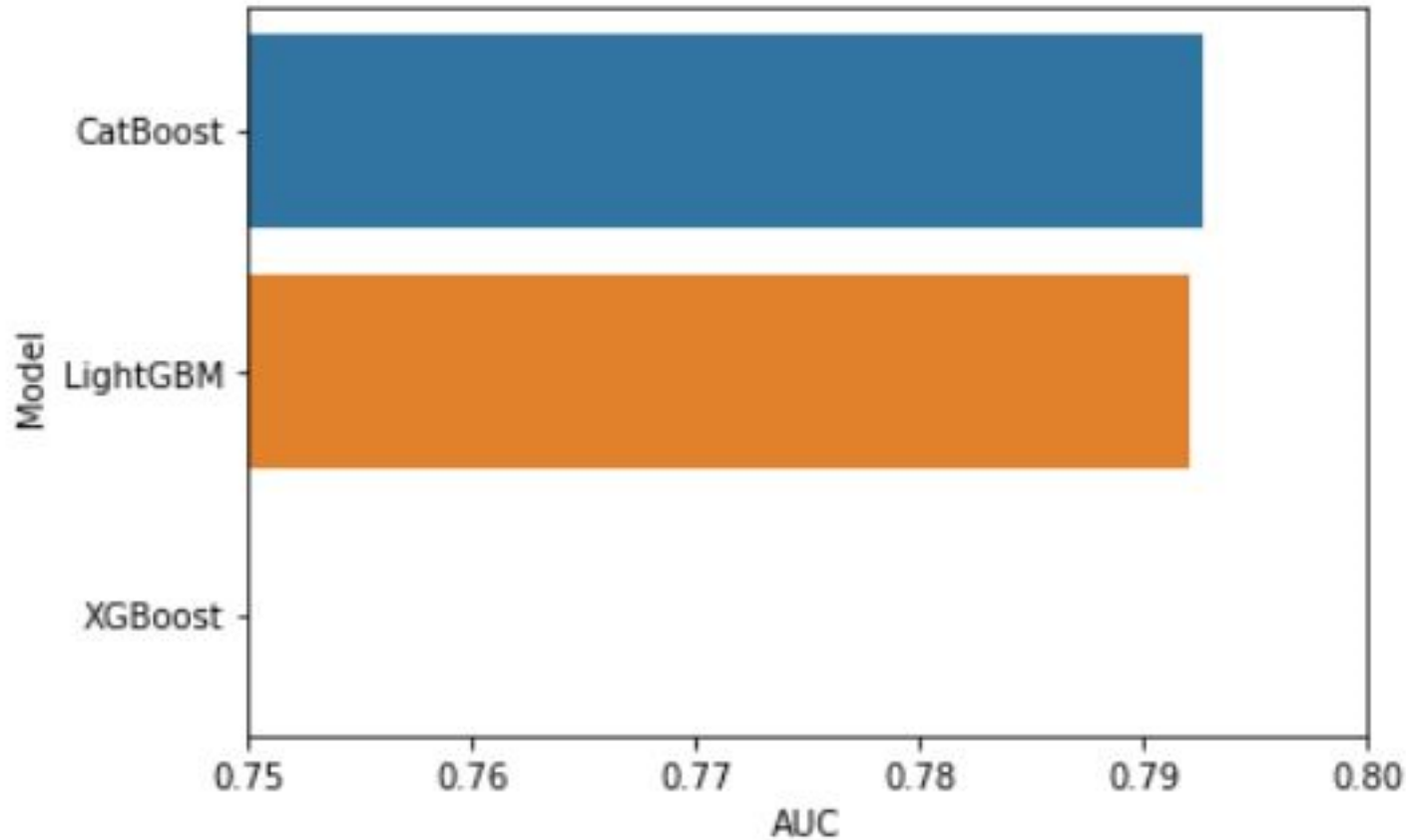


<부수적인 효과>

AUC도 높아져 성능 향상

03. 데이터 처리 및 분석 과정

3차 데이터 처리 및 분석 과정 - Preview



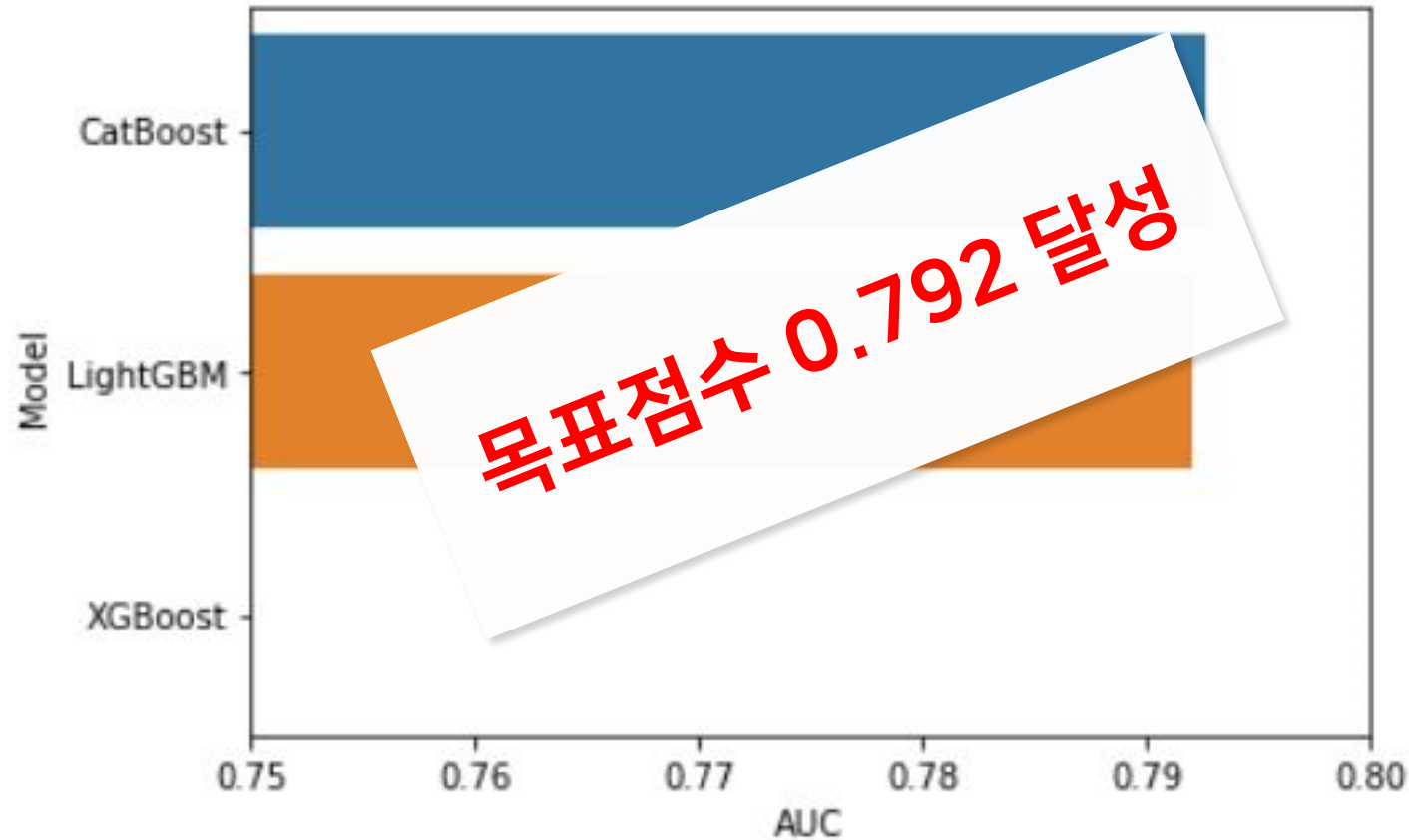
▶ CatBoost 0.79275

▶ LightGBM 0.79205

▶ XGBoost ...진행중...

03. 데이터 처리 및 분석 과정

3차 데이터 처리 및 분석 과정 - Preview



▶ CatBoost 0.79275

▶ LightGBM 0.79205

▶ XGBoost ...진행중...

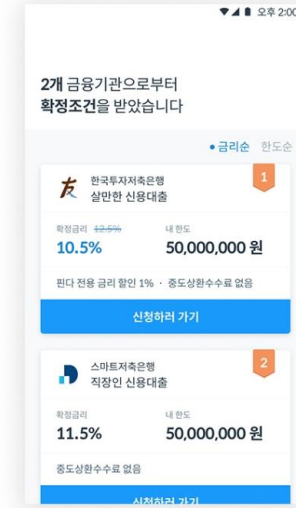
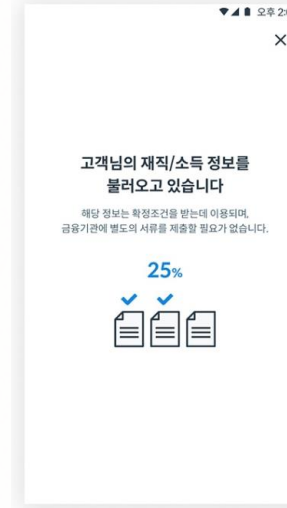
The background of the slide features a dark, grayscale image of a house with a gabled roof and several windows. In the foreground, a set of keys with a circular ring is visible, resting on a dark surface. The overall tone is professional and related to real estate or home ownership.

04

Home Credit Default Risk

서비스 활용 방안 및 기대 효과

04. 서비스 활용 방안



신용정보
데이터

+

개인 데이터

- ✓ 개인별 맞춤 신용대출 서비스 제공
- ✓ 대출 가능액 범위에 따라 개인별 대출상환 능력 예측 모델 개발

04. 서비스 기대 효과



회사 측면

비용절감
새로운 고객 유입



고객 측면

편리함
신속한 대출 가능

참고 문헌 및 사이트

HC 시작하기: <https://www.kaggle.com/willkoehrsen/start-here-a-gentle-introduction>

Manual Feature Engineering

<https://www.kaggle.com/willkoehrsen/introduction-to-manual-feature-engineering>

Automated Feature Engineering

<https://www.kaggle.com/willkoehrsen/automated-feature-engineering-basics>

lightgbm 시작: <https://www.kaggle.com/jsaguiar/lightgbm-with-simple-features>

"LightGBM: A Highly Efficient Gradient Boosting Decision Tree", Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu, "Neural Information Processing Systems 2017."

lightgbm 공식 홈페이지 <https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMClassifier.html>

Catboost 공식 홈페이지 <https://catboost.ai/>

<https://data-newbie.tistory.com/131>

xgboost 파라미터 설명 <http://machinelearningkorea.com/>

xgboost 사용법 <https://statkcle.github.io/model/model-python-xgboost-hyper.html>

xgboost 공식 홈페이지 <https://xgboost.ai/>

A close-up photograph of a person's hands holding a dark-colored Visa credit card. The card is held horizontally, with the Visa logo and the word "new" visible. The background is a dark, textured surface, likely a workbench, with various tools and objects scattered around, including a metal rod, a wooden handle, and some small metal pieces. A semi-transparent grey square with rounded corners is overlaid in the center of the image, containing the text "Q & A" in white.

Q & A