Bigdata fintech 4th

기계학습 4조

김태훈 남경혜 방수영 이상윤 정재영

기존 고객 대상 아웃바운드 콜의

자동차 보험 계약 성사 예측 모형

목차

CONTENTS

분석
목적

탐색
및
전처리

모델링

평가
및
결론

챕터1
CHAPTER1

분석 목적

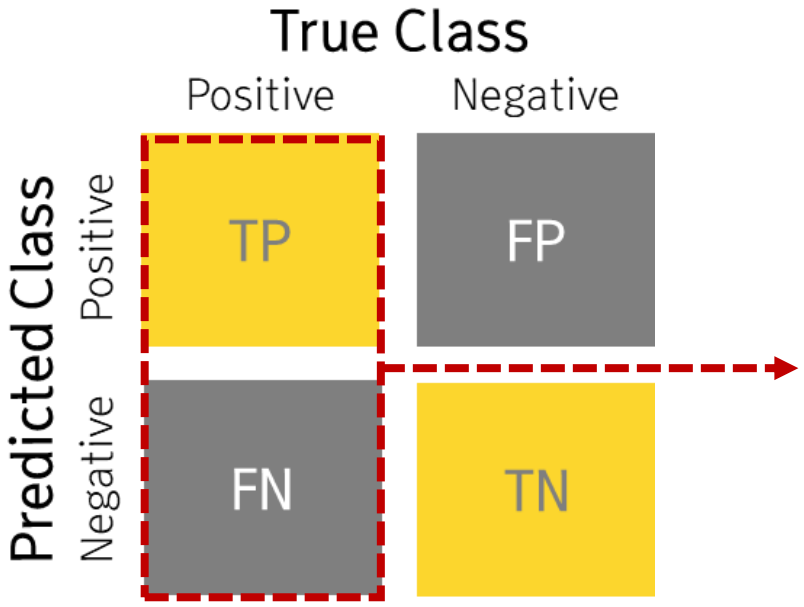
기존 은행이 보험으로 사업을 다각화하면서, 기존 고객들에게 보험 마케팅 콜을 했을 때의 계약 성사 여부를 학습하여 생성한 데이터를 바탕으로 새로운 마케팅에서 **계약 성사 가능성이 높은 고객을 분류 하는 것**이 목적

아웃바운드(outbound)콜의 **'선택'** 과 **'집중'** 이 가능해짐
→ 실적과 직결되기 때문에 실제 현업에서 중요하게 여기는 지표

아웃바운드(Outbound) : 상품 구매를 유도하는 전화 영업의 형태(전화를 거는 것)

인바운드(Inbound) : 고객에게서 걸려온 문의를 처리 및 응대(전화를 받는 것)

사전 지표 선정

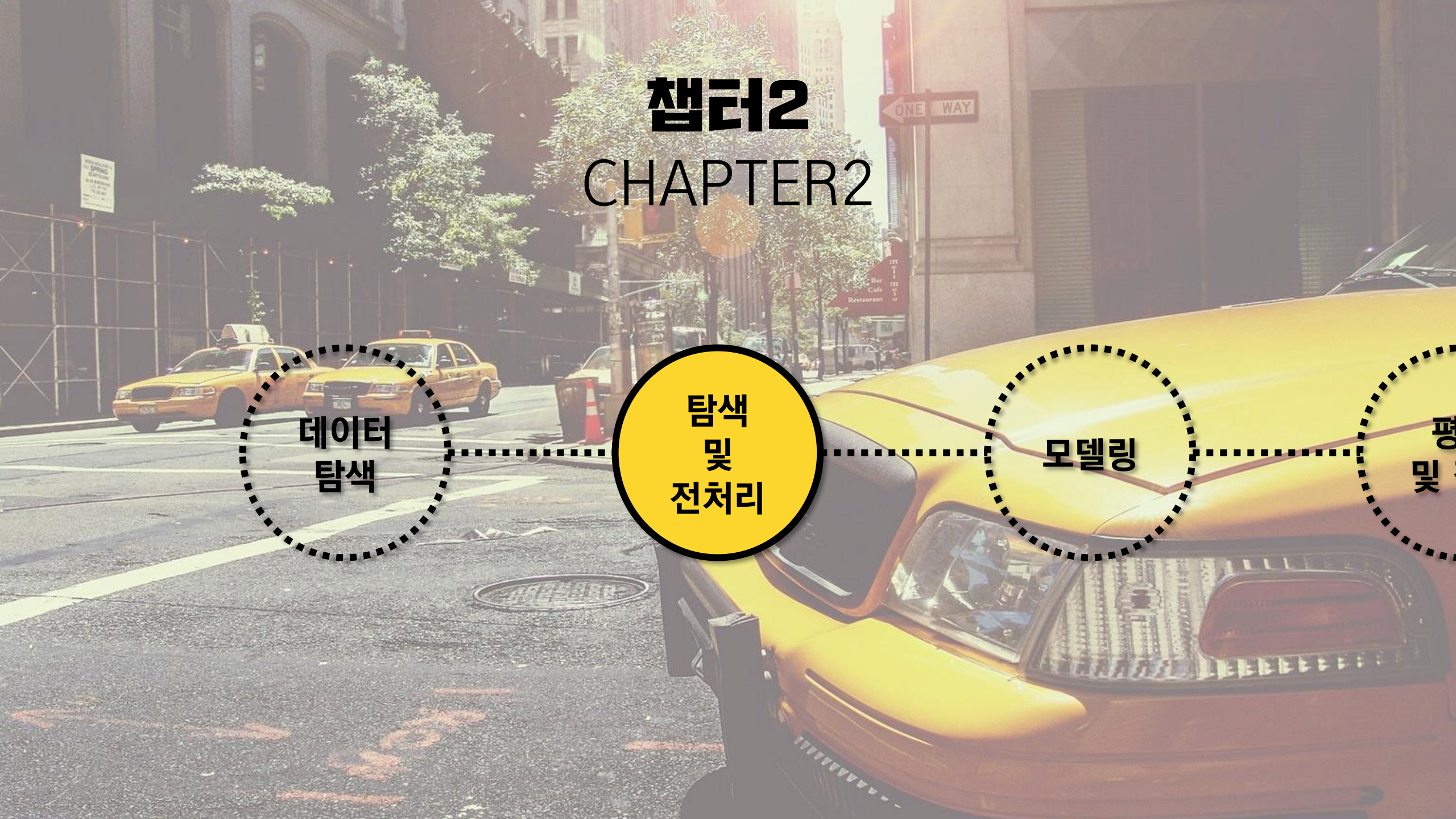


$$Recall = \frac{TP}{TP + FN}$$

계약이 성사 될 고객들에게
아웃바운드 콜이 갈 확률이 커진다

But, 아웃바운드 콜 대상 고객의 pool이 아주 커져 분류 의미가 사라짐.

Recall과 accuracy를 주요 지표로 삼아 모델의 퍼포먼스를 평가하자


챕터2
CHAPTER2
데이터
탐색
탐색
및
전처리
모델링

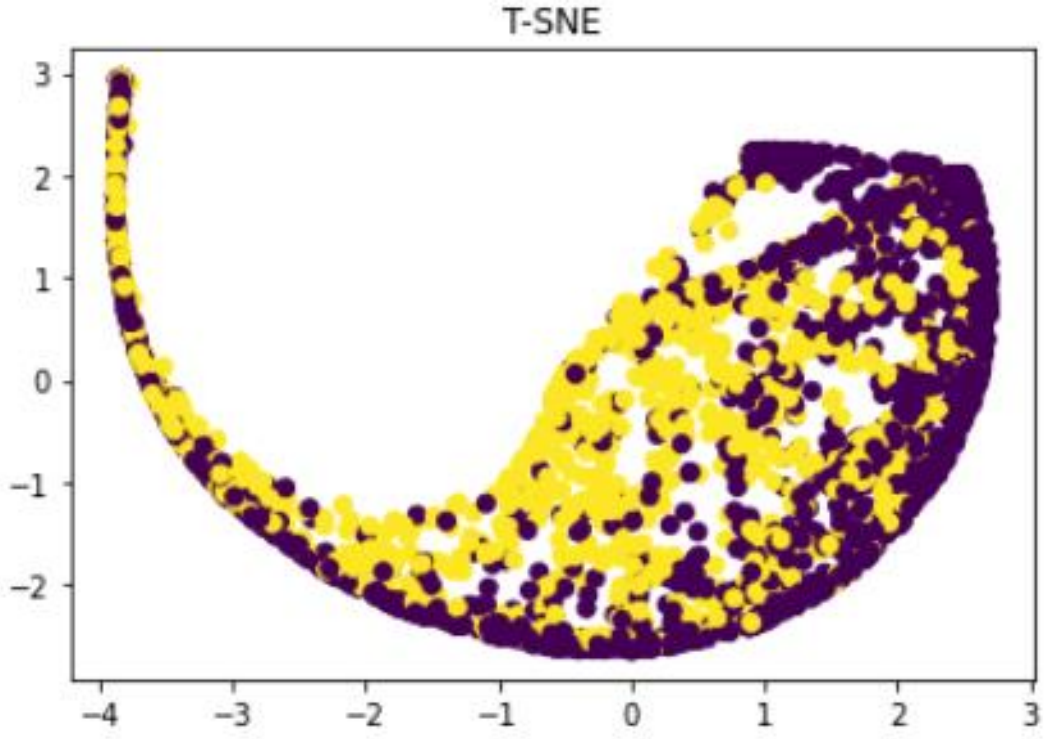
Data 탐색

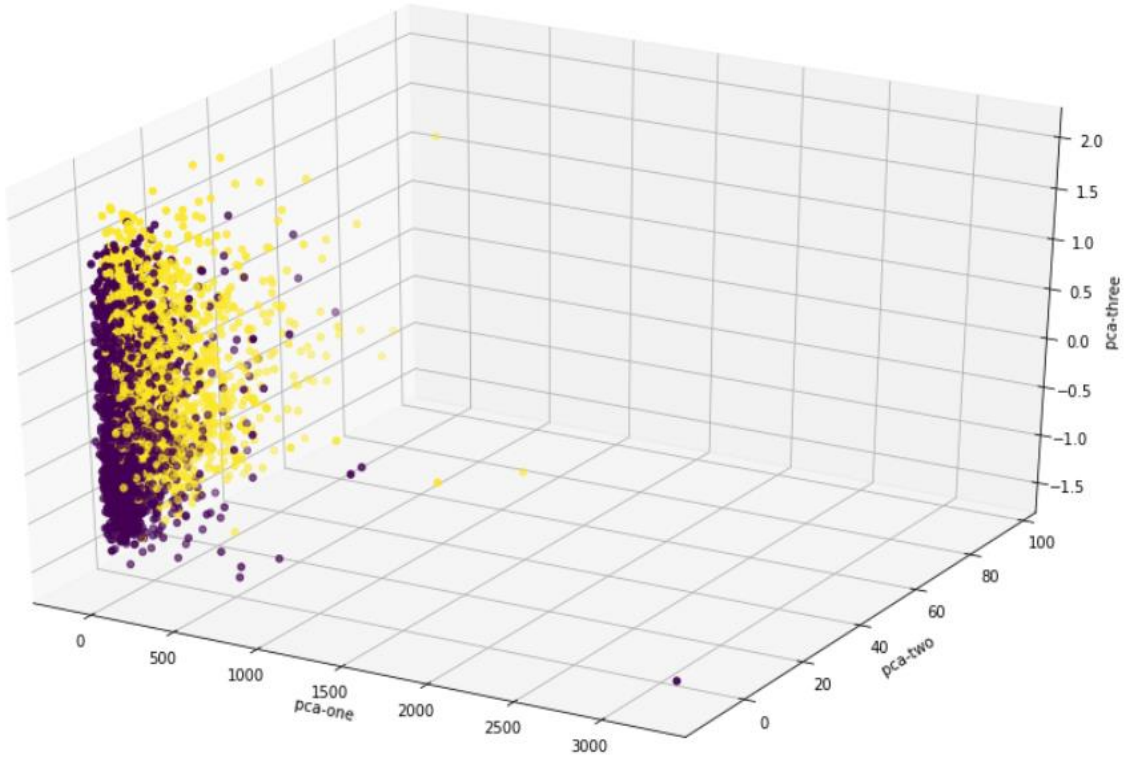
변수명	데이터형태	상세설명
ID	이산형	고객 식별 번호
Age	연속형	고객의 나이
Job	범주형	고객의 직업(admin/ blue-collar/…/등 10개 범주)
Marital	범주형	결혼여부(divorced/married/single)
Education	범주형	학력(primary/secondary/tertiary)
Default	범주형	파산여부(yes/no)
Balance	연속형	평균잔고
HHisurance	범주형	가계보험가입여부(yes/no)
Carloan	범주형	자동차대출여부(yes/no)
Communication	범주형	상담방식 (cellular/telephone/NA)
Lastcontactmonth	범주형	최근 접촉 월
Lastcontactday	범주형	최근 접촉 일
Callstart	연속형	통화 시작 시각
Callend	연속형	통화 종료 시각
Noofcontacts	연속형	현재 마케팅에서의 접촉 횟수
Dayspassed	연속형	이전 마케팅에서의 접촉 후 경과 시간
Prevattempts	연속형	현재 마케팅 이전 접촉 횟수
Outcome	범주형	이전 마케팅으로 인한 결과(성공/실패/보류/NA)
Carinsurance	범주형	보험가입여부 **종속변수 Y**

총 데이터 수 : 4000행 * 19열

출처 : Kaggle/car-insurance-cold-calls

데이터시각화 T-SNE/PCA를 이용하여 구현해본 데이터 분포





변수 생성 및 삭제

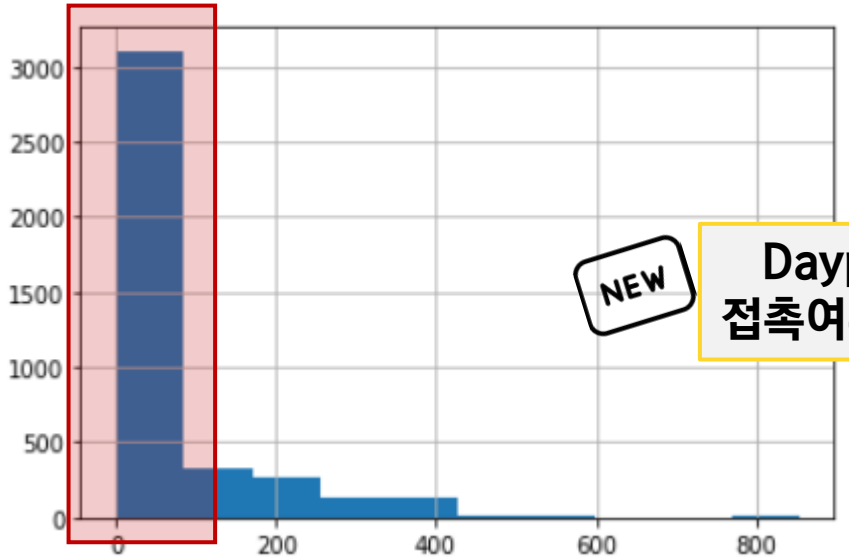
Callstart
전화를 받은 시각

Callend
전화를 끊은 시각

NEW **Call_duration**
콜지속시간

통화 시작/종료 시각이 아닌
통화 지속 시간이 **관심도를 반영**

Dayspassed(접촉횟수)



접촉 후 경과 시간이 아닌 **접촉 여부**가
유의미한 변수라고 판단

▶ Callstart, Callend, Daypassed 변수 삭제/Call_duration, Daypass 변수 생성

결측치 확인 및 처리

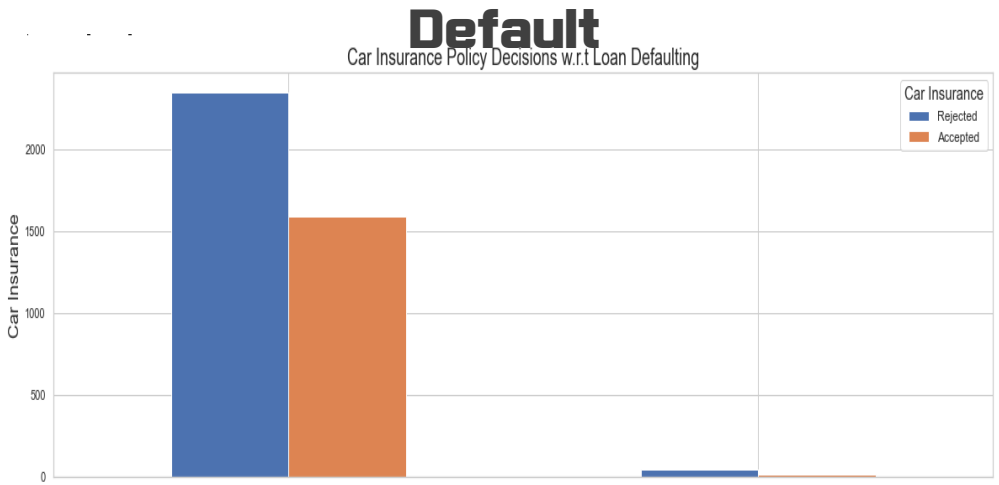
변수명	missing_value
ID	0
Age	19
Job	0
Marital	169
Education	0
Default	0
Balance	0
Hhnsurance	0
Carloan	902
Communication	0
Lastcontactmonth	0
Lastcontactday	0
Callstart	0
Callend	0
Noofcontacts	0
Dayspassed	0
Prevattempts	0
Outcome	3042

→ **최빈값**으로 대체

결측치 비율 20% 이상이나,
MCAR(Missing completely at Random)으로 판단

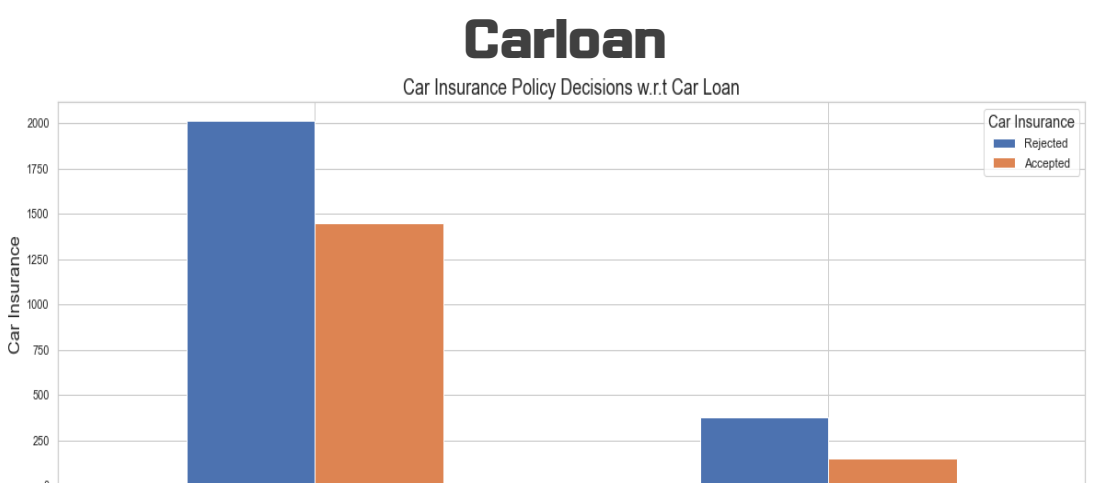
결측치의 비율이 **약 76%**인
'Outcome(소득)' 은 **삭제**하는 것이 맞다고 판단

범주별 종속변수 분포 확인- Default/Carloan(재정상태 관련 변수)



0(파산 안함) 1(파산함)

Car Insurance	0	1	Percentage Enrolled
Default			
0	2352	1590	40.33
1	44	14	24.14



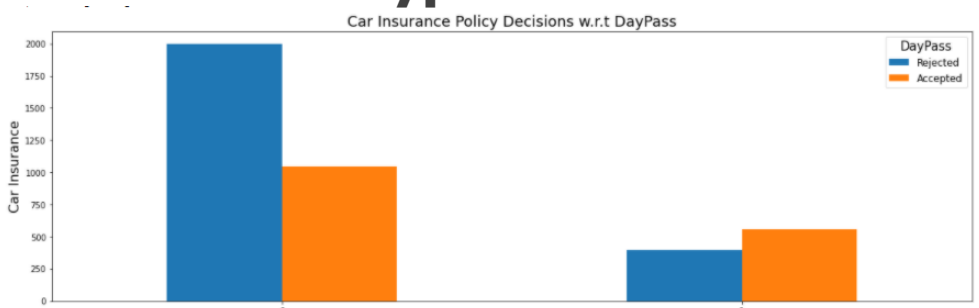
0(대출 없음) 1(대출 있음)

Car Insurance	0	1	Percentage Enrolled
CarLoan			
0	2016	1452	41.87
1	380	152	28.57

▶ 상대적으로 재정상황이 양호한 default ==0 & car loan == 0이 계약 성사율이 높았음

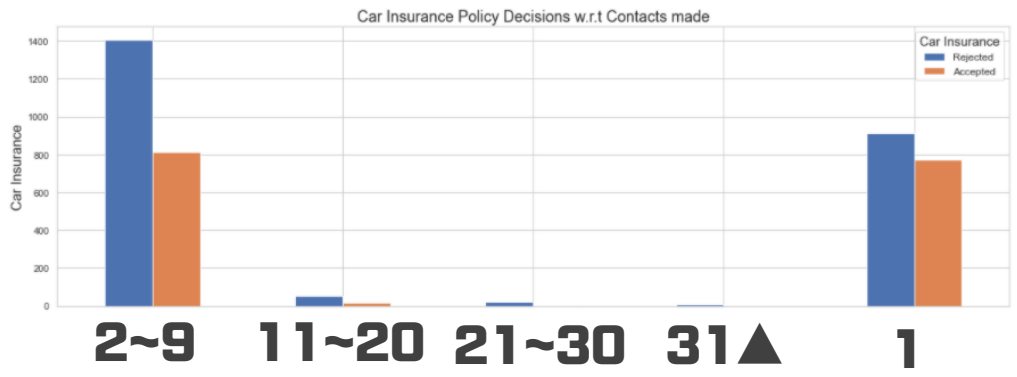
범주별 종속변수 분포 확인- Daypass, NoOfContact(마케팅 관련 지표)

Daypass(접촉여부)



DayPass	Car Insurance 0	1	Percentage Enrolled
0	1997	1045	34.35
1	399	559	58.35

NoOfContact(접촉횟수)

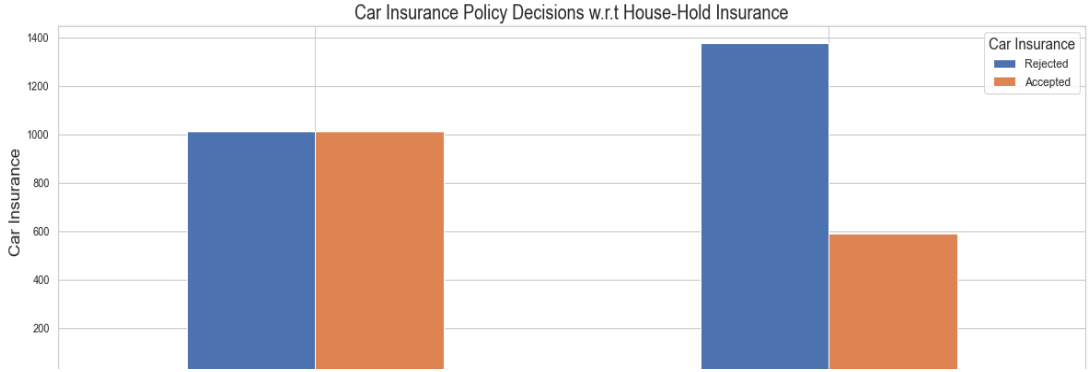


NoOfContacts_Category	Car Insurance 0	1	Percentage Enrolled
Contacted More than once	1406	814	36.67
Contacted more than 10 times	54	14	20.59
Contacted more than 20 times	19	3	13.64
Contacted more than 30 times	5	0	0.00
Contacted once	912	773	45.88

▶ 접촉을 하지 않는 경우보다 **접촉을 하는 경우가 성사율이 높으나**,
접촉횟수가 늘어날수록 성사율이 낮아지는 것을 확인할 수 있음

범주별 종속변수 분포 확인- HHInsurance

HHInsurance(가계보험 가입여부)



0(가입 안함) 1(가입함)

Car Insurance	0	1	Percentage Enrolled
HHInsurance			
0	1016	1013	49.93
1	1380	591	29.98

가계보험?
개인 · 가족의 건강 · 재산의 상실 등 위험에 대비하여 또는 가계의 유지와 생활의 안정을 위하여 개인이 계약하는 보험(기업보험의 반대)
출처:두산백과

▼

즉, 보장을 받고 있는 보험이 없다는 것.
그러므로 **아웃바운드 콜의 효과 증대 예상**

변수별 상관관계 확인


Age 1 0.0037 0.094 -0.18 -0.012 0.014 0.029 -0.017 -0.0025 0.013 0.0054
Default 0.0037 1 -0.054 -0.015 0.12 0.027 0.03 -0.033 -0.032 0.0036 -0.048
Balance 0.094 -0.054 1 -0.072 -0.082 0.016 -0.0052 -0.0021 0.0013 0.016 0.028
HHInsurance -0.18 -0.015 -0.072 1 0.088 -0.024 -0.0059 0.081 0.027 0.02 0.012
CarLoan -0.012 0.12 -0.082 0.088 1 0.011 0.021 -0.042 -0.021 0.042 -0.058
LastContactDay 0.014 0.027 0.016 -0.024 0.011 1 0.14 -0.074 -0.073 -0.021 -0.08
NoOfContacts 0.029 0.03 -0.0052 -0.0059 0.021 0.14 1 -0.1 -0.057 -0.061 -0.12
DaysPassed -0.017 -0.033 -0.0021 0.081 -0.042 -0.074 -0.1 1 0.5 -0.019 0.83
PrevAttempts -0.0025 -0.032 0.0013 0.027 -0.021 -0.073 -0.057 0.5 1 -0.0062 0.62
Call_Duration 0.013 0.0036 0.016 0.02 0.042 -0.021 -0.061 -0.019 -0.0062 1 -0.032
DayPass 0.0054 -0.048 0.028 0.012 -0.058 -0.08 -0.12 0.83 0.62 -0.032 1
Age Default Balance HHInsurance CarLoan LastContactDay NoOfContacts DaysPassed PrevAttempts Call_Duration DayPass
1.0 0.8 0.6 0.4 0.2 0.0

범주형 변수 처리 : Onehot 인코딩

범주형 변수

JOB

Marital

education

LastContactMonth

더미 변수

job_bluecollar/job_entreprenrur/job_housemaid/job_retired/job_management/
job_selfemployed/job_services/job_student/job_technician/job_unemployed(총10개)

Marital_married/Marital_single/Marital_divorced(총 3개)

Education_primary/Education_secondary/Education_tertiary(총 3개)

LastContactMonth_jan – dec(총 12개)

최종변수 확인 (1/4)

변수명	데이터형태	상세설명
Age	연속형	고객의 나이
default	범주형	파산여부(yes/no)
balance	연속형	평균잔고
HHinsurance	범주형	가계보험가입여부(yes/no)
Carloan	범주형	자동차대출여부(yes/no)
Communication	범주형	상담방식 (cellular/telephone)
lastcontactmonth	범주형	최근 접촉 월
Lastcontactday	범주형	최근 접촉 일
noofcontacts	연속형	해당 캠페인 기간 동안의 접촉 수
dayspassed	연속형	이전 캠페인으로부터 고객의 마지막 접촉 이후 지난 기간
prevattempts	연속형	캠페인 이전에 접촉 수

최종변수 확인 (2/4)

변수명	데이터형태	상세설명
job_bluecollar	더미형	고객의 직업(생산직 종사자)
job_entreprenrur	더미형	고객의 직업(기업가)
job_housemaid	더미형	고객의 직업(주부)
job_retired	더미형	고객의 직업(은퇴자)
job_management	더미형	고객의 직업(경영인)
job_selfemployed	더미형	고객의 직업(자영업)
job_services	더미형	고객의 직업(서비스직 종사자)
job_student	더미형	고객의 직업(학생)
job_technician	더미형	고객의 직업(기술자)
job_unemployed	더미형	고객의 직업(실직자)

최종변수 확인 (3/4)

변수명	데이터형태	상세설명
Marital_married	더미형	혼인여부(기혼)
Marital_single	더미형	혼인여부(미혼)
Marital_divorced	더미형	혼인여부(이혼)
Education_primary	더미형	고객의 학력수준(고졸이하)
Education_secondary	더미형	고객의 학력수준(대졸)
Education_tertiary	더미형	고객의 학력수준(대학원졸이상)
LastContactMonth_jan	더미형	마지막접촉 월(1월)
LastContactMonth_feb	더미형	마지막접촉 월(2월)
LastContactMonth_mar	더미형	마지막접촉 월(3월)
LastContactMonth_apr	더미형	마지막접촉 월(4월)

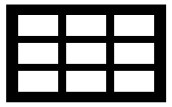
최종변수 확인 (4/4)

변수명	데이터형태	상세설명
LastContactMonth_may	더미형	마지막접촉 월(5월)
LastContactMonth_jun	더미형	마지막접촉 월(6월)
LastContactMonth_jul	더미형	마지막접촉 월(7월)
LastContactMonth_aug	더미형	마지막접촉 월(8월
LastContactMonth_sep	더미형	마지막접촉 월(9월)
LastContactMonth_oct	더미형	마지막접촉 월(10월)
LastContactMonth_nov	더미형	마지막접촉 월(11월)
LastContactMonth_dec	더미형	마지막접촉 월(12월)

총 데이터 수 : 4000행 * 40열

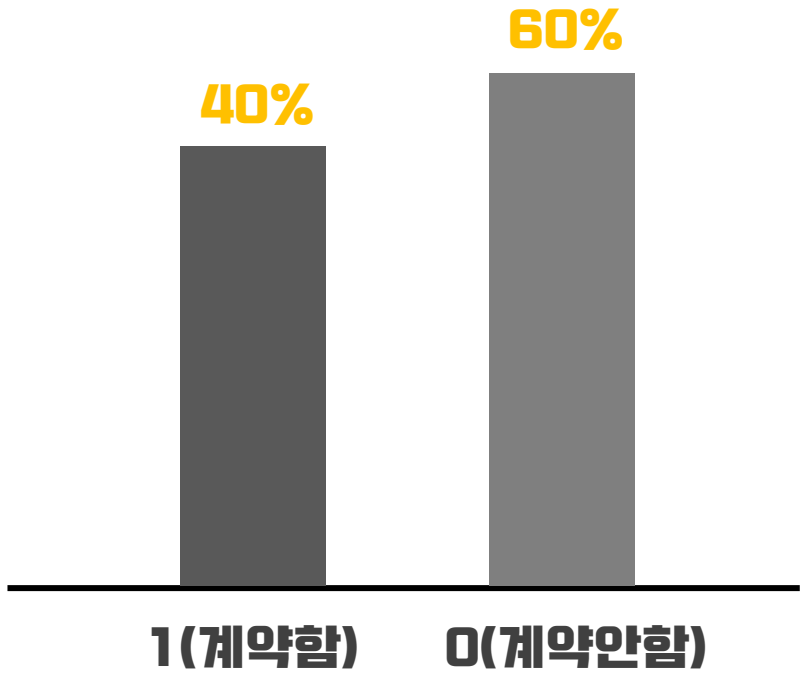
Scailing

목적 : 계약을 할 고객인지 분류 작업 → 각 Feature의 값이 일정 범위에 있어야 함 → 스케일링 필요

Data탐색  ▶ Age, Balance, Call_Duration, NoOfContacts, LastContactDay (이상 연속형 변수) ▶ **MinMaxscale** 

0과 1사이의 값으로 조정

Resampling/DataSplitting



불균형도가 크지 않아
굳이 resampling을 하지 않는 것이 낫다고 판단

Train/test set를 **8:2 비율로 split**
(교차검증을 할 것이기 때문에 validation set은 불필요)

차원 축소

모형의 복잡도를 낮춰 예측 모델의 **정확도와 학습 속도를 개선할 목적**

Feature extraction 데이터 내의 중복되고 상관없는 feature를 제거하고 종속변수에 유의한 feature를 선택

- PCA
- kernelPCA(linear)
- kernelPCA(rbf)

Feature selection 기존 feature들의 조합으로 서로 중복되지 않고 종속변수에 유의한 feature를 생성

- SelectKBest


챕터3
CHAPTER3
터
탐색
및
전처리
모델링
평가
및
결론

모델링

- Raw
- PCA
- kernelPCA(linear)
- kernelPCA(rbf)
- SelectKBest

차원축소

×

- Logistic regression
- Ridge
- Bagging
- RandomForest
- SVM(SVC)
- XGBoosting
- AdaBoosting
- Gaussian Naïve Bayes

분류모델

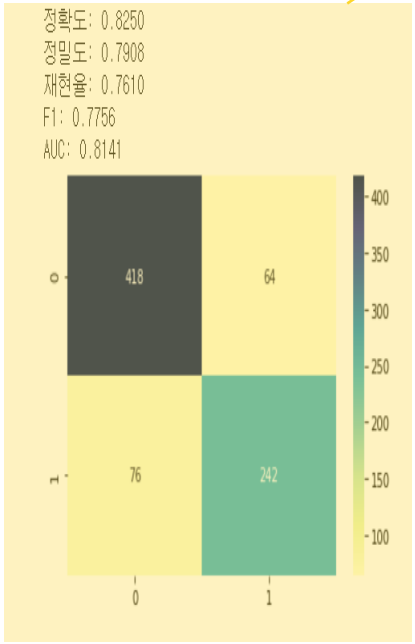
=

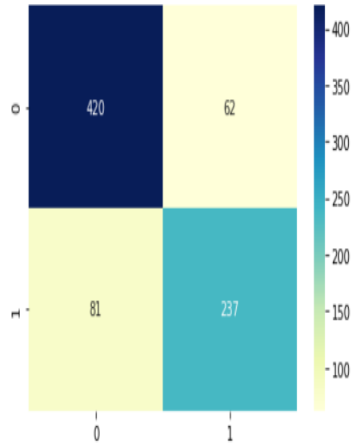
총 40가지 조합 비교

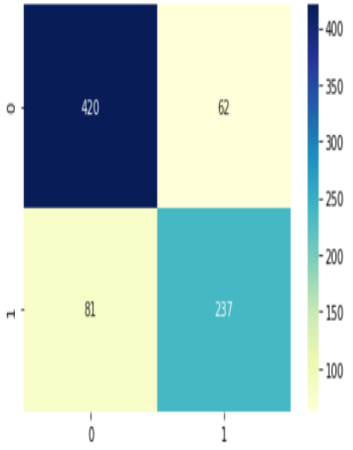
Logistic regression

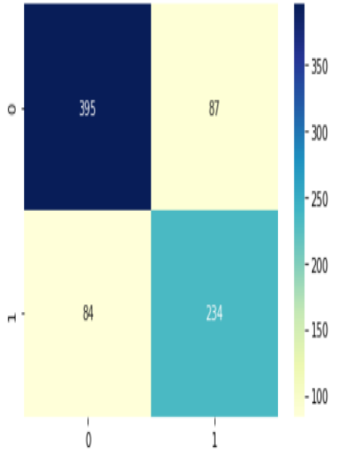
정확도(accuracy) : 0.8250 / 재현율(recall) : 0.7610

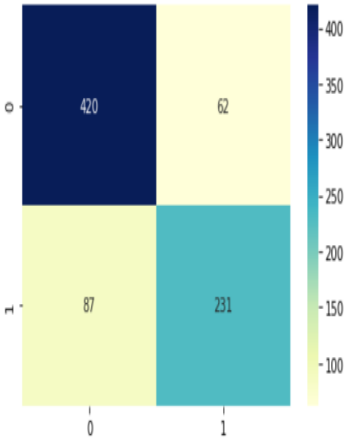
차원 축소	Raw (BEST)	PCA	kernelPCA(linear)	kernelPCA(rbf)	SelectKBest
결과	정확도: 0.8250 정밀도: 0.7908 재현율: 0.7610 F1: 0.7756 AUC: 0.8141	정확도: 0.8213 정밀도: 0.7926 재현율: 0.7453 F1: 0.7682 AUC: 0.8083	정확도: 0.8213 정밀도: 0.7926 재현율: 0.7453 F1: 0.7682 AUC: 0.8083	성확도: 0.7863 정밀도: 0.7290 재현율: 0.7358 F1: 0.7324 AUC: 0.7777	성확도: 0.8137 정밀도: 0.7884 재현율: 0.7264 F1: 0.7561 AUC: 0.7989







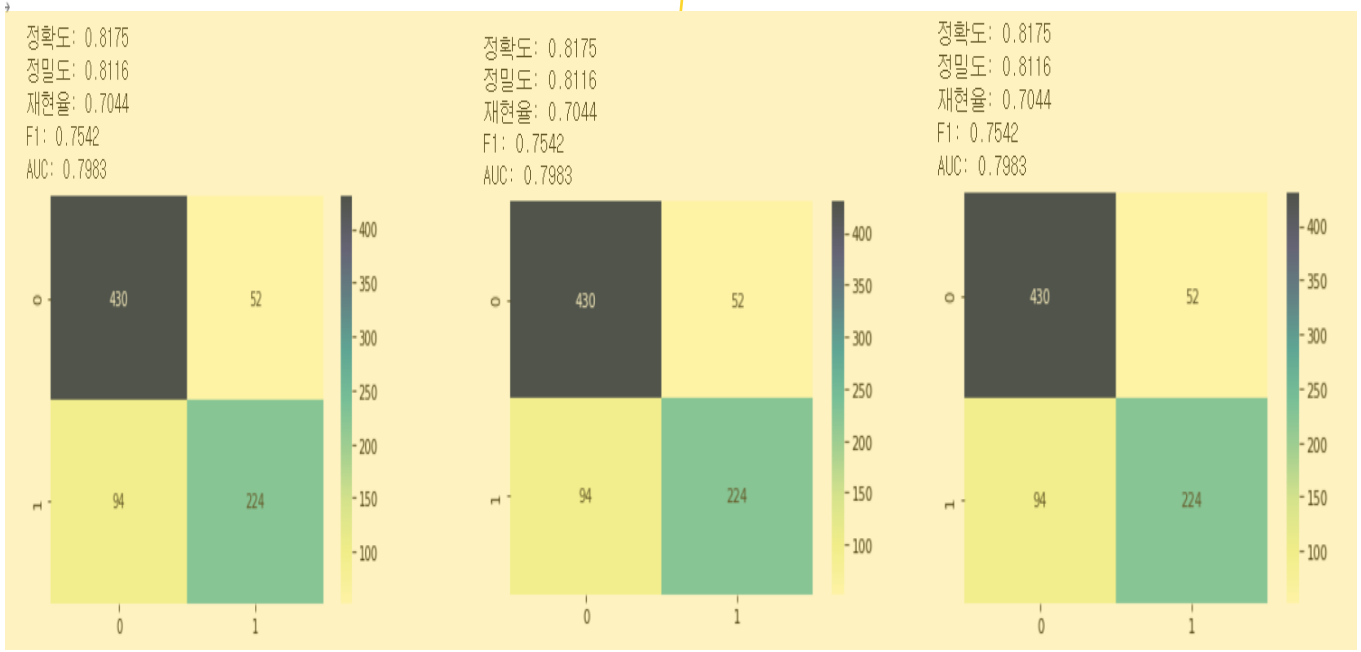


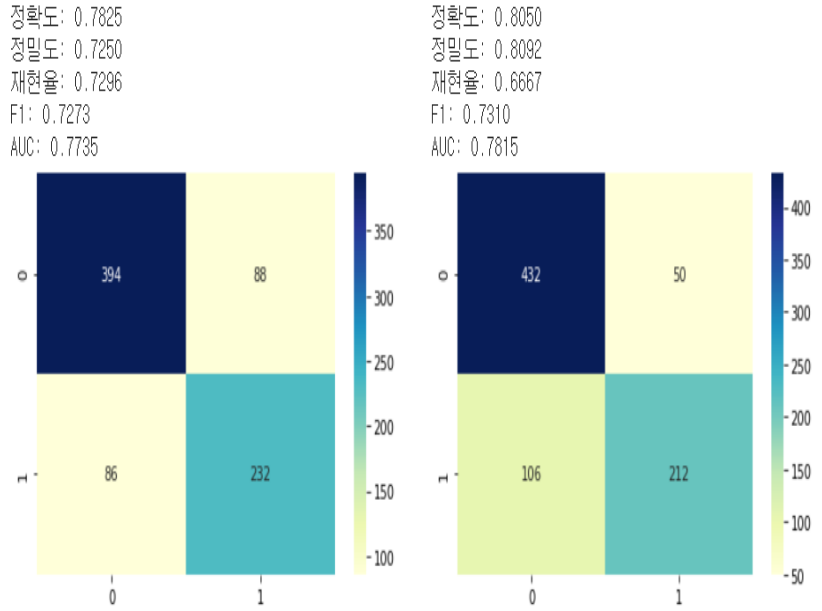


Ridge

정확도 (accuracy) : 0.8175 / 재현율(recall) : 0.7044

차원 축소	Raw (BEST)	PCA (BEST)	kernelPCA(linear) (BEST)	kernelPCA(rbf)	SelectKBest
결과	정확도: 0.8175 정밀도: 0.8116 재현율: 0.7044 F1: 0.7542 AUC: 0.7983	정확도: 0.8175 정밀도: 0.8116 재현율: 0.7044 F1: 0.7542 AUC: 0.7983	정확도: 0.8175 정밀도: 0.8116 재현율: 0.7044 F1: 0.7542 AUC: 0.7983	정확도: 0.7825 정밀도: 0.7250 재현율: 0.7296 F1: 0.7273 AUC: 0.7735	정확도: 0.8050 정밀도: 0.8092 재현율: 0.6667 F1: 0.7310 AUC: 0.7815

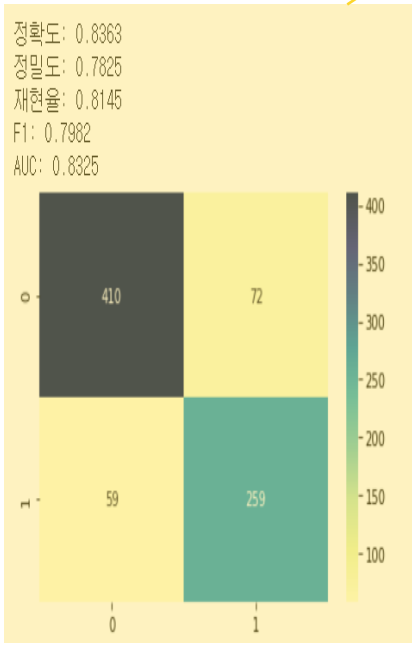


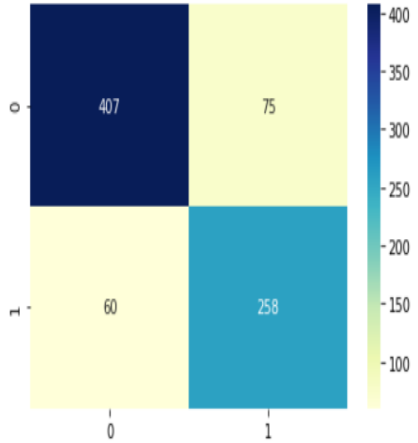


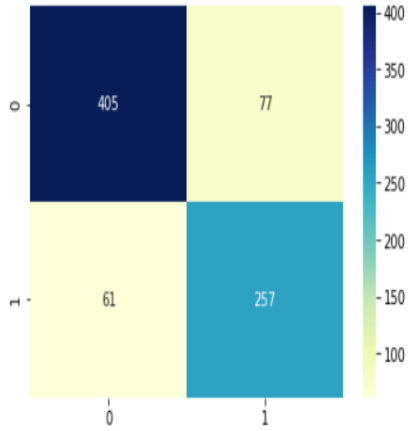
Bagging

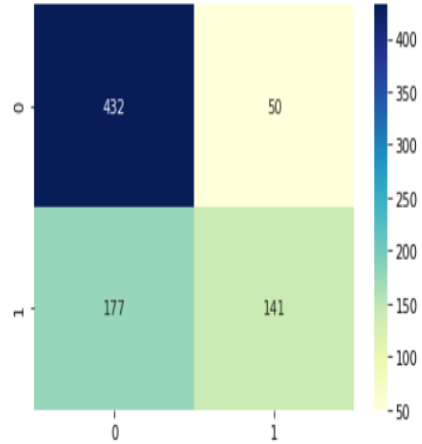
정확도(accuracy) : 0.8363 / 재현율(recall) : 0.7982

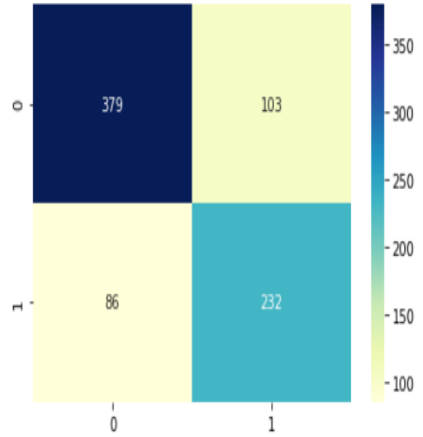
차원 축소	Raw (BEST)	PCA	kernelPCA(linear)	kernelPCA(rbf)	SelectKBest
결과	정확도: 0.8363 정밀도: 0.7825 재현율: 0.8145 F1: 0.7982 AUC: 0.8325	정확도: 0.8313 정밀도: 0.7748 재현율: 0.8113 F1: 0.7926 AUC: 0.8279	정확도: 0.8275 정밀도: 0.7695 재현율: 0.8082 F1: 0.7883 AUC: 0.8242	정확도: 0.7163 정밀도: 0.7382 재현율: 0.4434 F1: 0.5540 AUC: 0.6698	정확도: 0.7638 정밀도: 0.6925 재현율: 0.7296 F1: 0.7106 AUC: 0.7579







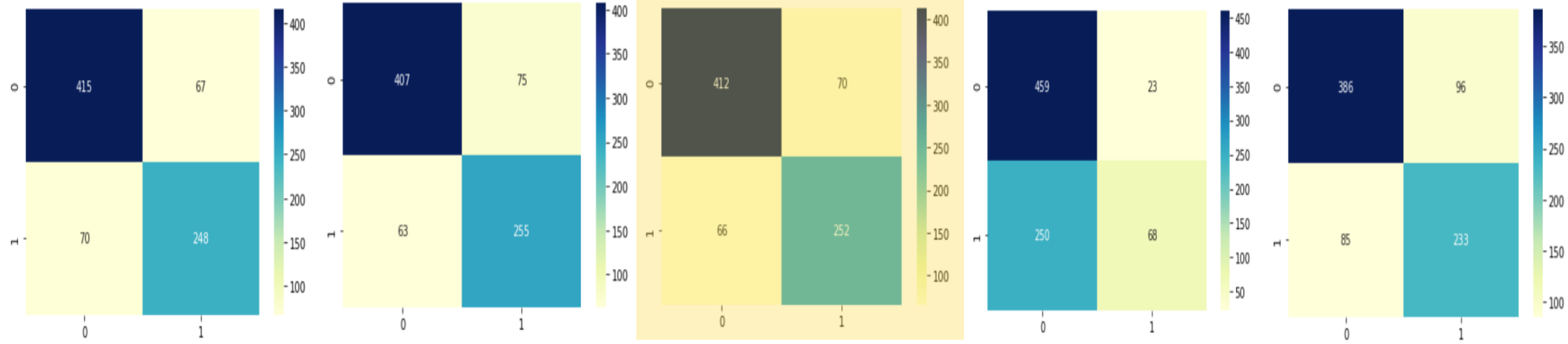




RandomForest

정확도(accuracy) : 0.8300 / 재현율(recall) : 0.7925

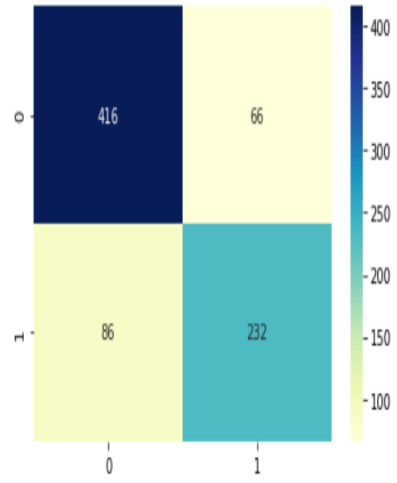
차원 축소	Raw	PCA	kernelPCA(linear)	kernelPCA(rbf)	SelectKBest
결과	정확도: 0.8287 정밀도: 0.7873 재현율: 0.7799 F1: 0.7836 AUC: 0.8204	정확도: 0.8275 정밀도: 0.7727 재현율: 0.8019 F1: 0.7870 AUC: 0.8231	정확도: 0.8300 정밀도: 0.7826 재현율: 0.7925 F1: 0.7875 AUC: 0.8236	정확도: 0.6587 정밀도: 0.7473 재현율: 0.2138 F1: 0.3325 AUC: 0.5831	정확도: 0.7738 정밀도: 0.7082 재현율: 0.7327 F1: 0.7202 AUC: 0.7668

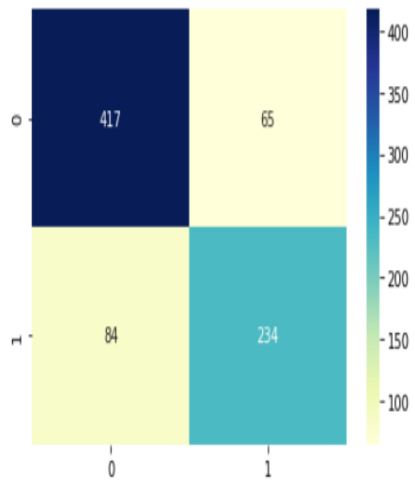


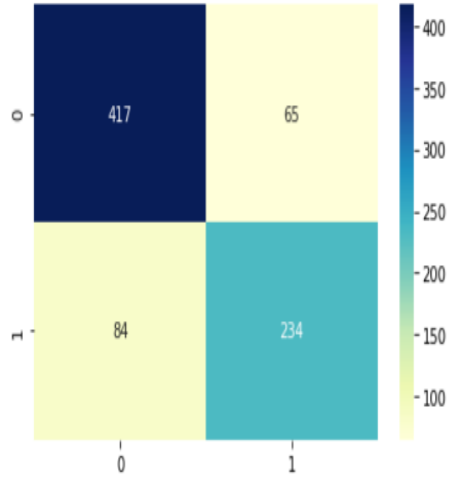
SVM(SVC)

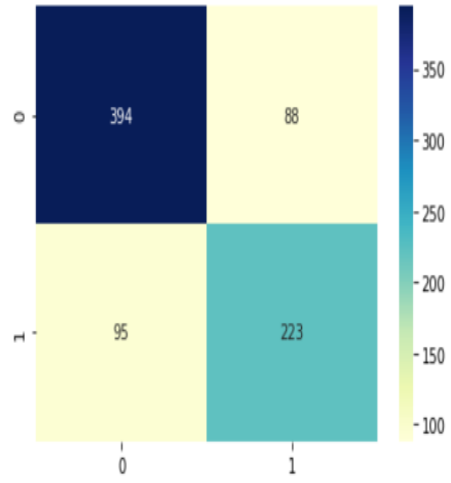
정확도(accuracy) : 0.8175 / 재현율(recall) : 0.7704

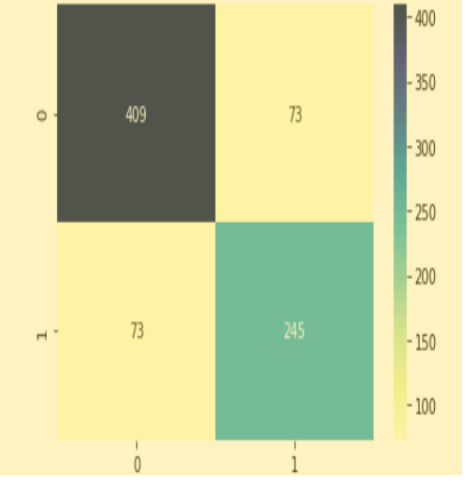
차원 축소	Raw	PCA	kernelPCA(linear)	kernelPCA(rbf)	SelectKBest (BEST)
결과	정확도: 0.8100 정밀도: 0.7785 재현율: 0.7296 F1: 0.7532 AUC: 0.7963	정확도: 0.8137 정밀도: 0.7826 재현율: 0.7358 F1: 0.7585 AUC: 0.8005	정확도: 0.8137 정밀도: 0.7826 재현율: 0.7358 F1: 0.7585 AUC: 0.8005	정확도: 0.7712 정밀도: 0.7170 재현율: 0.7013 F1: 0.7091 AUC: 0.7593	정확도: 0.8175 정밀도: 0.7704 재현율: 0.7704 F1: 0.7704 AUC: 0.8095







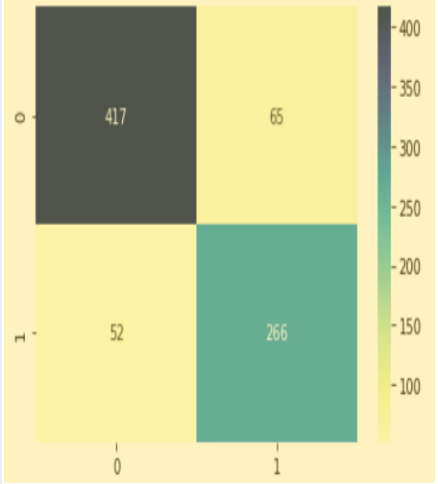


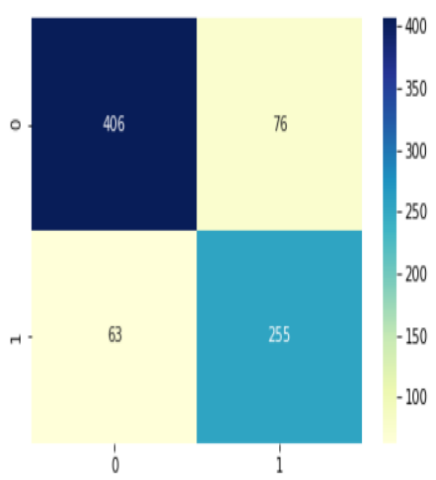


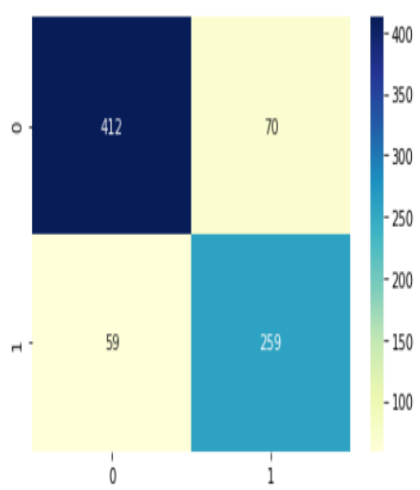
XGBoosting

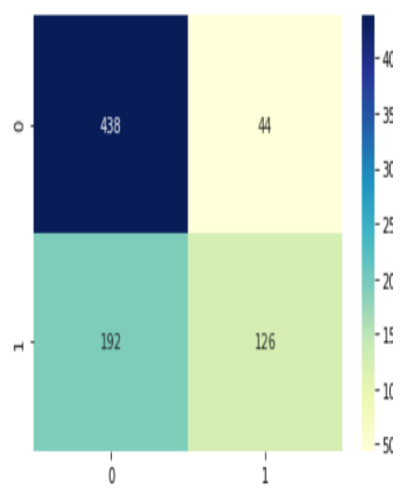
정확도(accuracy) : 0.8538 / 재현율(recall) : 0.8365

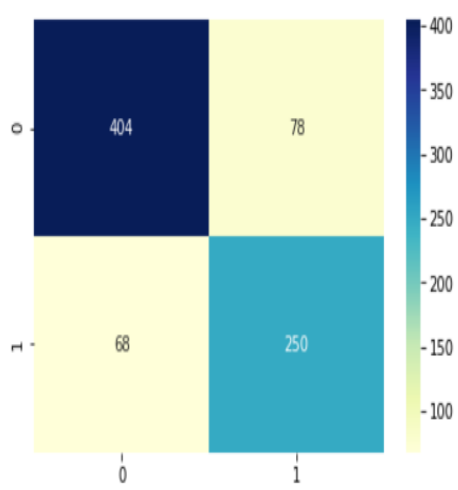
차원 축소	Raw (BEST)	PCA	kernelPCA(linear)	kernelPCA(rbf)	SelectKBest
결과	정확도: 0.8538 정밀도: 0.8036 재현율: 0.8365 F1: 0.8197 AUC: 0.8508	정확도: 0.8263 정밀도: 0.7704 재현율: 0.8019 F1: 0.7858 AUC: 0.8221	정확도: 0.8387 정밀도: 0.7872 재현율: 0.8145 F1: 0.8006 AUC: 0.8346	정확도: 0.7050 정밀도: 0.7412 재현율: 0.3962 F1: 0.5164 AUC: 0.6525	정확도: 0.8175 정밀도: 0.7622 재현율: 0.7862 F1: 0.7740 AUC: 0.8122







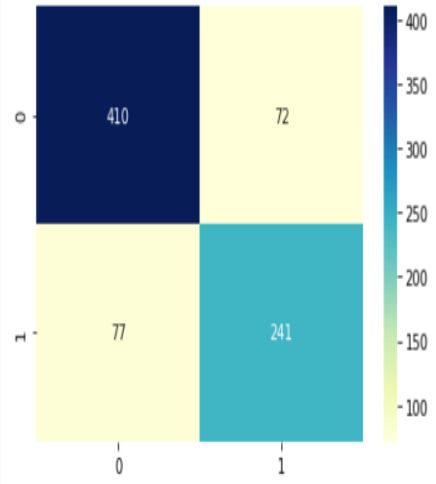


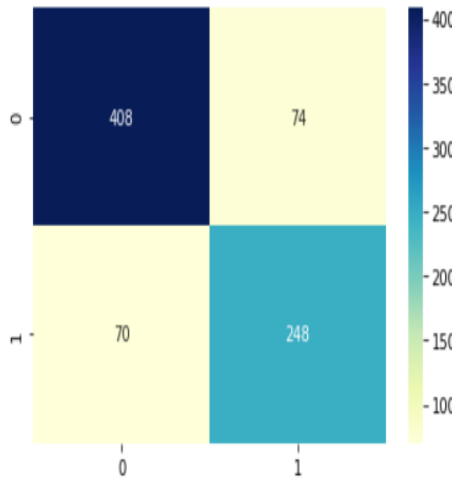


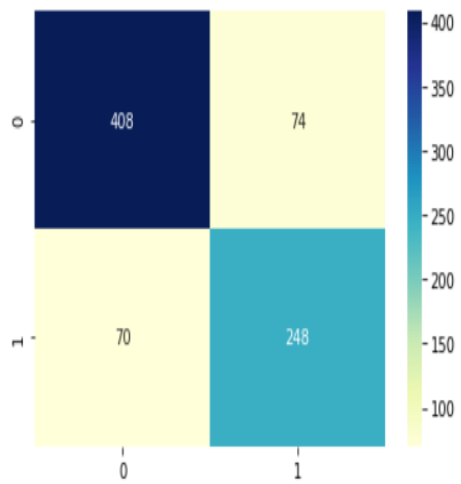
AdaBoosting

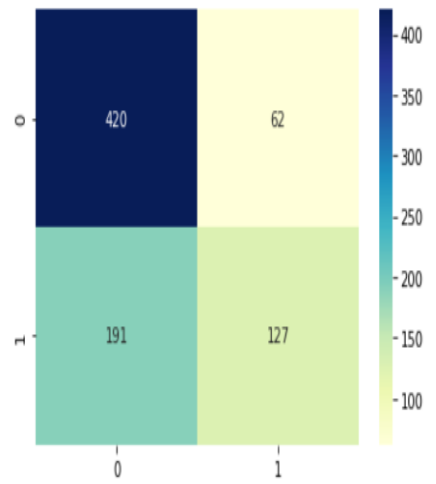
정확도(accuracy) : 0.8237 / 재현율(recall) : 0.7421

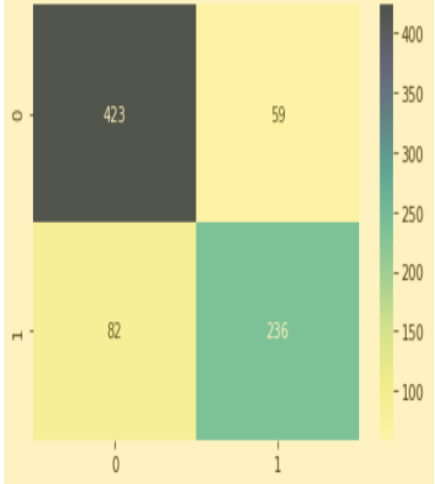
차원 축소	Raw	PCA	kernelPCA(linear)	kernelPCA(rbf)	SelectKBest
결과	정확도: 0.8137 정밀도: 0.7700 재현율: 0.7579 F1: 0.7639 AUC: 0.8042	정확도: 0.8200 정밀도: 0.7702 재현율: 0.7799 F1: 0.7750 AUC: 0.8132	정확도: 0.8200 정밀도: 0.7702 재현율: 0.7799 F1: 0.7750 AUC: 0.8132	정확도: 0.6837 정밀도: 0.6720 재현율: 0.3994 F1: 0.5010 AUC: 0.6354	정확도: 0.8237 정밀도: 0.8000 재현율: 0.7421 F1: 0.7700 AUC: 0.8099







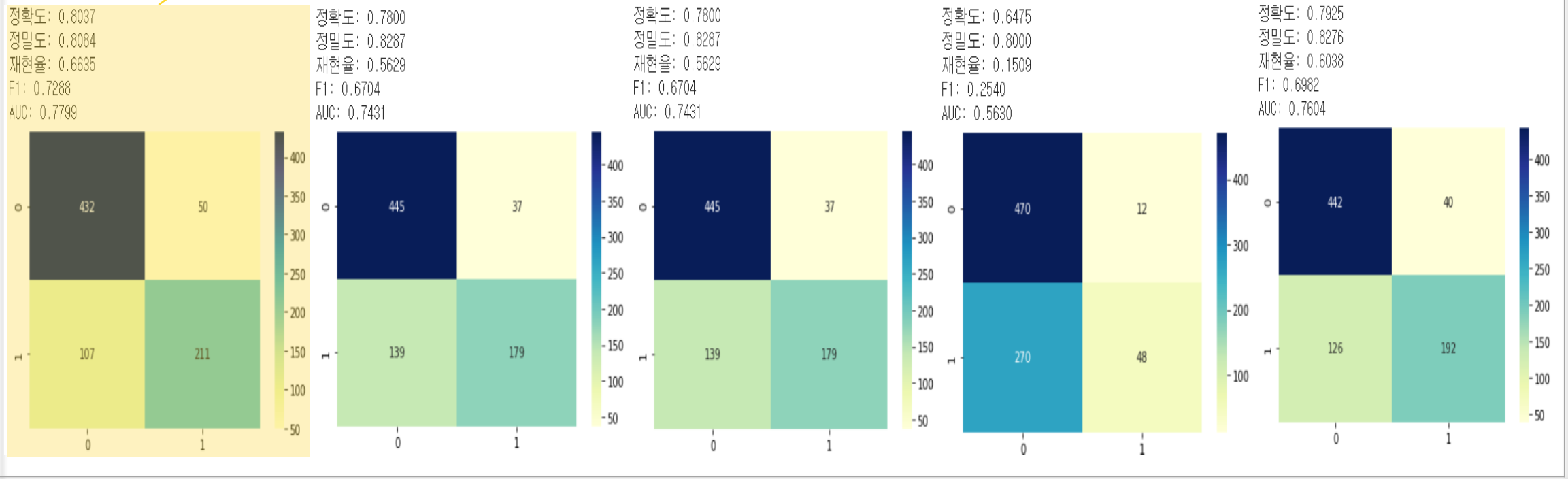




Gaussian Naïve Bayes

정확도(accuracy) : 0.8037 / 재현율(recall) : 0.6635

차원 축소	Raw (BEST)	PCA	kernelPCA(linear)	kernelPCA(rbf)	SelectKBest
결과	정확도: 0.8037 정밀도: 0.8084 재현율: 0.6635 F1: 0.7288 AUC: 0.7799	정확도: 0.7800 정밀도: 0.8287 재현율: 0.5629 F1: 0.6704 AUC: 0.7431	정확도: 0.7800 정밀도: 0.8287 재현율: 0.5629 F1: 0.6704 AUC: 0.7431	정확도: 0.6475 정밀도: 0.8000 재현율: 0.1509 F1: 0.2540 AUC: 0.5630	정확도: 0.7925 정밀도: 0.8276 재현율: 0.6038 F1: 0.6982 AUC: 0.7604



챕터4
CHAPTER4

주요지표비교

[Raw] – XGBoosting이 최우수

RAW	AdaBoosting	XGBoosting	Bagging	Random Forest	Logit	Ridge	SVM	Gaussian Naive Bayes
Accuracy	0.8137	0.8538	0.8363	0.8287	0.8250	0.8175	0.8100	0.8037
Recall	0.7579	0.8365	0.8145	0.7799	0.7610	0.7044	0.7296	0.6635
Precision	0.7700	0.8036	0.7825	0.7873	0.7908	0.8116	0.7785	0.8084
F1	0.7639	0.8197	0.7982	0.7836	0.7756	0.7542	0.7532	0.7288
AUC	0.8042	0.8508	0.8325	0.8204	0.8141	0.7983	0.7963	0.7799

[PCA] – Bagging이 최우수

PCA	AdaBoosting	XGBoosting	Bagging	Random Forest	Logit	Ridge	SVM	Gaussian Naive Bayes
Accuracy	0.8200	0.8263	0.8313	0.8275	0.8213	0.8175	0.8137	0.7800
Recall	0.7799	0.8019	0.8113	0.8019	0.7453	0.7044	0.7358	0.5629
Precision	0.7702	0.7704	0.7748	0.7727	0.7926	0.8116	0.7826	0.8287
F1	0.7750	0.7858	0.7926	0.7870	0.7682	0.7542	0.7585	0.6704
AUC	0.8132	0.8221	0.8279	0.8231	0.8083	0.7983	0.8005	0.7431

주요지표비교

[KernelPCA(Linear)] – XGBoosting이 최우수

KernelPCA(Linear)	AdaBoosting	XGBoosting	Bagging	Random Forest	Logit	Ridge	SVM	Gaussian Naive Bayes
Accuracy	0.8200	0.8387	0.8275	0.8300	0.8213	0.8175	0.8137	0.7800
Recall	0.7799	0.8145	0.8082	0.7925	0.7453	0.7044	0.7358	0.5629
Precision	0.7702	0.7872	0.7695	0.7826	0.7926	0.8116	0.7826	0.8287
F1	0.7750	0.8006	0.7883	0.7875	0.7682	0.7542	0.7585	0.6704
AUC	0.8132	0.8346	0.8242	0.8236	0.8083	0.7983	0.8005	0.7431

[KernelPCA(rbf)] – Logit이 최우수, 모형 불문 전반적 performance 저하

KernelPCA(rbf)	AdaBoosting	XGBoosting	Bagging	Random Forest	Logit	Ridge	SVM	Gaussian Naive Bayes
Accuracy	0.6837	0.7050	0.7163	0.6587	0.7863	0.7825	0.7712	0.6475
Recall	0.3994	0.3962	0.4434	0.2138	0.7358	0.7296	0.7013	0.1509
Precision	0.6720	0.7412	0.7382	0.7473	0.7290	0.7250	0.7170	0.8000
F1	0.5010	0.5164	0.5540	0.3325	0.7324	0.7273	0.7091	0.2540
AUC	0.6354	0.6525	0.6698	0.5831	0.7777	0.7735	0.7593	0.5630

주요지표비교

[selectKBest] – XGBoosting이 최우수

selectKBest	AdaBoosting	XGBoosting	Bagging	Random Forest	Logit	Ridge	SVM	Gaussian Naive Bayes
Accuracy	0.8237	0.8175	0.7638	0.7738	0.8137	0.8050	0.8175	0.7925
Recall	0.7421	0.7862	0.7296	0.7327	0.7264	0.6667	0.7704	0.6038
Precision	0.8000	0.7622	0.6925	0.7082	0.7884	0.8092	0.7704	0.8276
F1	0.7700	0.7740	0.7106	0.7202	0.7561	0.7310	0.7704	0.6982
AUC	0.8099	0.8122	0.7579	0.7668	0.7989	0.7815	0.8095	0.7604

Best model

Accuracy Best	Recall Best	Best combination
0.8538	0.8365	**Raw Data & XGBoosting**

추후 과제

Raw data와 차원축소 데이터의 XGBoosting의 퍼포먼스가 **아주 작은 차이**를 보임

▼

좀 더 **세밀한 하이퍼 파라미터 튜닝**으로 accuracy와 recall 상승 가능성 모색 필요

큐앤에이