Ybigta Machine Learning Project

IEEE-CIS Fraud Detection

Powered by kaggle

Team 4

고예희
문승현
박현지
염정운

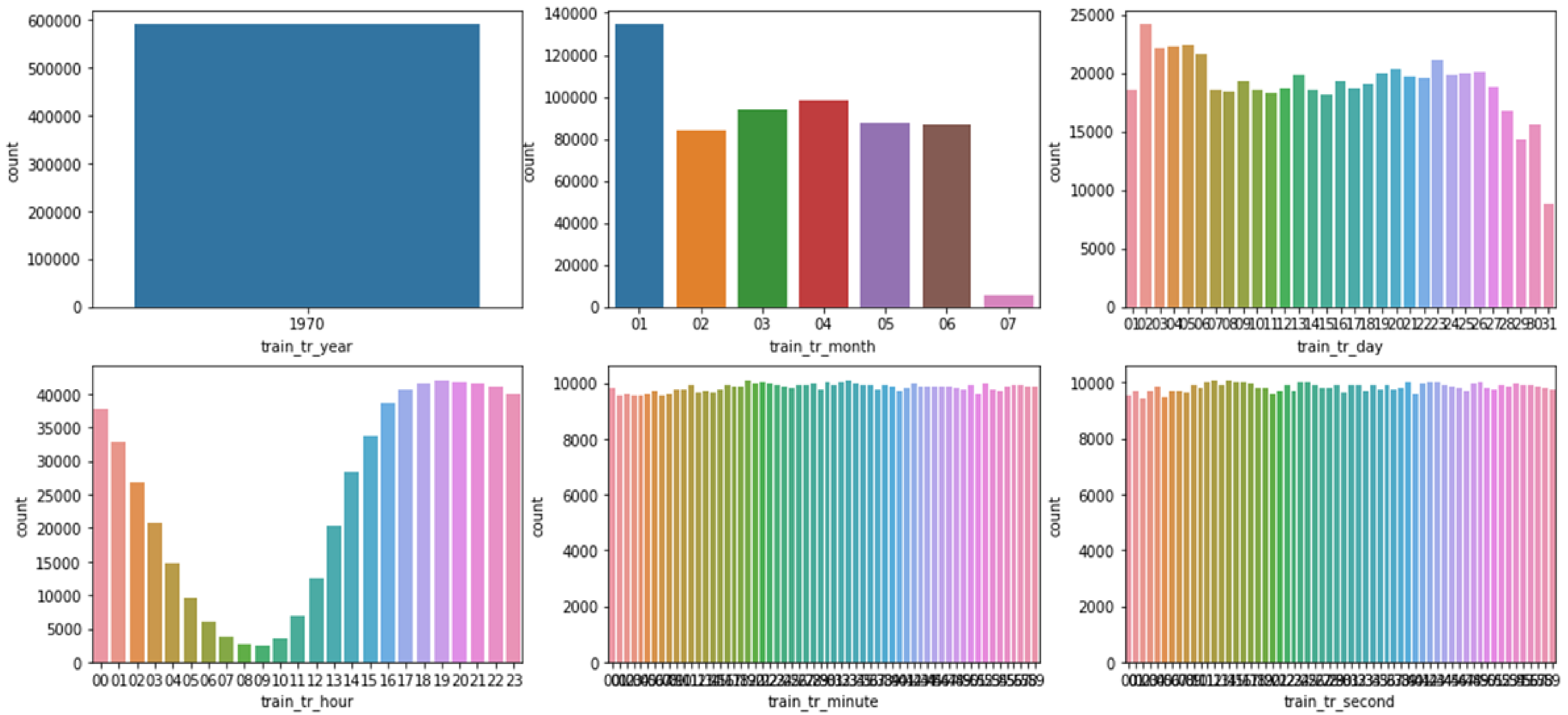

BIG DATA
VÉLOCITÉ
ANALYSES
TERAOCTETS
CLASSIQUES
MARCHÉ
ARCHITECTURE
ANALYTICS
ANALYSE
ANALYSÉES FRAMEWORK
COMPLEXES
VOLUME
MINING
ACTEURS
RELATIONNELLES
VISUALISATION
OUTILS
TECHNOLOGIES
SYSTÈMES
MODÈLES
RECHERCHE
PORTENT
VARIÉTÉ
ACTUELLE
WEB
RÉEL
PUIS
BASES
PARALLÈLES
BASÉES
AUJOURD HUI
BRUTES
DEVIENNENT
NOUVEAUX
JOUR
GÉRER
HEURE
EMC
ILS

CONTENTS

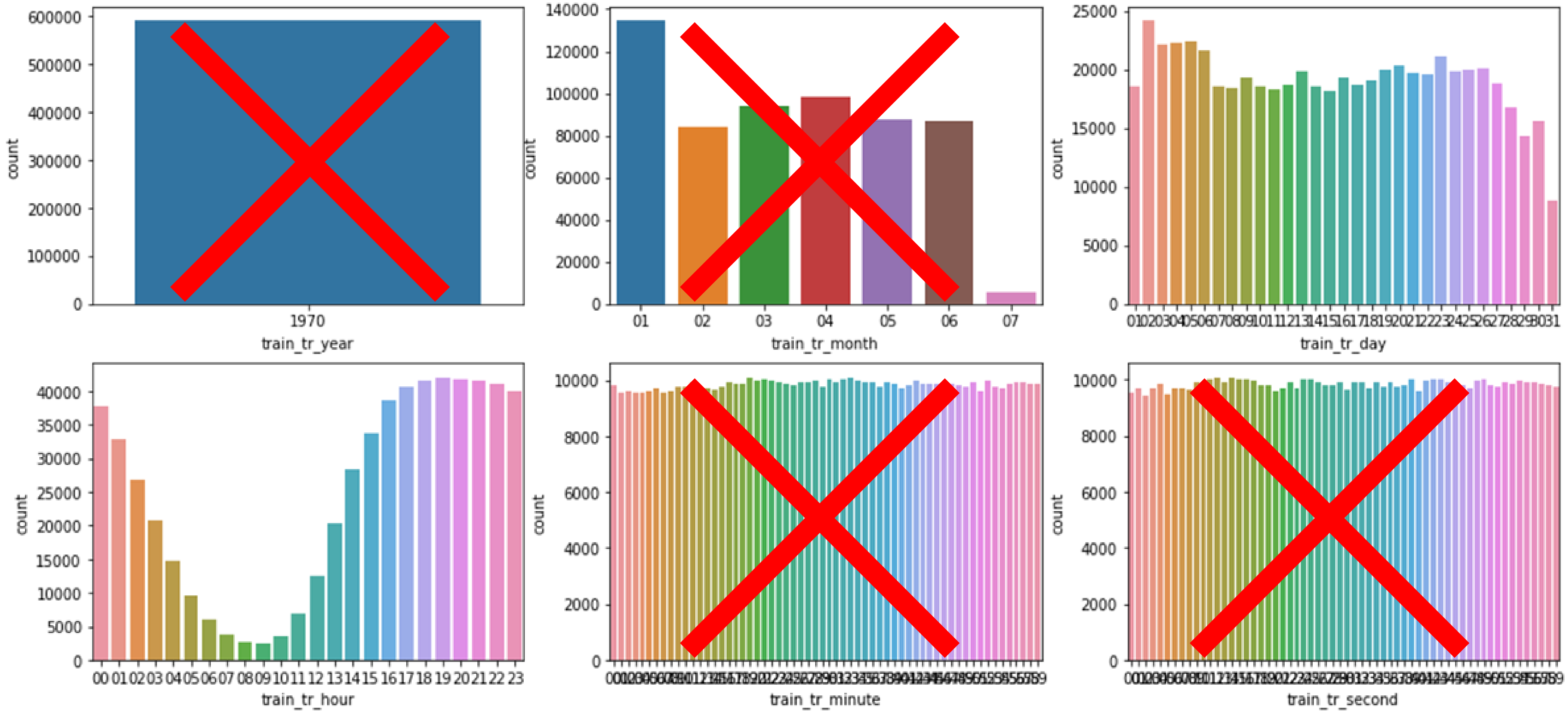
1. Variables

Variable selection, FE, Feedbacks etc…

TransactionDT



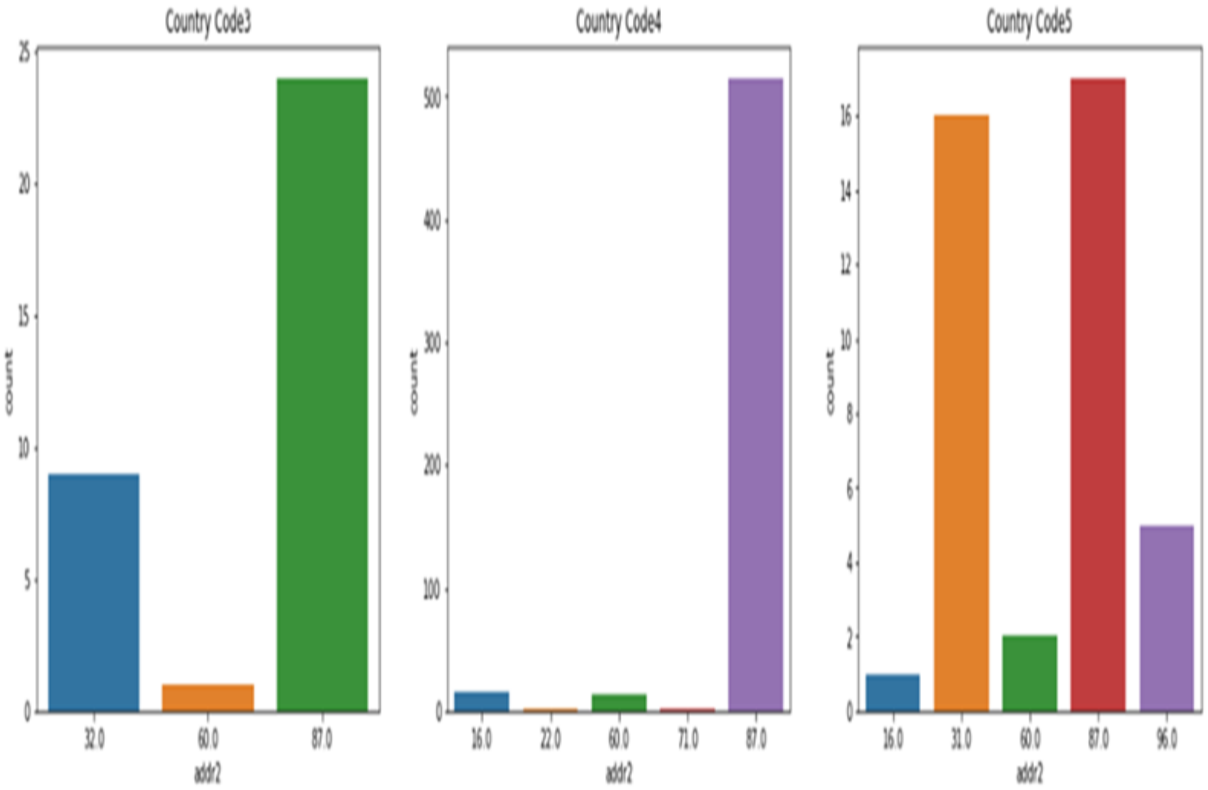
TransactionDT



Addr_2 & P_emaildomain

Addr 2 : 국가 코드라 추측되는 이상치 값

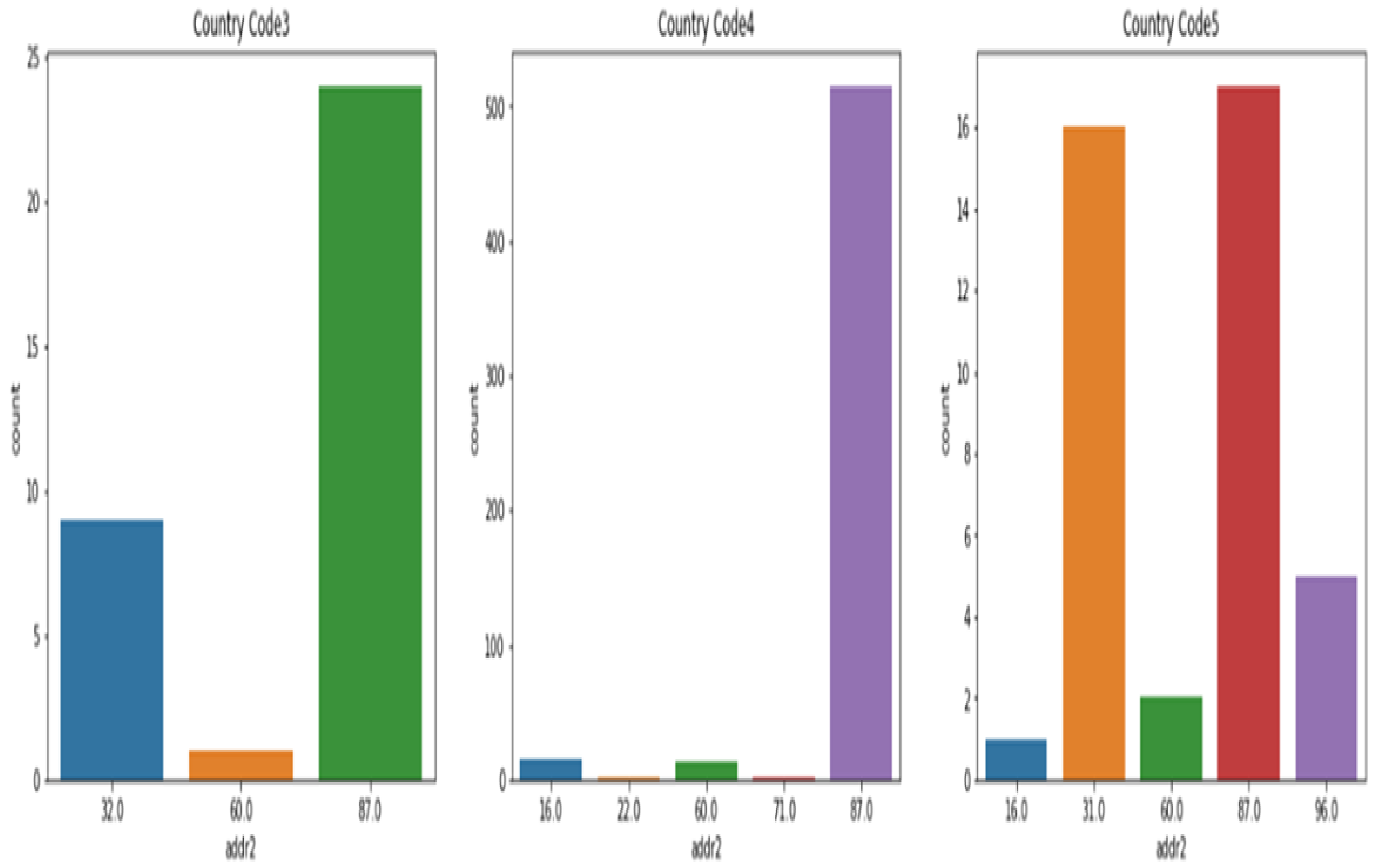
P emaildomain: 확실한 국가코드



Name	TRUE	FALSE	Overall	Percent	Rank	Service info		Region	Country_coc
aim.com	275	40	315	12.6984127	4	AOL		global	1
anonymous.com	36139	859	36998	2.32174712	20	익명		anonymous	9
aol.com	27672	617	28289	2.18105978	25	AOL		global	1
att.net	4003	30	4033	0.74386313	37	미국 통신사		usa	2
bellsouth.net	1856	53	1909	2.77632268	17	미국 통신사		usa	2
cableone.net	156	3	159	1.88679245	29	미국 통신사		usa	2
centurylink.net	205	0	205	0	43	미국 통신사		usa	2
cfl.rr.com	172	0	172	0	43	미국 통신사		usa	2
charter.net	791	25	816	3.06372549	15	미국 통신사		usa	2
comcast.net	7642	246	7888	3.11866126	14	미국 통신사		usa	2
cox.net	1364	29	1393	2.08183776	28	미국 통신사		usa	2
earthlink.net	503	11	514	2.14007782	26	미국 통신사		usa	2
embarqmail.com	251	9	260	3.46153846	11	미국 통신사		usa	2
frontier.com	272	8	280	2.85714286	16	미국 통신사		usa	2
frontiernet.net	190	5	195	2.56410256	19	미국 통신사		usa	2
gmail	485	11	496	2.21774194	23	구글		global	1
gmail.com	218412	9943	228355	4.35418537	9	구글		global	1
gmx.de	149	0	149	0	43	독일 메일회사		germany	4
hotmail.co.uk	112	0	112	0	43	마이크로소프트		uk	3
hotmail.com	42854	2396	45250	5.29502762	8	마이크로소프트		global	1
hotmail.de	43	0	43	0	43	마이크로소프트		germany	4
hotmail.es	285	20	305	6.55737705	6	마이크로소프트		spain	6
hotmail.fr	295	0	295	0	43	마이크로소프트		france	5
icloud.com	6070	197	6267	3.14344982	13	애플		global	1
juno.com	316	6	322	1.86335404	30	미국 통신사		usa	2
live.com	2957	84	3041	2.76224926	18	마이크로소프트		global	1
live.com.mx	708	41	749	5.47396529	7	마이크로소프트		mexico	7
live.fr	56	0	56	0	43	마이크로소프트		france	5
mac.com	422	14	436	3.21100917	12	애플		global	1
mail.com	453	106	559	18.9624329	2	독일 메일회사		germany	4
me.com	1495	27	1522	1.7739816	31	애플		global	1
msn.com	4002	90	4092	2.19941349	24	마이크로소프트		global	1
netzero.com	230	0	230	0	43	미국 통신사		usa	2
netzero.net	195	1	196	0.51020408	39	미국 통신사		usa	2
optonline.net	994	17	1011	1.68150346	32	미국 통신사		usa	2
outlook.com	4614	482	5096	9.45839874	5	마이크로소프트		global	1
outlook.es	381	57	438	13.0136986	3	스페인		spain	6
prodigy.net.mx	206	1	207	0.48309179	40	크라우드펀딩 회사		mexico	7

Addr_2 & P_emaildomain

Region from email	Code
Null	0
Global	1
USA	2
UK	**3**
Germany	**4**
France	**5**
Spain	6
Mexico	7
Japan	8
anonymous	9



DeviceInfo

Windows
iOS Device
MacOS
Trident/7.0
rv:11.0
rv:57.0
SM-J700M Build/MMB29K
SM-G610M Build/MMB29K
SM-G531H Build/LMY48B
rv:59.0
SM-G935F Build/NRD90M
SM-G955U Build/NRD90M
SM-G532M Build/MMB29T
ALE-L23 Build/HuaweiALE-L23
SM-G950U Build/NRD90M
SM-G930V Build/NRD90M
rv:58.0
rv:52.0
SAMSUNG
SM-G950F Build/NRD90M

LG
MacOS
Moto
SM
Trident/7.0
Windows
iOS Device
others
rv
unknown

One hot encoding

One-hot Encoding

ProductCD card4 card6

1. Country Code

2. Device Type

3. Device Info

1. Variables

train_transaction.csv

TransactionID TransactionDT

TransactionAmt ProductCD

card1 card2 card3 card4 card5 card6

Addr1 addr2 dist1 dist2

P_emaildomain R_emaildomain

C1-C14 D1-D15 M1-M19 Vxxx

이용변수

TransactionID isFraud
TransactionAmt card1
Card 2 card3 card5

변수변형

TransactionDT DeviceInfo

새로운변수

Addr2 Region

One-hot Encoding

ProductCD card4 card6 DeviceType
DeviceInfo

1. Resampling

-수많은 Resampling 종류

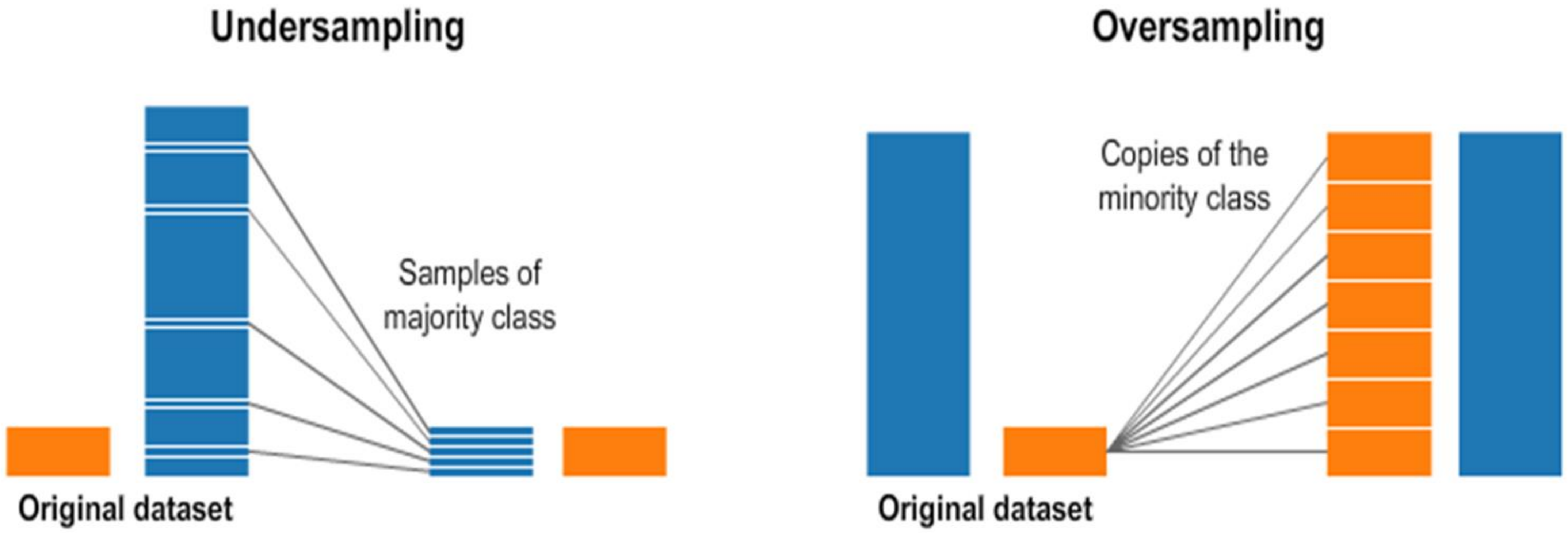
-패키지 사용의 문제점

2. CDMV…

-어떤 기법 적용

-NA 처리 기준

(1) Resampling



(1) Resampling

Imbalanced-learn 패키지

1. Under Sampling

- Random Under Sampler
- Tomek links method

2. Over Sampling

- Random Over Sampler
- Smote

(1) Resampling

Imbalanced-learn 패키지

NA값 처리 X

XGBoost

LightGBM

(1) Resampling

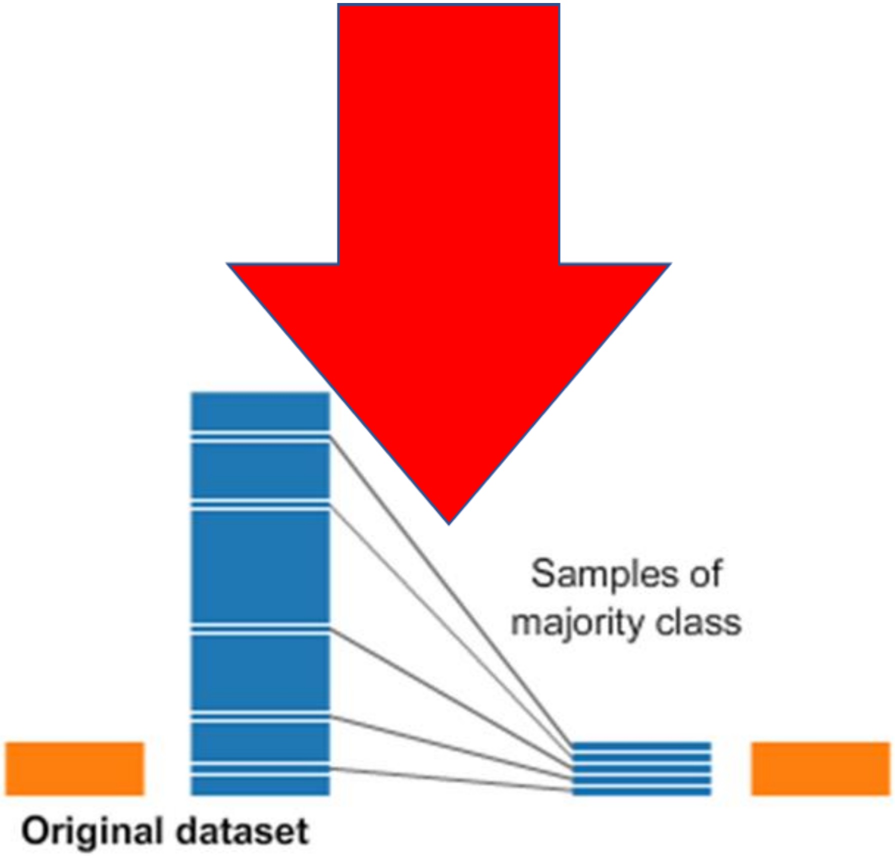
Imbalanced-learn 패키지

NA값 처리 X

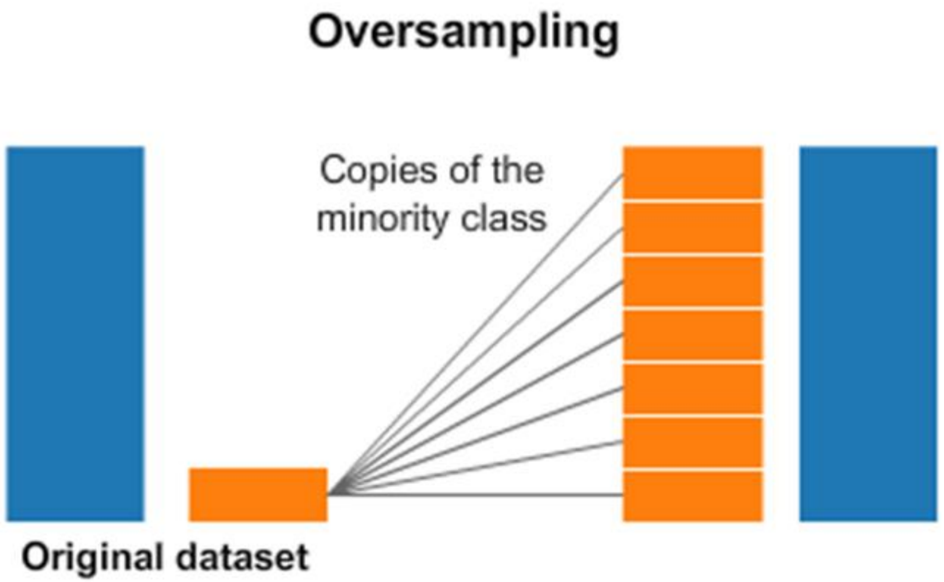
~~XGBoost~~

~~LightGBM~~

(1) Resampling



AUC Score



(2) PCA

수많은 V들…. (1 ~ 339) → PCA

(2) PCA

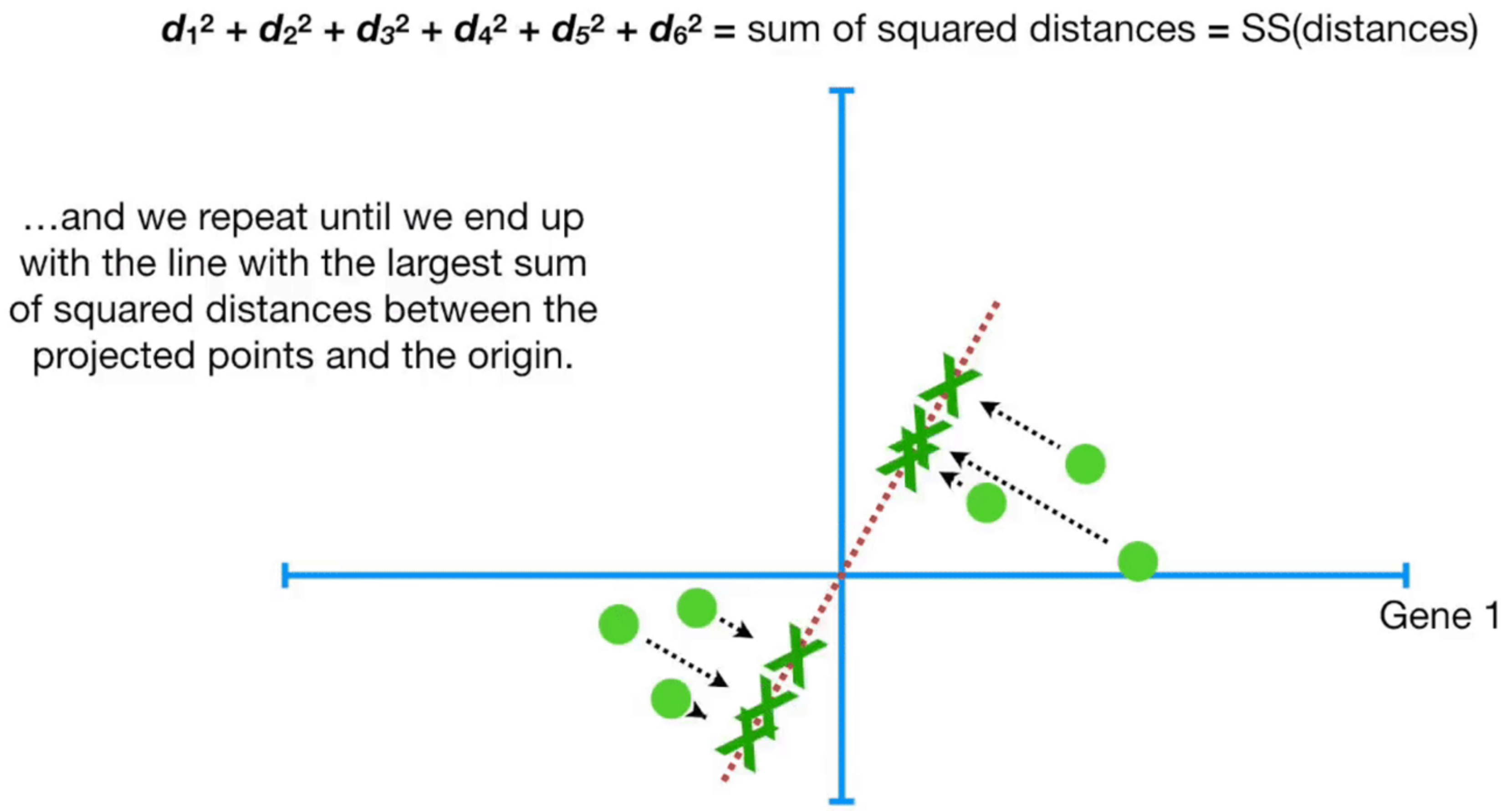
PCA

1. 변수들 간의 상관관계, 연관성

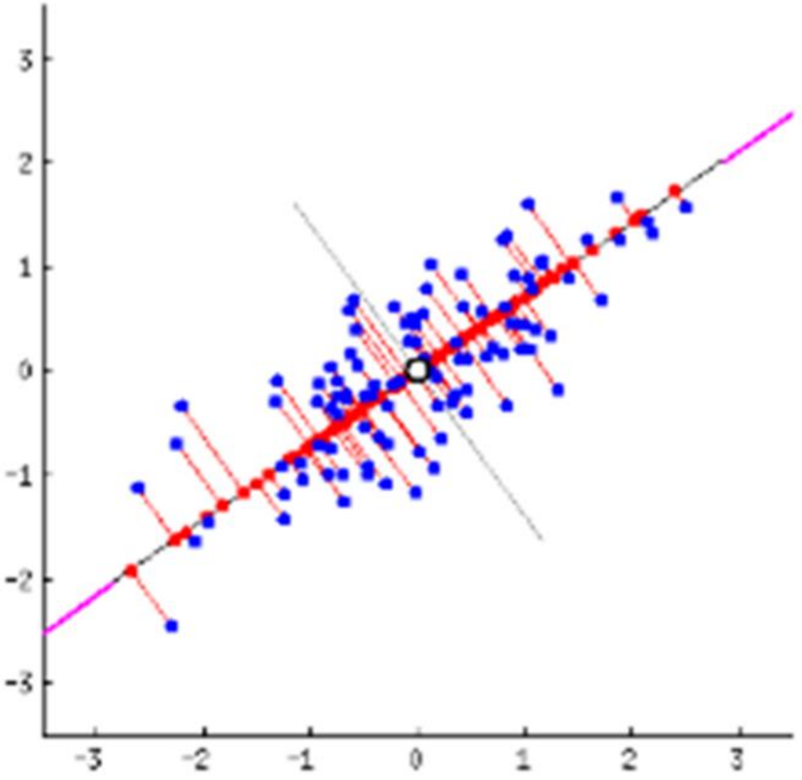
2. 선형결합 – 데이터의 변동성

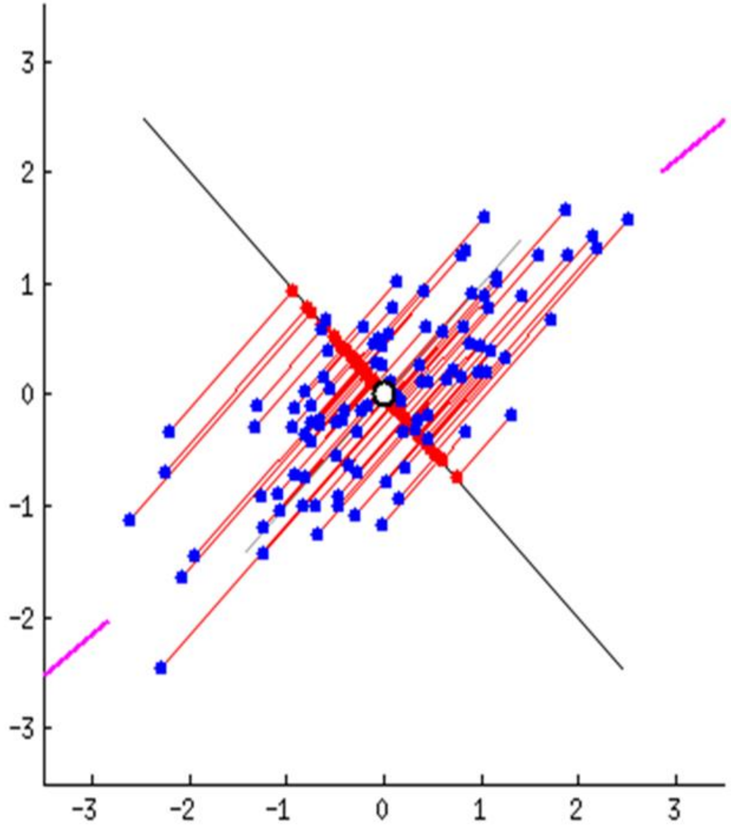
3. 고차원 데이터의 패턴 찾기

(2) PCA


$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$ = sum of squared distances = SS(distances)
...and we repeat until we end up with the line with the largest sum of squared distances between the projected points and the origin.
Gene 1

(2) PCA





(2) PCA

PCA ? → 예측성능 감소

(2) PCA

PCA

데이터의 차원축소 외

컴퓨터 비전 분야
-> 얼굴인식

2. Score

Confusion Matrix, Recall, Precision, F1, ROC-AUC

Fraud detection

Fraud(1) / not Fraud(0)

Transactic	isFraud	Transactic	Transactic	ProductCI	card1	card2	card3	card4	card5	card6	addr2	C1	D1	D10	D15	V12	V53	V75	V95
2987000	0	86400	68.5	0	13926	321	150	0	142	0	87	1	14	13	0	1	1	1	0
2987001	0	86401	29	0	2755	404	150	1	102	0	87	1	0	0	0	0	0	0	0
2987002	0	86469	59	0	4663	490	150	2	166	1	87	1	0	0	315	1	1	1	0
2987003	0	86499	50	0	18132	567	150	1	117	1	87	2	112	84	111	1	1	1	1
2987004	0	86506	50	1	4497	514	150	1	102	0	87	1	0	0	0	1	1	1	0
2987005	0	86510	49	0	5937	555	150	2	226	1	87	1	0	0	0	1	1	1	0
2987006	0	86522	159	0	12308	360	150	2	166	1	87	1	0	0	0	1	1	1	0
2987007	0	86529	422.5	0	12695	490	150	2	226	1	87	1	0	0	0	0	0	0	0
2987008	0	86535	15	1	2803	100	150	2	226	1	87	1	0	0	0	1	1	1	0
2987009	0	86536	117	0	17399	111	150	1	224	1	87	2	61	40	318	1	0	1	0
2987010	0	86549	75.887	2	16496	352	117	1	134	0	0	1	1	0	0	0	0	0	0
2987011	0	86555	16.495	2	4461	375	185	1	224	1	0	1	0	0	0	0	0	0	0
2987012	0	86564	50	0	3786	418	150	2	226	1	87	4	72	107	107	1	1	1	0
2987013	0	86585	40	0	12866	303	150	2	226	1	87	6	46	45	45	0	0	0	0
2987014	0	86596	10.5	0	11839	490	150	2	226	1	87	1	0	0	0	0	0	0	0
2987015	0	86618	57.95	0	7055	555	150	2	226	1	87	4	0	465	0	0	1	1	0
2987016	0	86620	30	1	1790	555	150	2	226	1	87	1	0	0	0	1	1	1	0
2987017	0	86668	100	1	11492	111	150	1	219	0	87	1	0	0	0	1	1	1	0
2987018	0	86725	47.95	0	4663	490	150	2	166	1	87	1	0	0	0	1	1	1	0
2987019	0	86730	186	0	7005	111	150	2	226	1	87	2	62	50	62	0	1	1	0
2987020	0	86761	39	0	7875	314	150	1	224	1	87	1	0	0	0	1	1	1	0
2987021	0	86769	159.95	0	11401	543	150	1	117	1	87	127	485	485	109	0	0	1	0
2987022	0	86786	50	1	1724	583	150	2	226	0	87	1	0	0	0	1	1	1	0
2987023	0	86808	107.95	0	2392	360	150	1	166	1	87	1	0	0	0	1	1	1	0
2987024	0	86821	73.95	0	10112	360	150	2	166	1	87	3	66	0	65	0	1	1	0
2987025	0	86944	107.95	0	15385	111	150	1	224	1	87	4	0	26	26	1	0	1	1
2987026	0	86945	184	0	17868	148	150	2	226	1	87	2	0	244	244	0	0	0	0
2987027	0	86972	47.95	0	11307	321	150	2	226	1	87	4	0	0	0	1	1	1	0
2987028	0	86973	20	0	8431	269	150	1	224	1	87	1	1	0	0	1	1	1	1
2987029	0	86979	36.99	0	12932	361	150	2	226	1	87	1	0	0	0	0	0	0	0
2987030	0	86994	35	0	13276	555	150	2	226	1	87	3	169	264	391	1	0	0	0
2987031	0	86998	363.89	0	6573	583	150	2	226	0	87	1	0	0	0	1	1	1	0
2987032	0	87008	200	0	7835	361	150	2	226	1	87	4	29	28	259	0	0	1	0
2987033	0	87078	40	0	8613	272	150	1	224	1	87	2	121	121	121	1	1	1	0
2987034	0	87135	107.95	0	17359	555	150	2	226	1	87	1	245	245	245	1	1	1	0
2987035	0	87140	107.95	0	9766	360	150	2	226	1	87	3	201	290	290	0	1	0	0
2987036	0	87149	77	0	4806	490	150	2	226	1	87	8	478	0	477	1	0	0	0
2987037	0	87161	21.95	0	13249	111	150	2	226	1	87	92	0	0	0	0	0	0	0
2987038	0	87172	25	3	5463	399	150	3	137	0	87	1	542	520	541	1	1	0	0
2987039	0	87202	39.95	0	12598	111	150	2	166	1	87	1	0	0	0	0	0	0	0
2987040	0	87209	75.887	2	13329	569	117	2	226	0	0	2	0	0	0	0	0	0	0
2987041	0	87220	774	0	9002	453	150	2	226	1	87	2	181	181	389	0	0	1	1
2987042	0	87233	117	0	13085	417	150	2	195	1	87	2	0	0	22	1	1	1	0

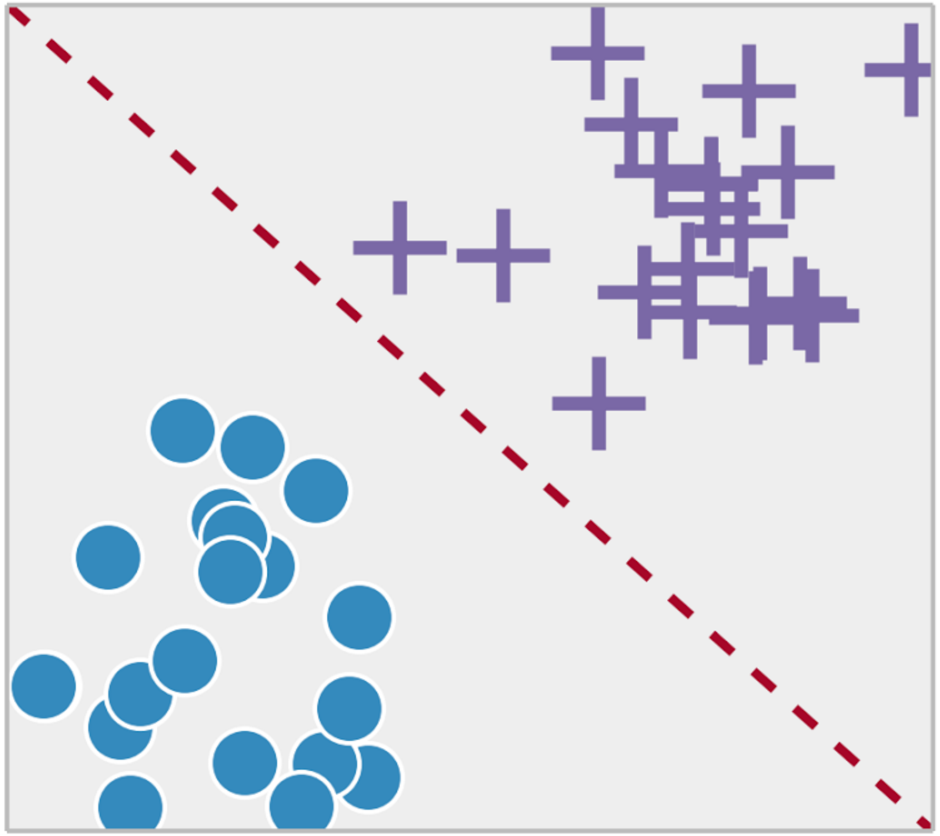
Transactic	V279	V281	exist	dt_month	dt_hour	ProductCI	ProductCI	ProductCI	ProductCI	ProductCI	card4_am	card4_dis	card4_ma	card4_visa	card6_cha	card6_cre	card6_det	card6_det	id_01	id_02	DeviceTyp	Device_inf	Country_c
2987000	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	1	0	0					0
2987001	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	1	0	0					1
2987002	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					1
2987003	1	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0					1
2987004	0	0	1	1	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	11.1674	2	1	1
2987005	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					1
2987006	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					1
2987007	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					4
2987008	0	0	1	1	0	0	1	0	0	0	0	0	0	1	0	0	1	0	1.79176	11.5023	2	2	9
2987009	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0					1
2987010	3	3	1	1	0	1	0	0	0	0	0	0	1	0	0	1	0	0	1.79176	12.1633	1	3	1
2987011	0	0	1	1	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1.79176	12.3097	1	0	1
2987012	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					2
2987013	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					1
2987014	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					1
2987015	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					0
2987016	0	0	1	1	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	8.91731	1	4	1
2987017	0	0	1	1	0	0	1	0	0	0	0	0	1	0	0	1	0	0	1.79176	11.0209	1	3	1
2987018	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					1
2987019	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					1
2987020	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0					1
2987021	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0					1
2987022	0	0	1	1	0	0	1	0	0	0	0	0	0	1	0	1	0	0	2.77259	11.7425	0	0	1
2987023	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0					1
2987024	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					1
2987025	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0					1
2987026	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					1
2987027	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					1
2987028	1	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0					0
2987029	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					1
2987030	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					1
2987031	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	1	0	0					1
2987032	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					1
2987033	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0					1
2987034	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					1
2987035	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					1
2987036	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					1
2987037	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					1
2987038	0	0	1	1	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	10.3724	2	0	0
2987039	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					1
2987040	0	0	1	1	0	1	0	0	0	0	0	0	0	1	0	1	0	0	2.3979	11.6622	1	3	1
2987041	1	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					1
2987042	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0					1

TransactionDT TransactionAmt ProductCD card1 card2 card3 card4 card5 card6 addr2 C1 D1 D10 D15 V12 V53 V75 V95 V279 V281 exist dt_month dt_hour ProductCD_C ProductCD_H ProductCD_R ProductCD_S ProductCD_W card4_american express card4_discover card4_mastercard card4_visa card6_charge card card6_credit card6_debit card6_debit or credit id_01 id_02 DeviceType Device_info_clean Country_code

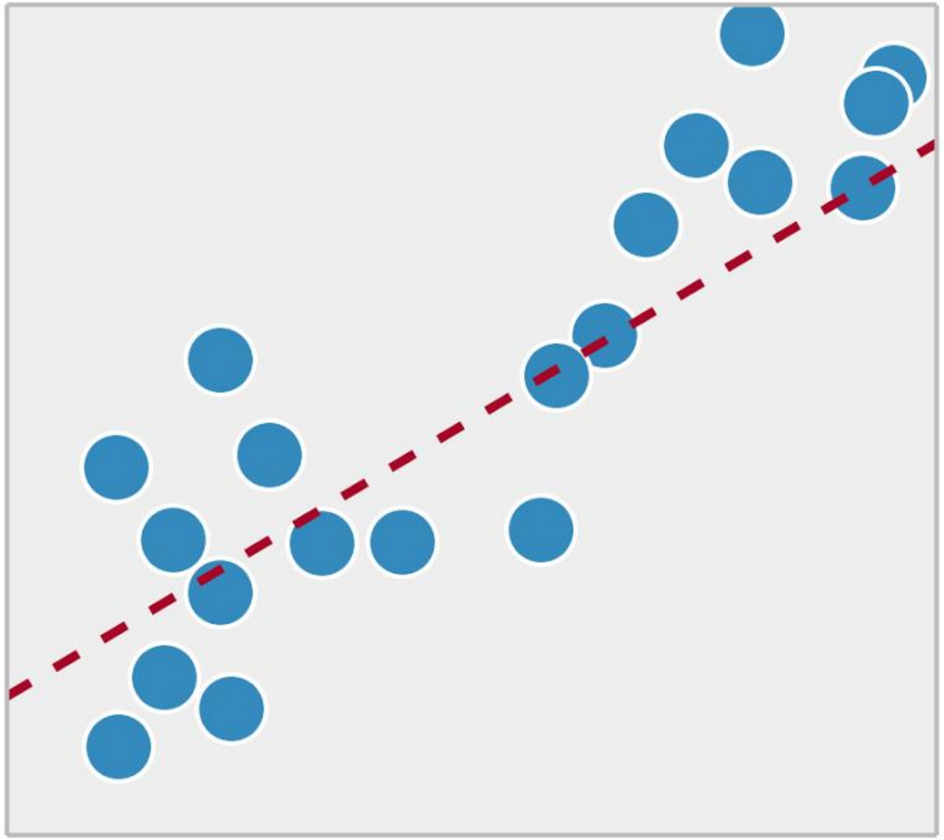
Fraud detection

Fraud(1) / not Fraud(0)

Classification

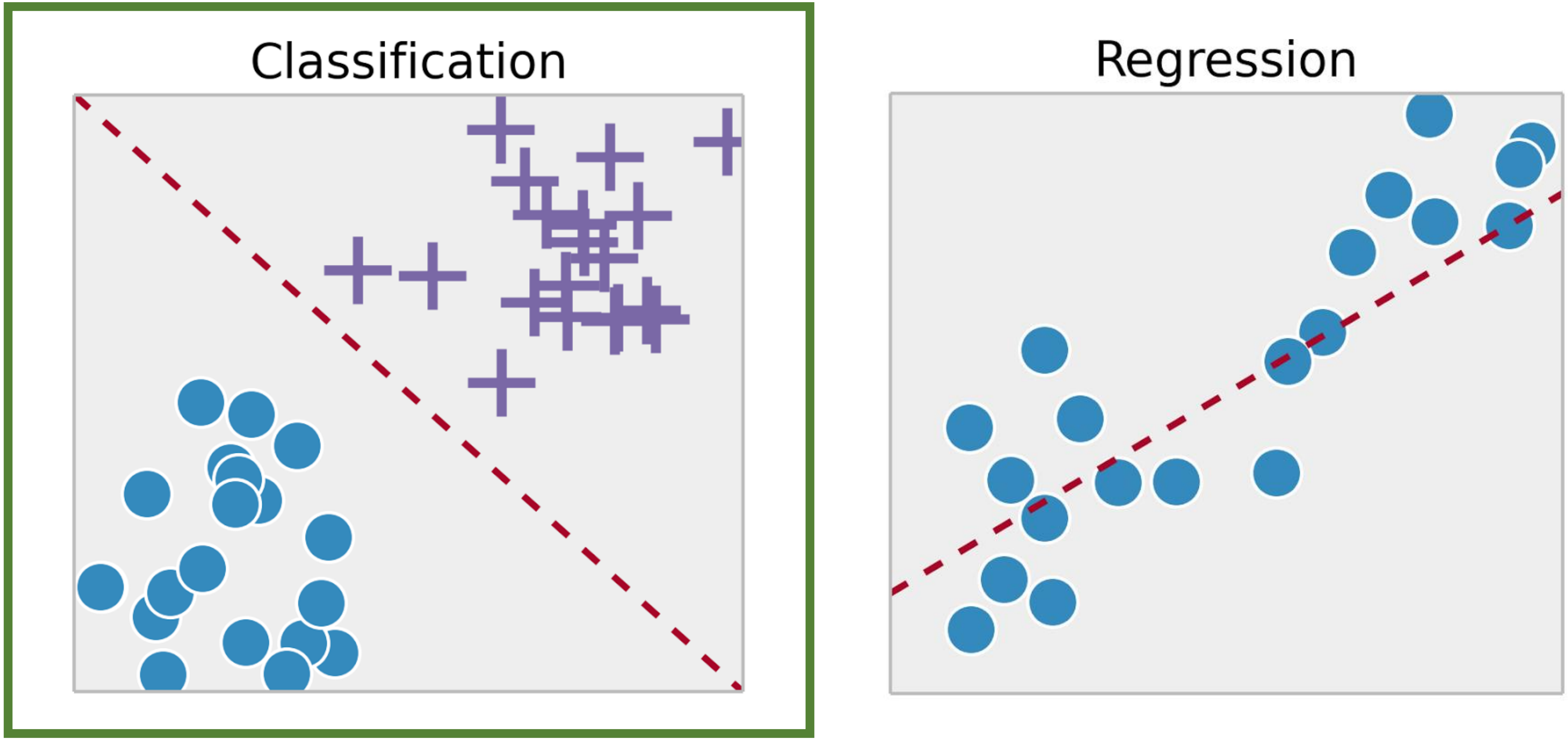


Regression



Fraud detection

Fraud(1) / not Fraud(0)



Fraud detection: Binary Classification

Fraud(1) / not Fraud(0)

실제값 / 예측값	Fraud(1)	Not Fraud(0)
Fraud(1)	True Positive (TP)	False Positive (FP)
Not Fraud(0)	False Negative (FN)	True Negative (TN)

Confusion Matrix (오차행렬)

Classification 결과(실제 값, 예측 값)를 행렬 형태로 나타냄.

Binary Classification의 경우 Class가 1, 0 두개이므로 2 x 2 형태로 표현

Fraud detection: Binary Classification

Fraud(1) / not Fraud(0)

실제값 / 예측값	Fraud(1)	Not Fraud(0)
Fraud(1)	True Positive (TP)	False Positive (FP)
Not Fraud(0)	False Negative (FN)	True Negative (TN)

Confusion Matrix (오차행렬)

True Positive

True Negative

False Positive

False Negative

올바른 예측

틀린 예측

Fraud detection: Binary Classification

Fraud(1) / not Fraud(0)

실제값 / 예측값	Fraud(1)	Not Fraud(0)
Fraud(1)	True Positive (TP)	False Positive (FP)
Not Fraud(0)	False Negative (FN)	True Negative (TN)

Confusion Matrix (오차행렬)

True Positive

True Negative

False Positive

False Negative

올바르게 예측하였고(True), 예측한 값이 1(Positive)이다

Fraud detection: Binary Classification

Fraud(1) / not Fraud(0)

실제값 / 예측값	Fraud(1)	Not Fraud(0)
Fraud(1)	True Positive (TP)	False Positive (FP)
Not Fraud(0)	False Negative (FN)	True Negative (TN)

Confusion Matrix (오차행렬)

True Positive

True Negative

False Positive

False Negative

올바르게 예측하였고(True), 예측한 값이 0(Negative)이다

Fraud detection: Binary Classification

Fraud(1) / not Fraud(0)

실제값 / 예측값	Fraud(1)	Not Fraud(0)
Fraud(1)	True Positive (TP)	False Positive (FP)
Not Fraud(0)	False Negative (FN)	True Negative (TN)

Confusion Matrix (오차행렬)

True Positive

True Negative

False Positive

False Negative

틀리게 예측하였고(False), 예측한 값이 1(Positive)이다

Fraud detection: Binary Classification

Fraud(1) / not Fraud(0)

실제값 / 예측값	Fraud(1)	Not Fraud(0)
Fraud(1)	True Positive (TP)	False Positive (FP)
Not Fraud(0)	False Negative (FN)	True Negative (TN)

Confusion Matrix (오차행렬)

True Positive

True Negative

False Positive

False Negative

틀리게 예측하였고(False), 예측한 값이 0(Negative)이다

How well Classified?

얼마나 잘 분류했는지 평가하기 위한 지표

Accuracy / F1 / ROC_AUC...

1) Accuracy

How well classified?

실제값 / 예측값	Fraud(1)	Not Fraud(0)
Fraud(1)	True Positive (TP)	False Positive (FP)
Not Fraud(0)	False Negative (FN)	True Negative (TN)

Ex) 100건의 거래 관측치

Total 100건

1) Accuracy: Balanced Data

How well classified?

실제값 / 예측값	Fraud(1)	Not Fraud(0)
Fraud(1)	True Positive (TP)	False Positive (FP)
Not Fraud(0)	False Negative (FN)	True Negative (TN)
	Fraud 50건	Not Fraud 50건

Ex) 100건의 거래 관측치

$$Accuracy = \frac{True\ Positive + True\ Negative}{OBS}$$

총 관측치(100건) 중
올바르게 예측한 비율

Total 100건

1) Accuracy: Balanced Data

How well classified?

실제값 \ 예측값	Fraud(1)	Not Fraud(0)
Fraud(1)	47	4
Not Fraud(0)	3	46
	Fraud 50건	Not Fraud 50건

Ex) 100건의 거래 관측치

$$Accuracy = \frac{True\ Positive + True\ Negative}{OBS}$$

총 관측치(100건) 중
올바르게 예측한 비율

Total 100건

1) Accuracy: Balanced Data

How well classified?

실제값 / 예측값	Fraud(1)	Not Fraud(0)
Fraud(1)	47	4
Not Fraud(0)	3	46
	Fraud 50건	Not Fraud 50건

Ex) 100건의 거래 관측치

$$Accuracy = \frac{True\ Positive + True\ Negative}{OBS}$$

총 관측치(100건) 중
올바르게 예측한 비율

$$\frac{47 + 46}{100} = 0.93$$

93% Accuracy

1) Accuracy: Imbalanced Data

How well classified?

실제값 / 예측값	Fraud(1)	Not Fraud(0)
Fraud(1)	True Positive (TP)	False Positive (FP)
Not Fraud(0)	False Negative (FN)	True Negative (TN)
	Fraud **3건**	Not Fraud **97건**

Total 100건

But… if?

1) Accuracy: Imbalanced Data

How well classified?

실제값 / 예측값	Fraud(1)	Not Fraud(0)	
Fraud(1)	True Positive (TP)	False Positive (FP)	Fraud 0건
Not Fraud(0)	False Negative (FN)	True Negative (TN)	Not Fraud 100건
	Fraud 3건	Not Fraud 97건	Total 100건

1) Accuracy: Imbalanced Data

How well classified?

예측값 \ 실제값	Fraud(1)	Not Fraud(0)	
Fraud(1)	0	0	Fraud 0건
Not Fraud(0)	3	97	Not Fraud 100건
	Fraud 3건	Not Fraud 97건	Total 100건

$$Accuracy = \frac{True\ Positive + True\ Negative}{OBS}$$

$$\frac{97+0}{100} = 0.97$$

97% Accuracy

사기 거래는
하나도 못 잡았는데...?

상상도 못한 정체

2) F1, ROC_AUC - Recall, Precision, Specificity

(sensitivity)

How well classified?

실제값 / 예측값	Fraud(1)	Not Fraud(0)
Fraud(1)	True Positive (TP)	False Positive (FP)
Not Fraud(0)	False Negative (FN)	True Negative (TN)

$$\text{Recall} = \frac{True\ Positive}{\text{True Positive} + \text{False Negative}}$$

(sensitivity)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + \text{False Positive}}$$

$$\text{Specificity} = \frac{True\ Negative}{True\ \text{Negative} + False\ \text{Positive}}$$

2) F1, ROC_AUC - Recall, Precision, Specificity
(sensitivity)

How well classified?

예측값 \ 실제값	Fraud(1)	Not Fraud(0)
Fraud(1)	True Positive (TP)	False Positive (FP)
Not Fraud(0)	False Negative (FN)	True Negative (TN)

$$\text{Recall} = \frac{True\ Positive}{\text{True Positive} + \text{False Negative}}$$

(sensitivity)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + \text{False Positive}}$$

$$\text{Specificity} = \frac{True\ Negative}{True\ \text{Negative} + False\ \text{Positive}}$$

실제 Fraud 를 올바르게 가려낸 비율

2) F1, ROC_AUC - Recall, Precision, Specificity

(sensitivity)

How well classified?

실제값 / 예측값	Fraud(1)	Not Fraud(0)
Fraud(1)	True Positive (TP)	False Positive (FP)
Not Fraud(0)	False Negative (FN)	True Negative (TN)

$$\text{Recall} = \frac{True\ Positive}{\text{True Positive} + \text{False Negative}}$$

(sensitivity)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + \text{False Positive}}$$

$$\text{Specificity} = \frac{True\ Negative}{True\ \text{Negative} + False\ \text{Positive}}$$

실제 Not Fraud 를 올바르게 가려낸 비율

2) F1, ROC_AUC - Recall, Precision, Specificity
(sensitivity)

How well classified?

실제값 / 예측값	Fraud(1)	Not Fraud(0)
Fraud(1)	True Positive (TP)	False Positive (FP)
Not Fraud(0)	False Negative (FN)	True Negative (TN)

$$\text{Recall (sensitivity)} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}$$

예측 Fraud 중 실제 Fraud 를 올바르게 예측한 비율

2) F1, ROC_AUC - Recall, Precision, Specificity
(sensitivity)

How well classified?

실제값 / 예측값	Fraud(1)	Not Fraud(0)
Fraud(1)	**0**	0
Not Fraud(0)	3	97

Fraud
3건

Not Fraud
97건

$$\text{Recall} = \frac{TP}{TP+FN} = 0$$

$$\text{Precision} = \frac{TP}{TP+FP} = 0$$

$$\text{Specificity} = \frac{TN}{TN+FP} = 0$$

3) F1

How well classified?

실제값 / 예측값	Fraud(1)	Not Fraud(0)
Fraud(1)	True Positive (TP)	False Positive (FP)
Not Fraud(0)	False Negative (FN)	True Negative (TN)

$$\text{Recall} = \frac{\mathit{True\ Positive}}{\text{True Positive} + \mathit{False\ Negative}}$$

$$\text{Precision} = \frac{\mathit{True\ Positive}}{\mathit{True\ Positive} + \boldsymbol{\mathit{False\ Positive}}}$$

Recall과 Precision은 **Trade-off** 이고,

분류 성능 평가에서 이 두 지표를 적절히

Balancing 해야 한다!

3) F1

How well classified?

실제값 \ 예측값	Fraud(1)	Not Fraud(0)
Fraud(1)	True Positive (TP)	False Positive (FP)
Not Fraud(0)	False Negative (FN)	True Negative (TN)

$$\text{Recall} = \frac{\mathit{True\ Positive}}{\text{True Positive} + \mathit{False\ Negative}}$$

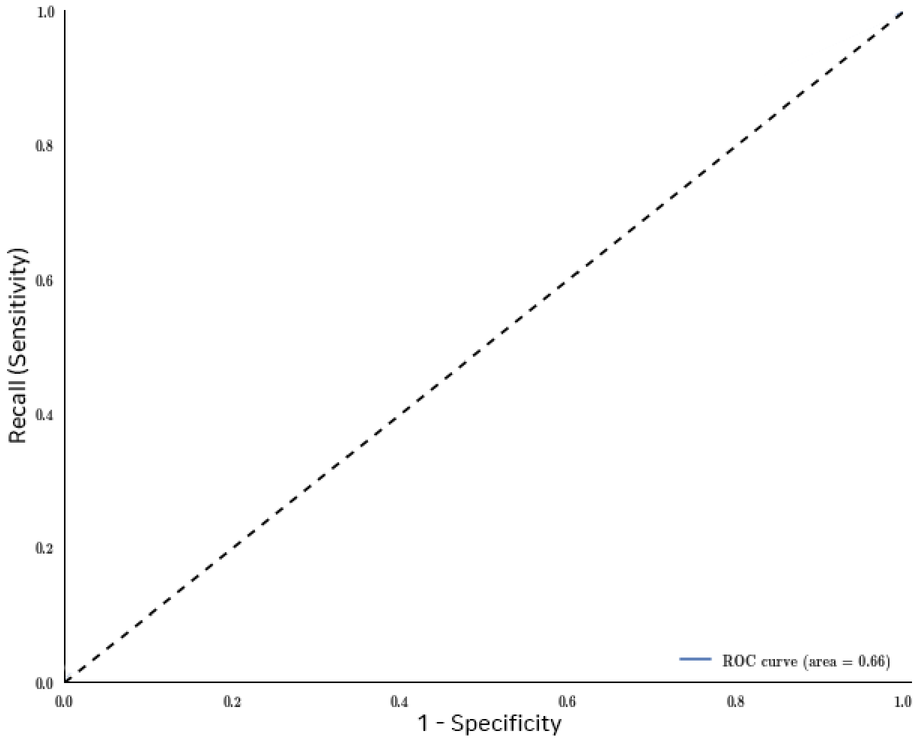
$$\text{Precision} = \frac{\mathit{True\ Positive}}{\mathit{True\ Positive} + \boldsymbol{False\ Positive}}$$

$$F_1\ score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$

[0,1] 값을 가지며, Recall과 Precision 모두를 고려한 지표
1에 가까울 수록 잘 분류되었다고 평가

4) ROC_AUC

How well classified?

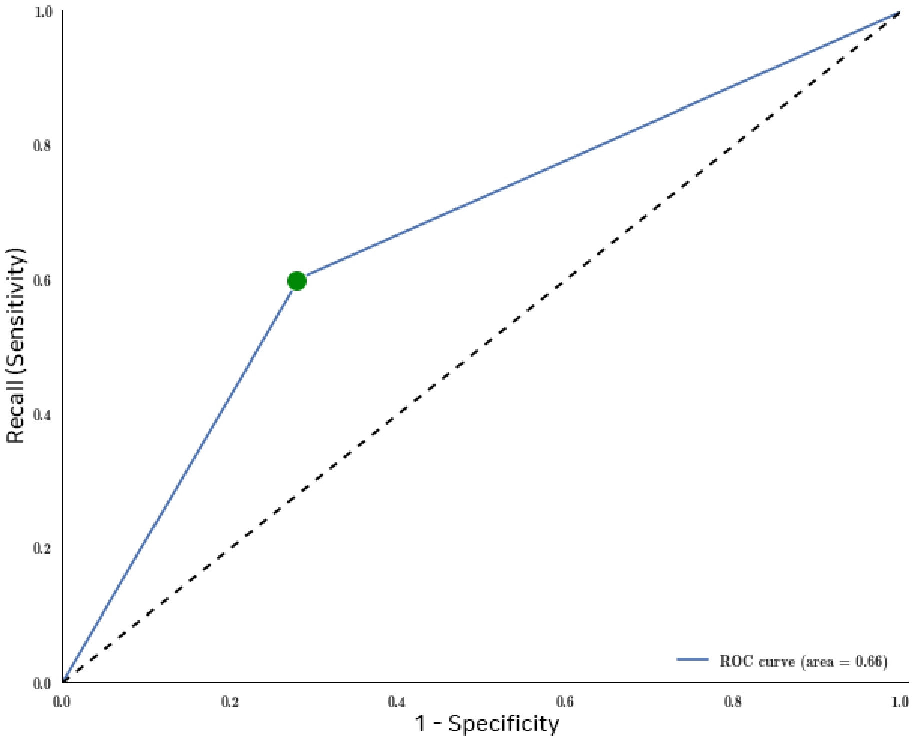


$$\text{Recall} = \frac{True\ Positive}{\text{True Positive} + \text{False Negative}}$$

$$\text{Specificity} = \frac{True\ Negative}{True\ \text{Negative} + False\ \text{Positive}}$$

4) ROC_AUC

How well classified?



$$\text{Recall} = \frac{True\ Positive}{\text{True Positive} + \text{False Negative}}$$

$$\text{Specificity} = \frac{True\ Negative}{True\ \text{Negative} + False\ \text{Positive}}$$

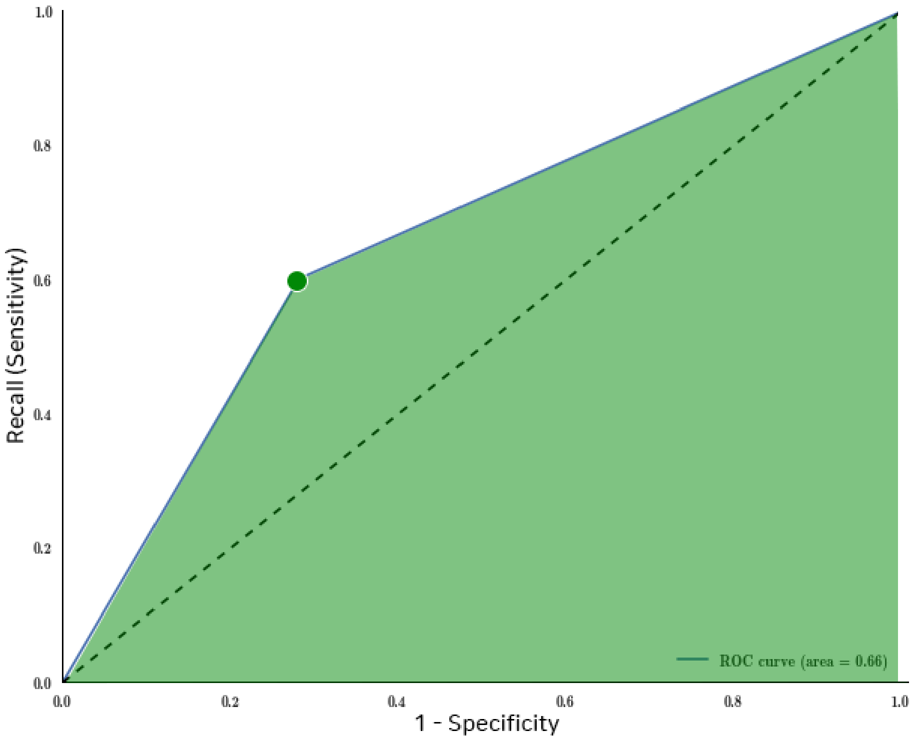
ROC Graph

(1-Specificity, Recall) 값을 좌표평면에 도식하고

그래프 원점, 종점과 연결한 Graph ($0 \le x,\ y \le 1$)

4) ROC_AUC

How well classified?



$$\text{Recall} = \frac{\mathit{True\ Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{Specificity} = \frac{\mathit{True\ Negative}}{\mathit{True}\ \text{Negative} + \mathit{False}\ \text{Positive}}$$

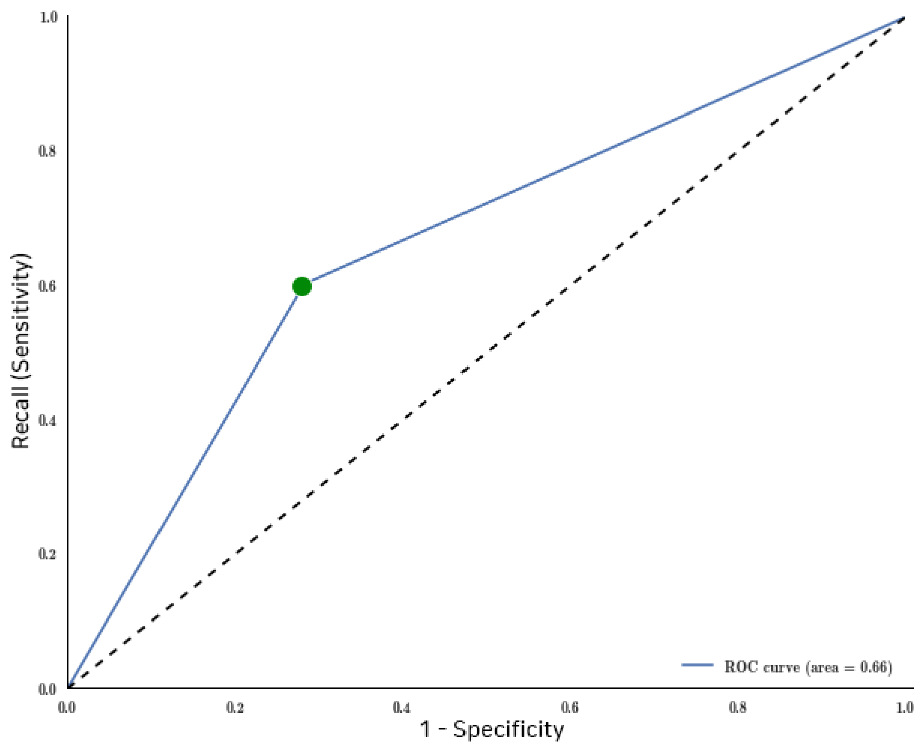
ROC_AUC

ROC Graph의 밑 면적 넓이

[0,1] 값을 가지며, Recall과 Specificity를 고려한 지표
1에 가까울 수록 잘 분류되었다고 평가

4) ROC_AUC

How well classified?



ROC Curve?

4) ROC_AUC

How well classified?

Recall, Precision, Specificity는 모두 Confusion Matrix를 통해 도출되는 값

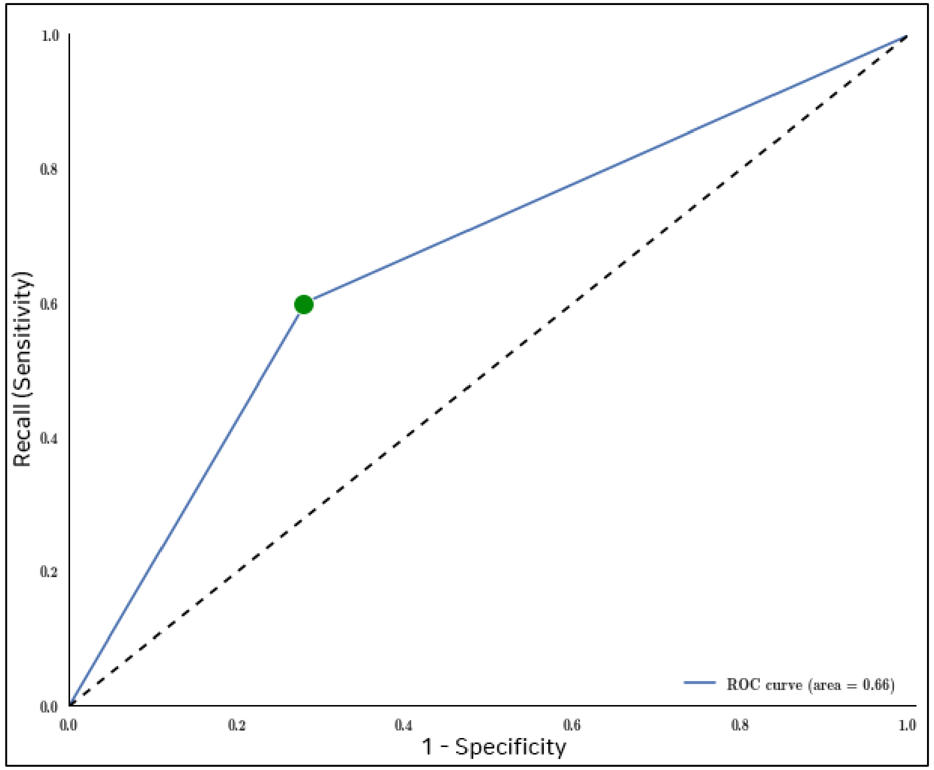
isFraud (실제 값)	Predict Probability (예측 확률)
1	0.7
1	0.6
0	0.4
0	0.3

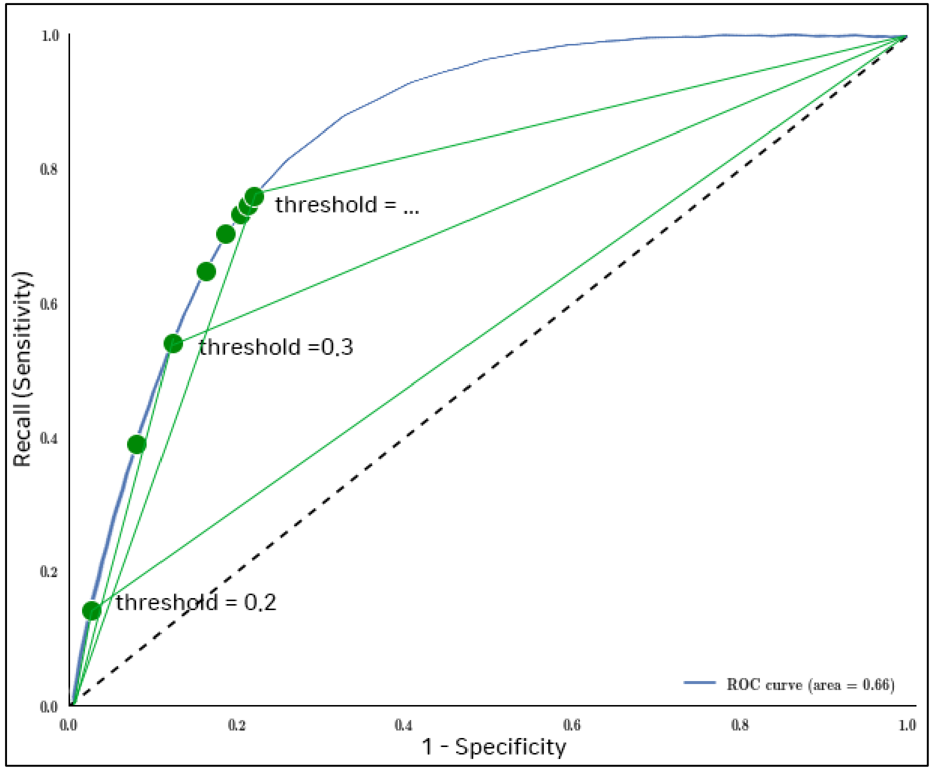
Prediction 1 (threshold = 0.5)	Prediction 2 (threshold = 0.2)	Prediction 3 (threshold = 0.8)
1	1	0
1	0	0
0	0	0
0	0	0

예측 확률을 도출하더라도, threshold 설정에 따라
하나의 모델에서 다수의 Confusion Matrix가 만들어진다

4) ROC_AUC

How well classified?





Threshold 선언 시

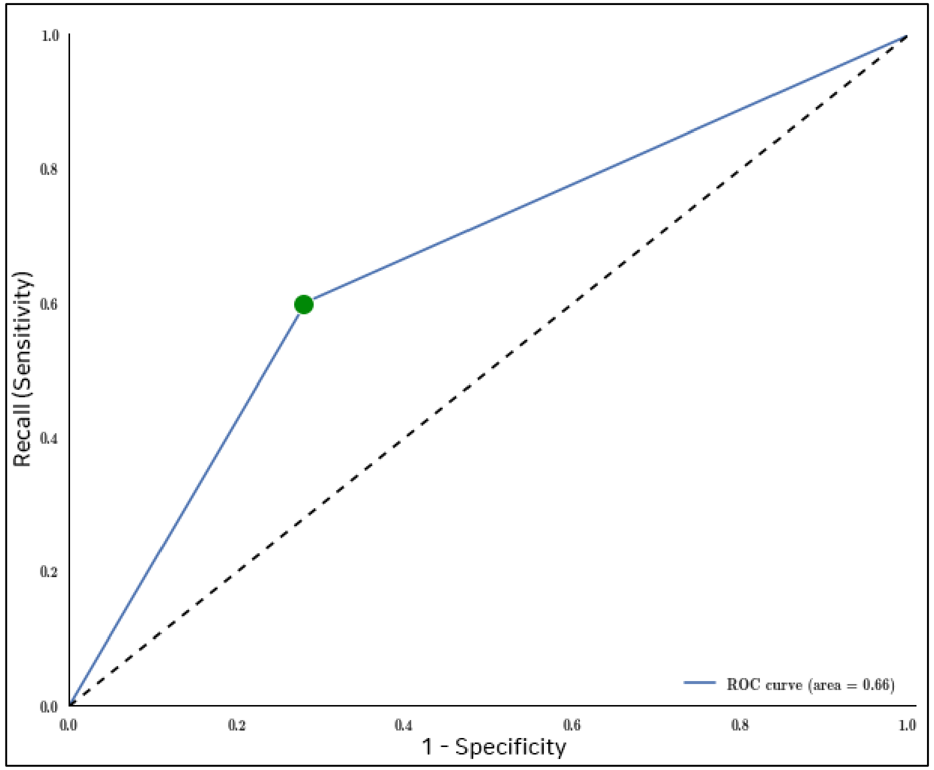
(1,0 binary 형태의 output)

Threshold 미 선언 시

(Probability 형태의 output)

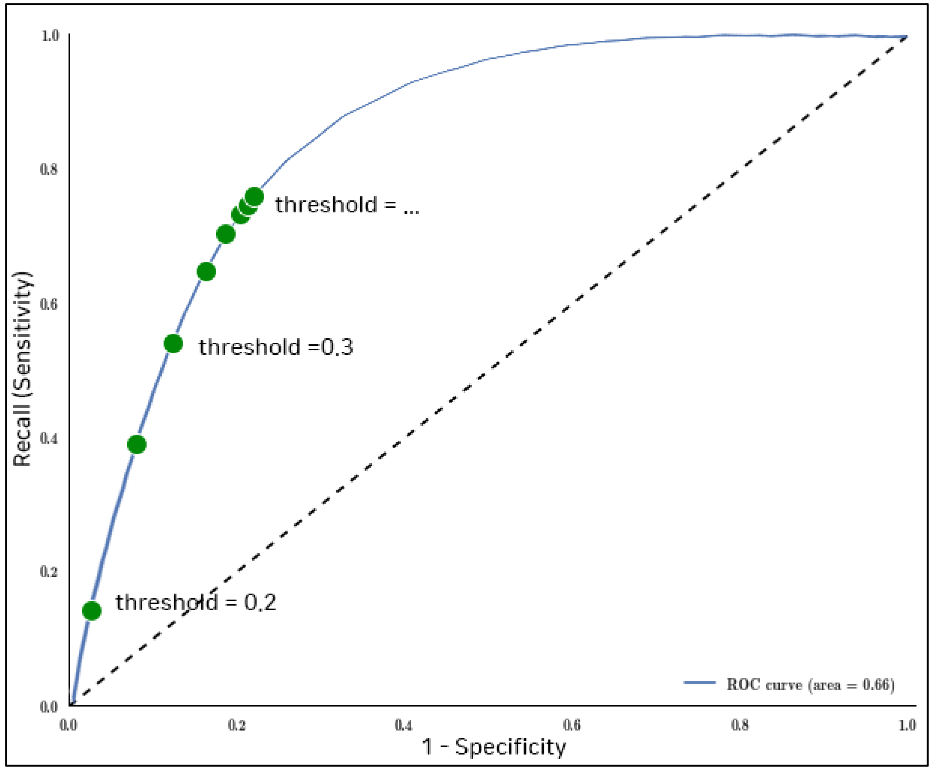
4) ROC_AUC

How well classified?



Threshold 선언 시

(1,0 binary 형태의 output)

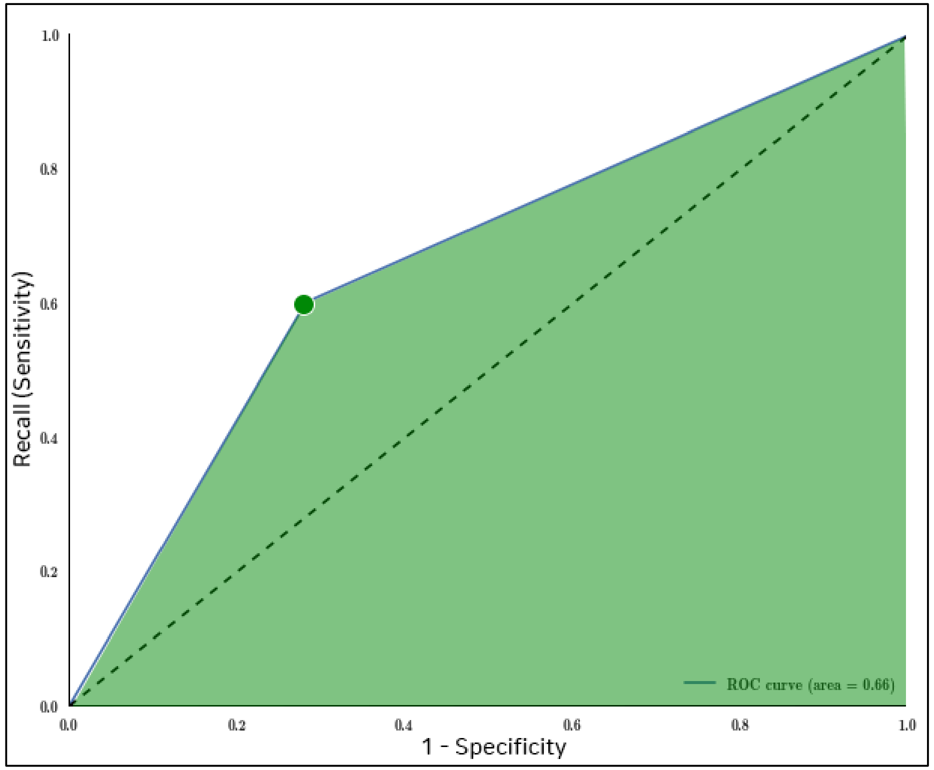


Threshold 미 선언 시

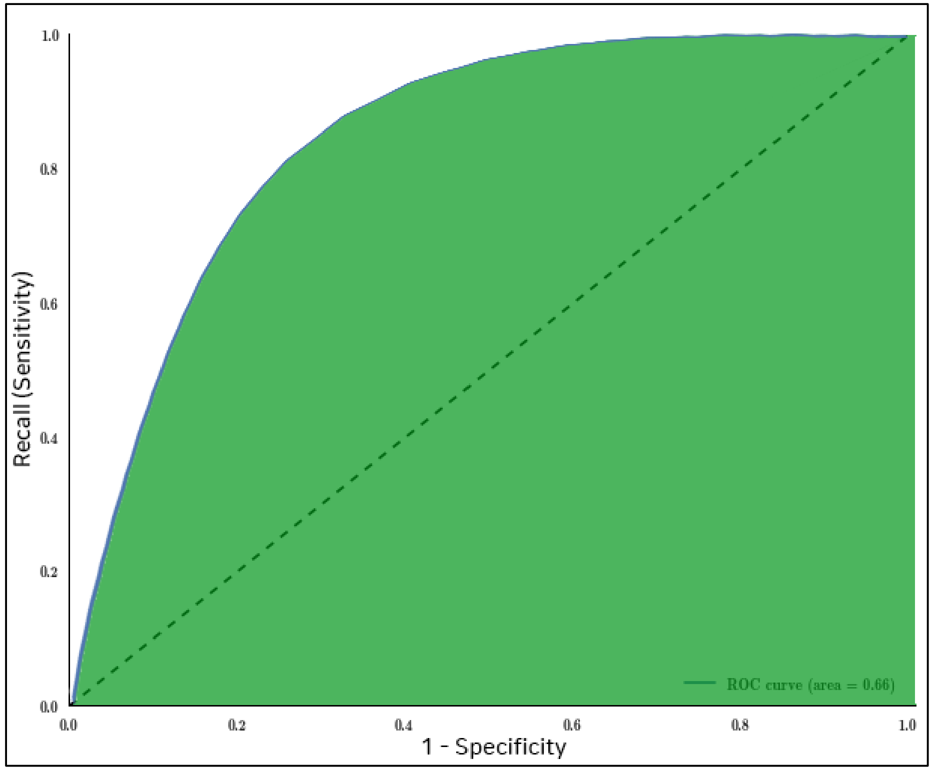
(Probability 형태의 output)

4) ROC_AUC

How well classified?



Kaggle Score



Threshold 선언 시
(1,0 binary 형태의 output)

Threshold 미 선언 시
(Probability 형태의 output)

3. Modeling & Tuning

Sampling, Classifier, Validation, Ensemble

1) Precautions: 완벽한 단일 평가 지표?

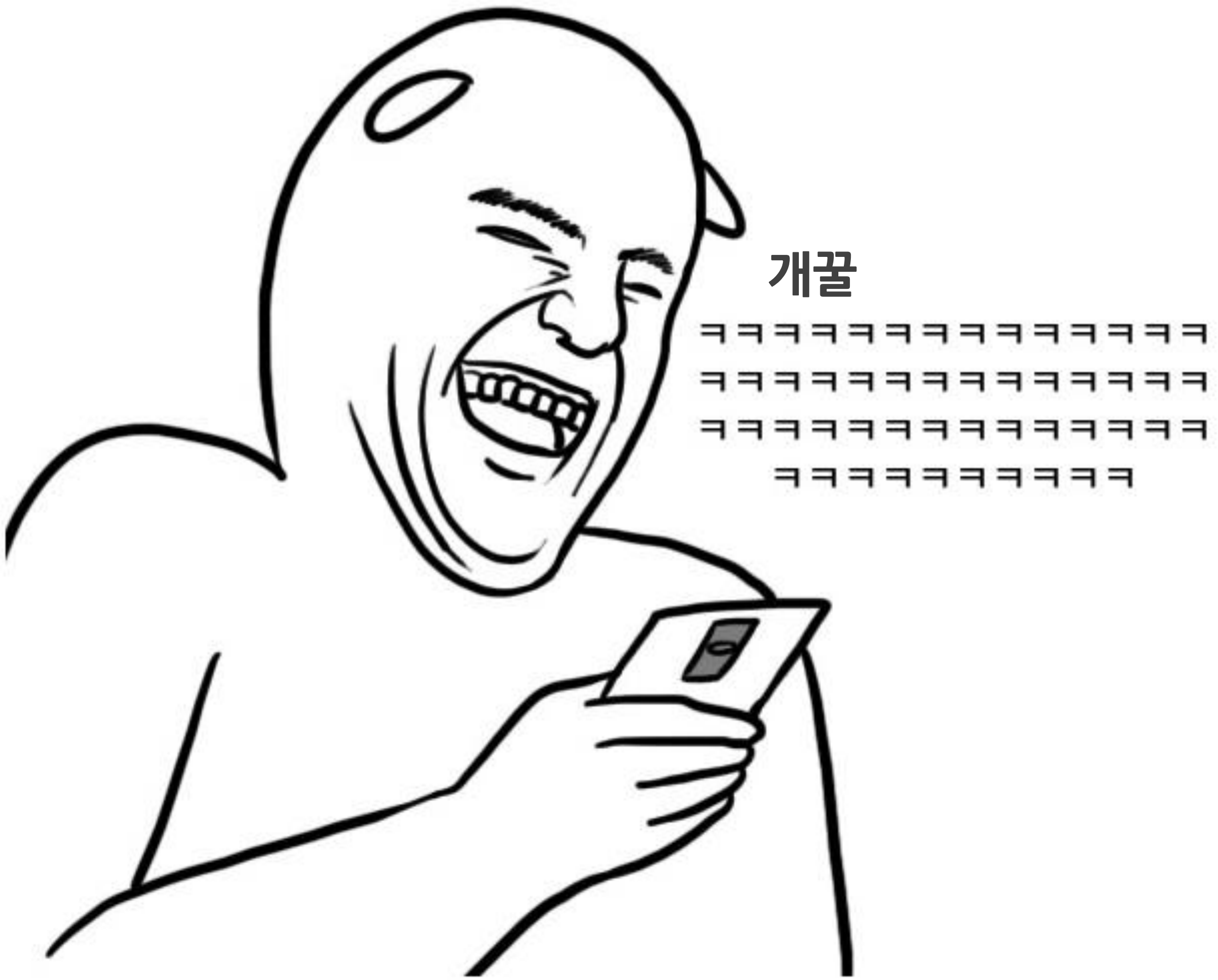
```python
# pred = clf.predict(X_test)
# print('F1:{}'.format(f1_score(y_test, pred)))

print('F1:{}'.format(f1_score(downsampled['isFraud'], oof > 0.5)),
      'ROC_AUC:{}'.format(roc_auc_score(downsampled['isFraud'], oof > 0.5)), sep = '\n')
```

```
F1:0.4333485013354437
ROC_AUC:0.8165067995934763
```

Submission and Description	Public Score
prob.csv 2 days ago by HelloWorld boom	0.8859

...?



개꿀
ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ

1) Precautions: 완벽한 단일 평가 지표?

	Fraud(1)	Not Fraud(0)
Fraud(1)	True Positive (TP)	False Positive (FP)
Not Fraud(0)	False Negative (FN)	True Negative (TN)

F1 score

	Fraud(1)	Not Fraud(0)
Fraud(1)	True Positive (TP)	False Positive (FP)
Not Fraud(0)	False Negative (FN)	True Negative (TN)

ROC_AUC score

1) Precautions: 완벽한 단일 평가 지표?

Goal:

F1, ROC_AUC score 모두 균등하게 높은 Model

2) Modeling

Data Sampling	Classifier	Cross Validation	Ensemble
Under Sampling	Random Forest	K-Fold	Voting
Over Sampling	LightGBM	Hold-out	Balanced Bagging
…	XGBoost	LOO	Stacking
	CatBoost	…	…
	…		

2) Modeling

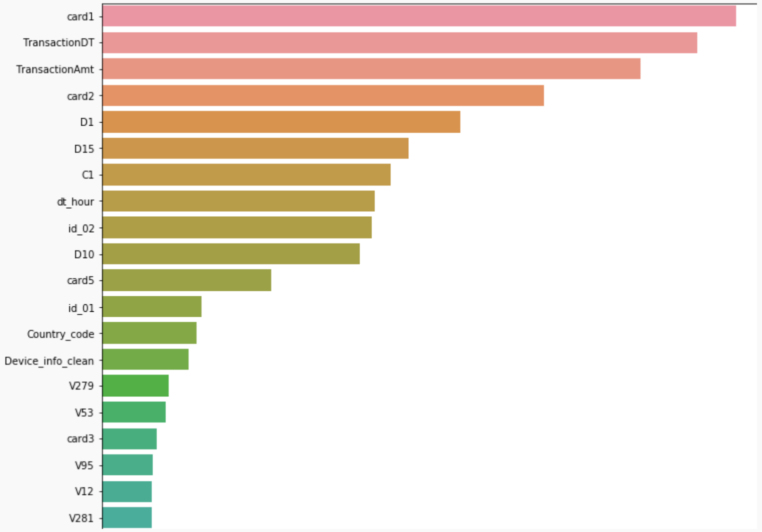
Final Model

Under Sampling	LightGBM	K-Fold	Soft Voting
Random Under Sampling Fraud = 20663 / Not Fraud = 20663 Number of Samples: 33	N estimators = 500 Learning rate = 0.15 Max depth = 9 …	K = 5 각 Fold별 Predict Probability의 Average	33개의 Sample에 대해 Soft Voting
Imbalanced Data의 효율적 학습	OHE 없이 Categorical Feature 처리 / NaN 처리	Sample 내에서의 Overfitting 방지	Data 전체 의견 수합

3) Output

1개 Sample에 대한 Validation Score

```
Precision:0.8825534820209376
Recall:0.8449707206117214
F1:0.8633485013354437
ROC_AUC:0.8665067995934763
```



Feature Importance

33개 모든 Sample에 대해

Prediction Probability 출력,

Soft Voting 진행

Submission and Description	Public Score
submission_average.csv 21 hours ago by HelloWorld helloworld	0.9100

최종 AUC score **0.91**

4) Limitations

1. Domain knowledge 활용 불가
2. Heuristic한 FE(결측치 처리 등)로 인해 성능 저하, 모델 적용이 제한 (Balanced Bagging Classifier 등)
3. Computing Power 문제로 최적의 Hyper Parameter Tuning, Sampling 제한

End of Document

Special thanks to

백상현, 유건욱, 이용하

Ybigta 15기