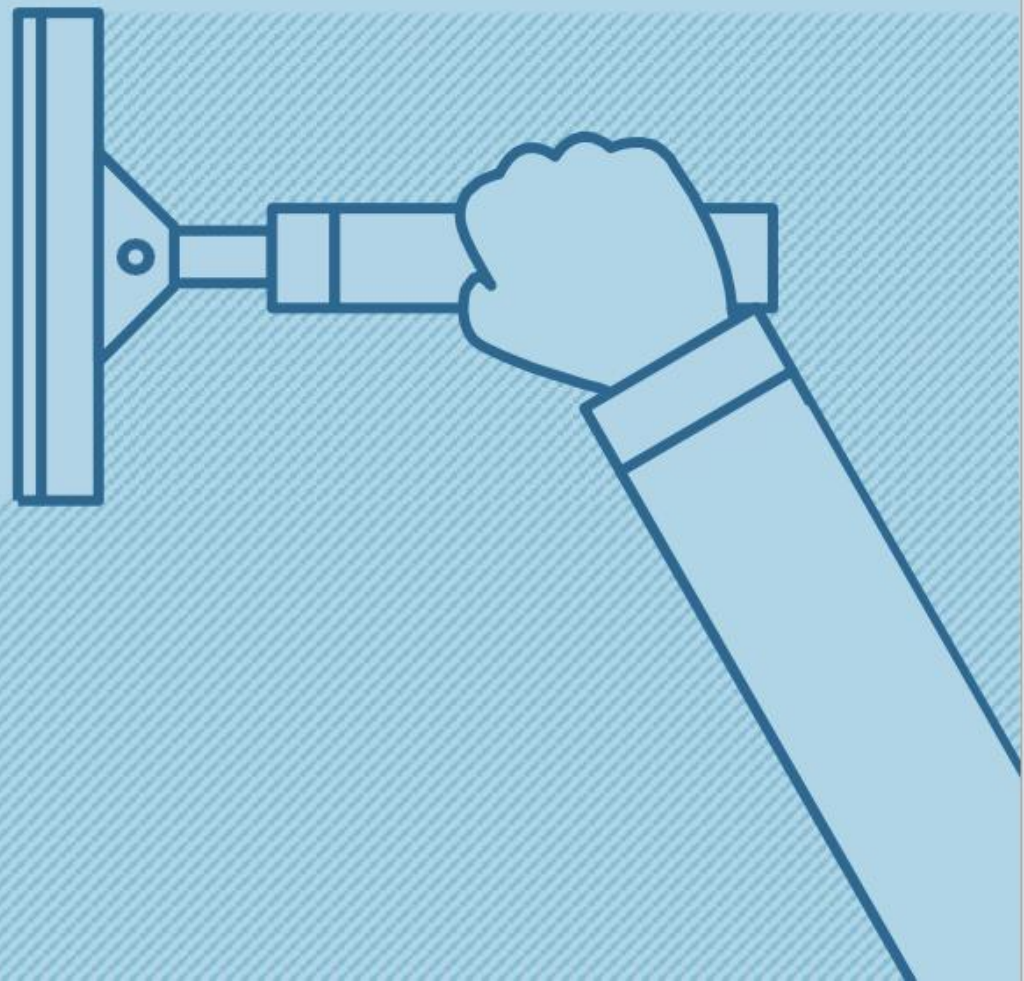
프로젝트 중간발표

3팀 강소영, 백진우, 우현수, 정현우, 정희영



CONTENTS

데이터 특징

전반적인 column별 특징

1. 데이터 특징

Transaction data

isFraud
TransactionDT
TransactionAmt
ProductCD
C_columns
D_columns
V_columns

(590540, 394)

&

Identity data

Id_01~39
DeviceInfo
DeviceType
.
.
.

(144233, 41)

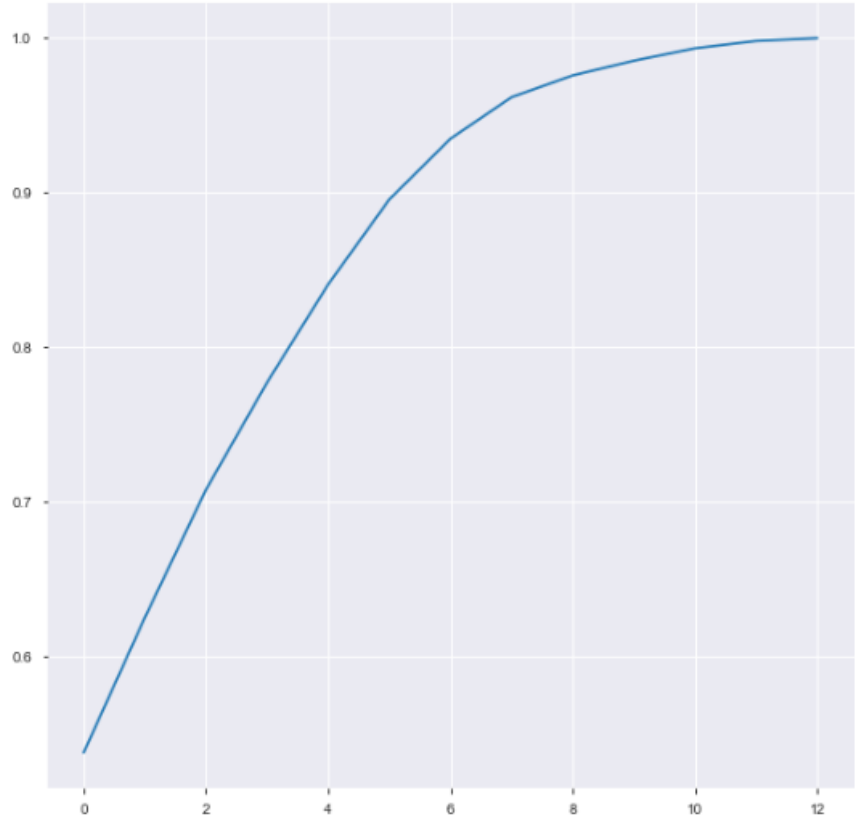
사기거래
예측

전처리 과정

PCA, Null 값 처리
및 column 재구성

2. 전처리 과정

identity Data: id_01 ~ id_38 / DeviceType / DeviceInfo



Identity data

➡ Binary encoding

DeviceType/DeviceInfo

➡ 외부 데이터를 끌어와서 상관관계 도출 시도

2. 전처리 과정

TransactionDT & TransactionAmt

	TransactionID	isFraud	TransactionDT	TransactionAmt	ProductCD	card1	card2	card3	card4	card5
0	2987000	0	86400	68.500	W	13926	NaN	150.0	discover	142.0
1	2987001	0	86401	29.000	W	2755	404.0	150.0	mastercard	102.0
2	2987002	0	86469	59.000	W	4663	490.0	150.0	visa	166.0
3	2987003	0	86499	50.000	W	18132	567.0	150.0	mastercard	117.0
4	2987004	0	86506	50.000	H	4497	514.0	150.0	mastercard	102.0
5	2987005	0	86510	49.000	W	5937	555.0	150.0	visa	226.0
6	2987006	0	86522	159.000	W	12308	360.0	150.0	visa	166.0
7	2987007	0	86529	422.500	W	12695	490.0	150.0	visa	226.0
8	2987008	0	86535	15.000	H	2803	100.0	150.0	visa	226.0
9	2987009	0	86536	117.000	W	17399	111.0	150.0	mastercard	224.0
10	2987010	0	86549	75.887	C	16496	352.0	117.0	mastercard	134.0
11	2987011	0	86555	16.495	C	4461	375.0	185.0	mastercard	224.0
12	2987012	0	86564	50.000	W	3786	418.0	150.0	visa	226.0
13	2987013	0	86585	40.000	W	12866	303.0	150.0	visa	226.0
14	2987014	0	86596	10.500	W	11839	490.0	150.0	visa	226.0

TransactionDT

➡ 거래 발생 시점 도출

TransactionAmt

➡ 사기거래의 특이점을 발견

2. 전처리 과정

TransactionDT & TransactionAmt

TransactionID	isFraud	TransactionDT	TransactionAmt	ProductCD	card1	card2	ca
2987203	1	89760	445.000	W	18268	583.0	1
2987240	1	90193	37.098	C	13413	103.0	1
2987243	1	90246	37.098	C	13413	103.0	1
2987245	1	90295	37.098	C	13413	103.0	1
2987288	1	90986	155.521	C	16578	545.0	1
2987367	1	92350	225.000	R	4425	562.0	1
2987405	1	92999	90.570	C	4504	500.0	1
2987630	1	97843	12.326	C	5812	408.0	1
2987683	1	99584	124.344	C	5812	408.0	1
2987736	1	100591	100.000	W	15063	NaN	1
2987779	1	102154	10.000	S	7481	364.0	1
2987780	1	102188	10.000	S	8732	360.0	1
2987781	1	102193	10.000	S	8732	360.0	1
2987869	1	106603	83.380	C	9026	545.0	1
2987923	1	108912	774.000	W	5033	269.0	1

TransactionDT

➡ 거래 발생 시점 도출

TransactionAmt

➡ 사기거래의 특이점을 발견

2. 전처리 과정

ProductCD

	TransactionID	isFraud	TransactionDT	TransactionAmt	ProductCD	card1	card2	card3	card4	card5
0	2987000	0	86400	68.500	W	13926	NaN	150.0	discover	142.0
1	2987001	0	86401	29.000	W	2755	404.0	150.0	mastercard	102.0
2	2987002	0	86469	59.000	W	4663	490.0	150.0	visa	166.0
3	2987003	0	86499	50.000	W	18132	567.0	150.0	mastercard	117.0
4	2987004	0	86506	50.000	H	4497	514.0	150.0	mastercard	102.0
5	2987005	0	86510	49.000	W	5937	555.0	150.0	visa	226.0
6	2987006	0	86522	159.000	W	12308	360.0	150.0	visa	166.0
7	2987007	0	86529	422.500	W	12695	490.0	150.0	visa	226.0
8	2987008	0	86535	15.000	H	2803	100.0	150.0	visa	226.0
9	2987009	0	86536	117.000	W	17399	111.0	150.0	mastercard	224.0
10	2987010	0	86549	75.887	C	16496	352.0	117.0	mastercard	134.0
11	2987011	0	86555	16.495	C	4461	375.0	185.0	mastercard	224.0
12	2987012	0	86564	50.000	W	3786	418.0	150.0	visa	226.0
13	2987013	0	86585	40.000	W	12866	303.0	150.0	visa	226.0
14	2987014	0	86596	10.500	W	11839	490.0	150.0	visa	226.0

ProductCD

➡ 더미변수화 시켰습니다.

2. 전처리 과정

Card ~ emaildomain columns

card1	card2	card3	card4	card5	card6	addr1	addr2	dist1	dist2	P_emaildomain	R_emaildomain	C1	C2	C3
13926	NaN	150.0	discover	142.0	credit	315.0	87.0	19.0	NaN	NaN	NaN	1.0	1.0	0.0
2755	404.0	150.0	mastercard	102.0	credit	325.0	87.0	NaN	NaN	gmail.com	NaN	1.0	1.0	0.0
4663	490.0	150.0	visa	166.0	debit	330.0	87.0	287.0	NaN	outlook.com	NaN	1.0	1.0	0.0
18132	567.0	150.0	mastercard	117.0	debit	476.0	87.0	NaN	NaN	yahoo.com	NaN	2.0	5.0	0.0
4497	514.0	150.0	mastercard	102.0	credit	420.0	87.0	NaN	NaN	gmail.com	NaN	1.0	1.0	0.0
5937	555.0	150.0	visa	226.0	debit	272.0	87.0	36.0	NaN	gmail.com	NaN	1.0	1.0	0.0
12308	360.0	150.0	visa	166.0	debit	126.0	87.0	0.0	NaN	yahoo.com	NaN	1.0	1.0	0.0
12695	490.0	150.0	visa	226.0	debit	325.0	87.0	NaN	NaN	mail.com	NaN	1.0	1.0	0.0
2803	100.0	150.0	visa	226.0	debit	337.0	87.0	NaN	NaN	anonymous.com	NaN	1.0	1.0	0.0
17399	111.0	150.0	mastercard	224.0	debit	204.0	87.0	19.0	NaN	yahoo.com	NaN	2.0	2.0	0.0
16496	352.0	117.0	mastercard	134.0	credit	NaN	NaN	NaN	NaN	gmail.com	gmail.com	1.0	4.0	0.0
4461	375.0	185.0	mastercard	224.0	debit	NaN	NaN	NaN	30.0	hotmail.com	hotmail.com	1.0	1.0	0.0
3786	418.0	150.0	visa	226.0	debit	204.0	87.0	NaN	NaN	verizon.net	NaN	4.0	2.0	0.0
12866	303.0	150.0	visa	226.0	debit	330.0	87.0	NaN	NaN	aol.com	NaN	6.0	5.0	0.0
11839	490.0	150.0	visa	226.0	debit	226.0	87.0	NaN	NaN	yahoo.com	NaN	1.0	1.0	0.0

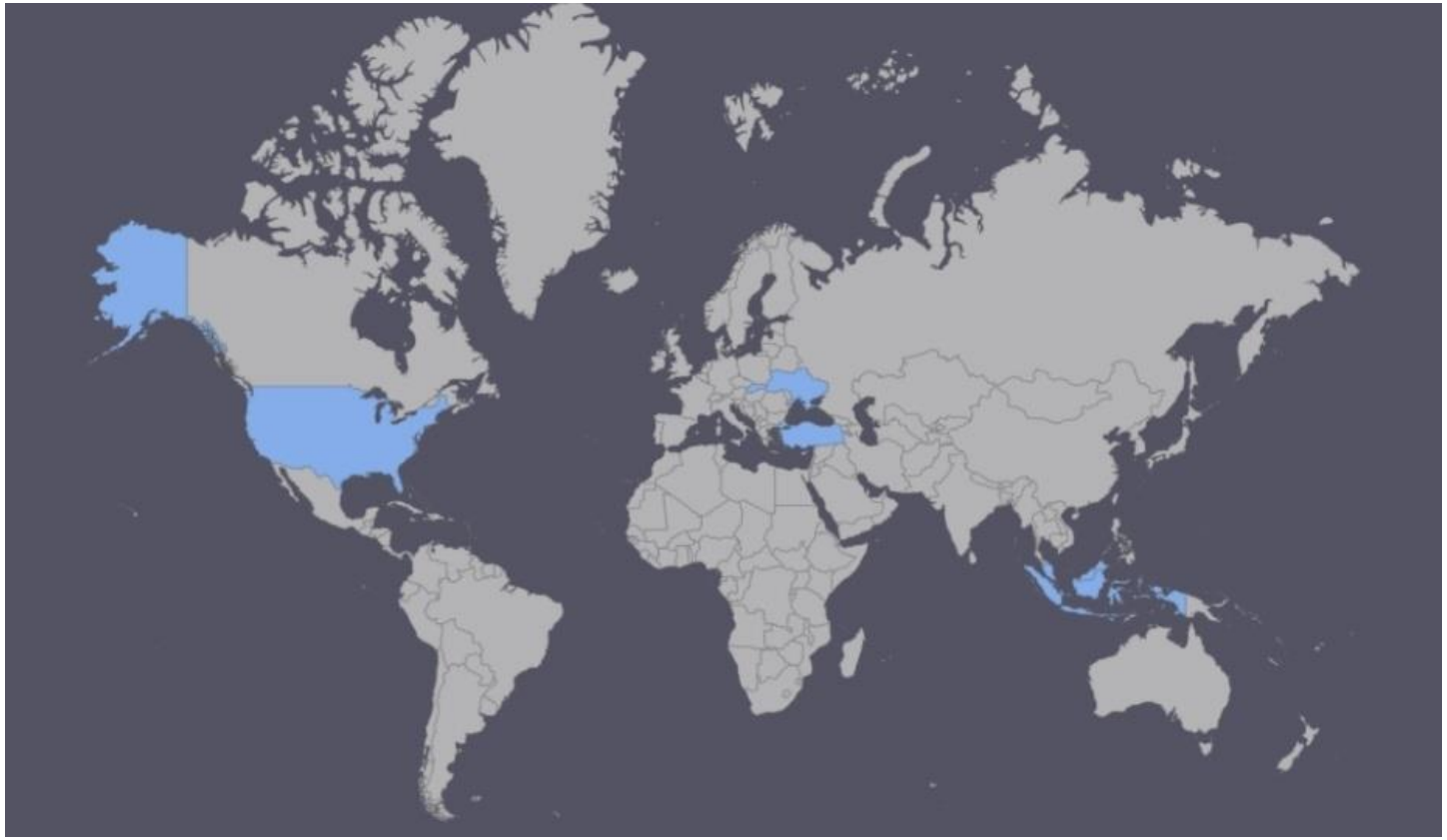
card, addr, dist, emaildomain

➡ 유의미한 상관관계를 보인 칼럼만 사용

'국가'를 의미하는 칼럼은?

2. 전처리 과정

Card ~ emaildomain columns



사기거래 Top 6 국가

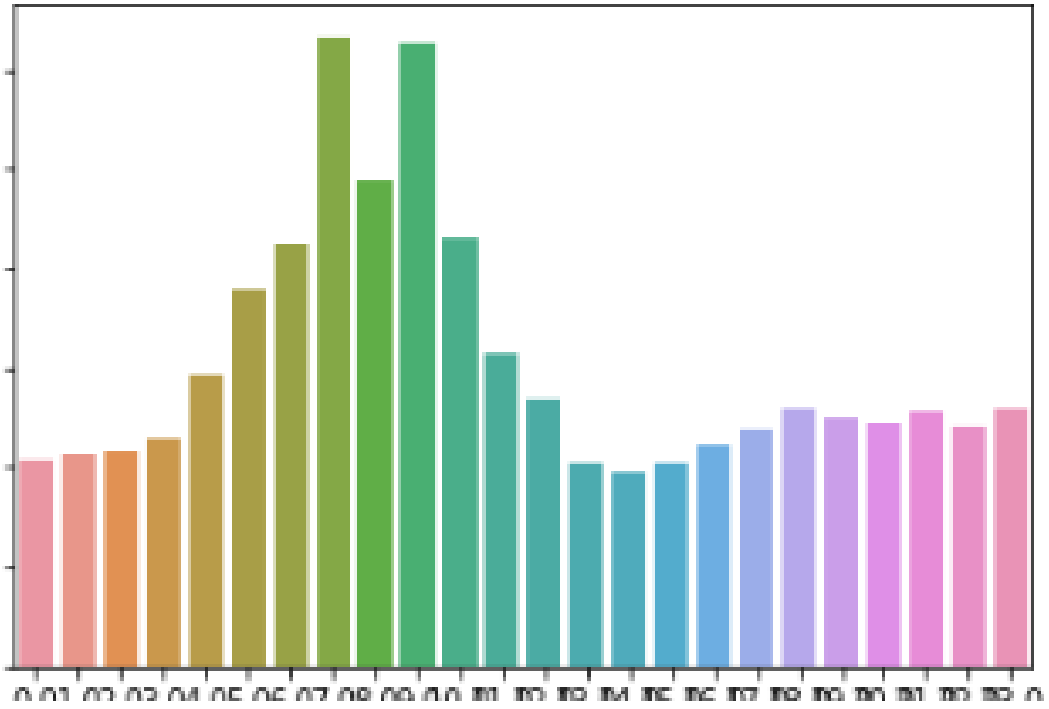
우크라이나　인도네시아　터키
유고슬라비아　말레이시아　미국

사기율이 높은 국가의

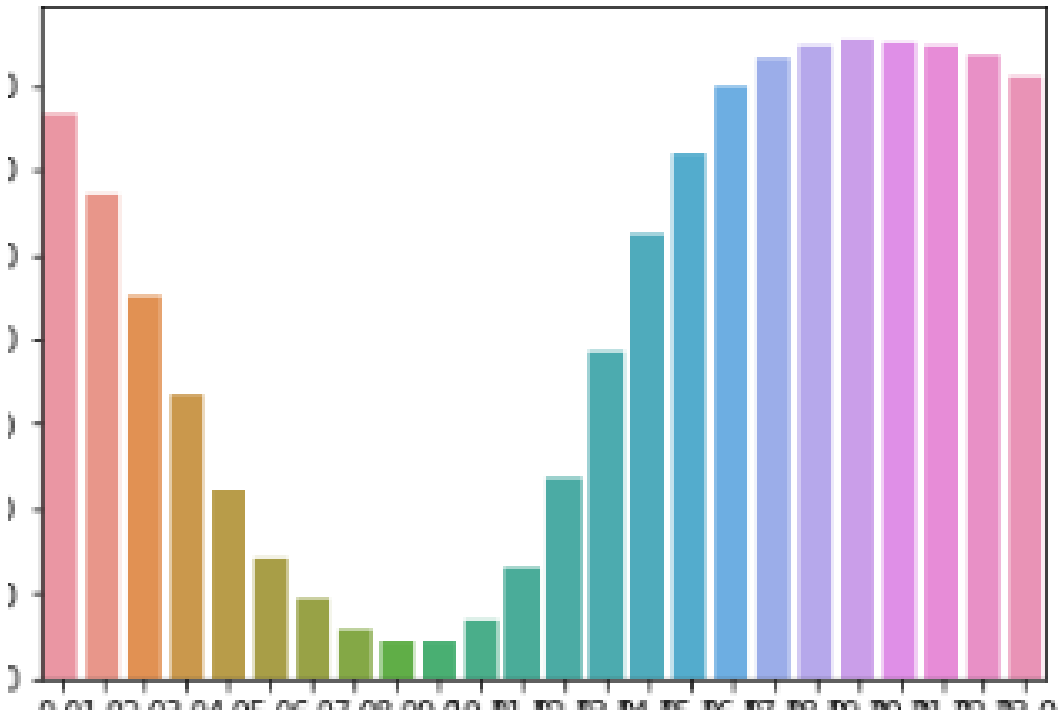
사기거래 추이 파악

2. 전처리 과정

Card ~ emaildomain columns



사기 카드거래



일반 카드거래

2. 전처리 과정

Card ~ emaildomain columns

	Percent	num
2	0.000000	1
3	0.071429	84
3	0.133333	90
4	0.156250	32
5	0.161812	309
1	0.166667	12
4	0.230769	13
4	0.289474	38

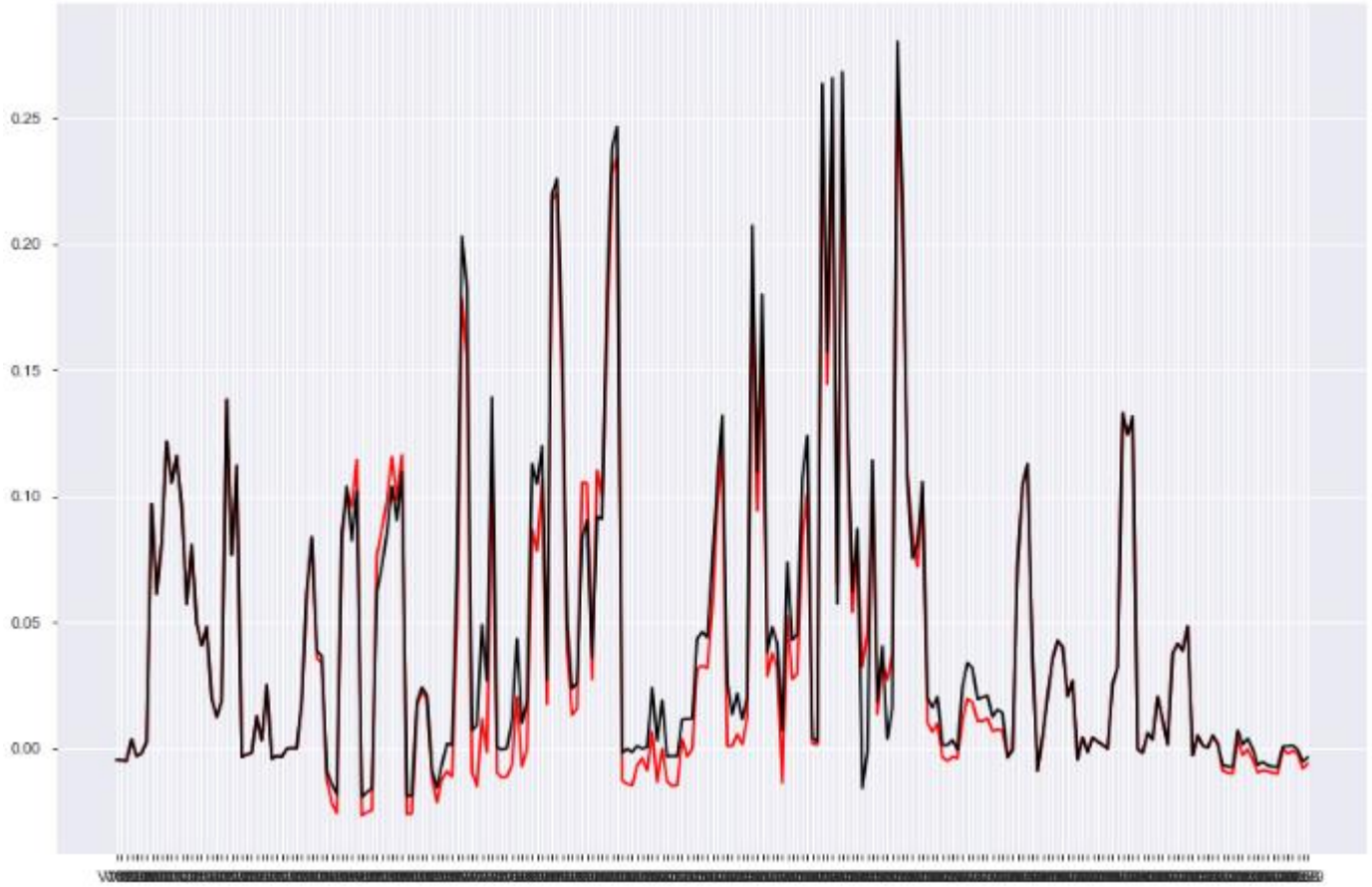
dist1, dist2

➡ Section 중에서 특정 section에서만 사기거래가 발생

➡ Binary column 추가

2. 전처리 과정

C, D, V-columns



C-columns

➡ isFraud와의 상관관계를 반영

D-columns

➡ 사용가능한 column들도 추림

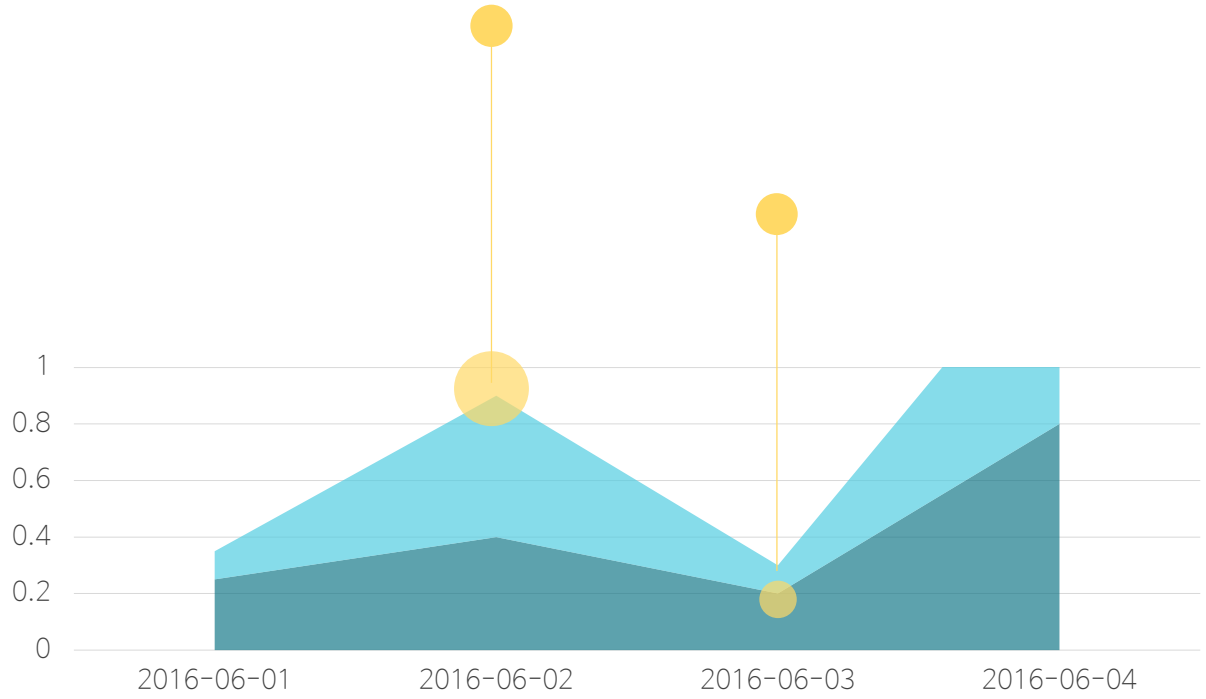
V-columns

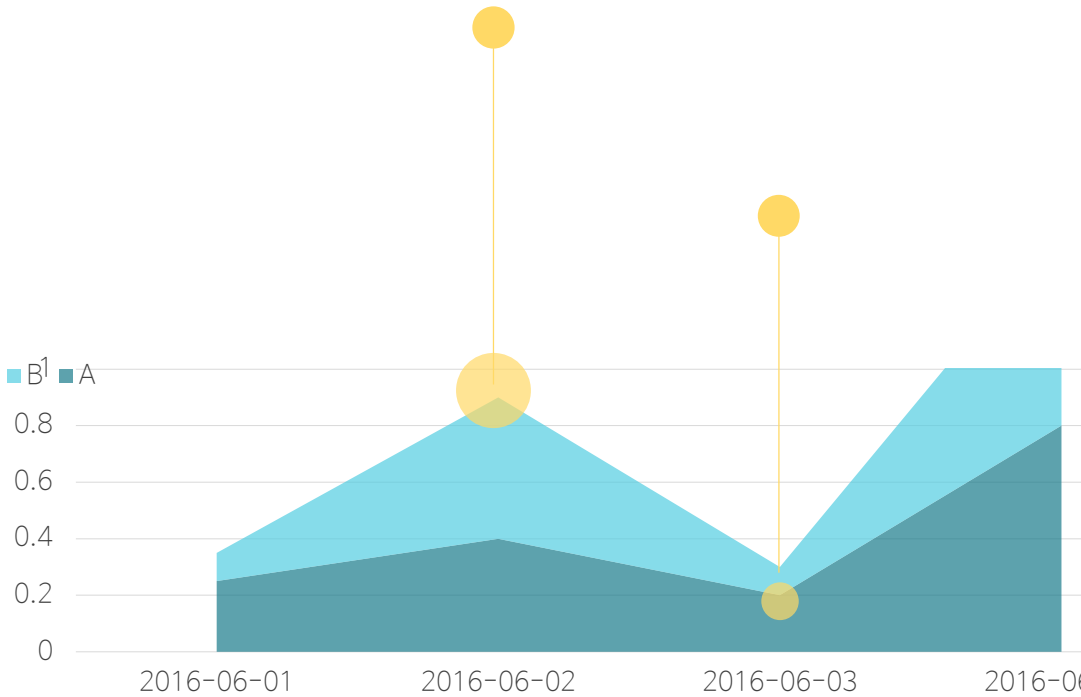
➡ Median imputation 후 PCA

사용 모델 및 성능평가

사용 모델 선정 및 성능

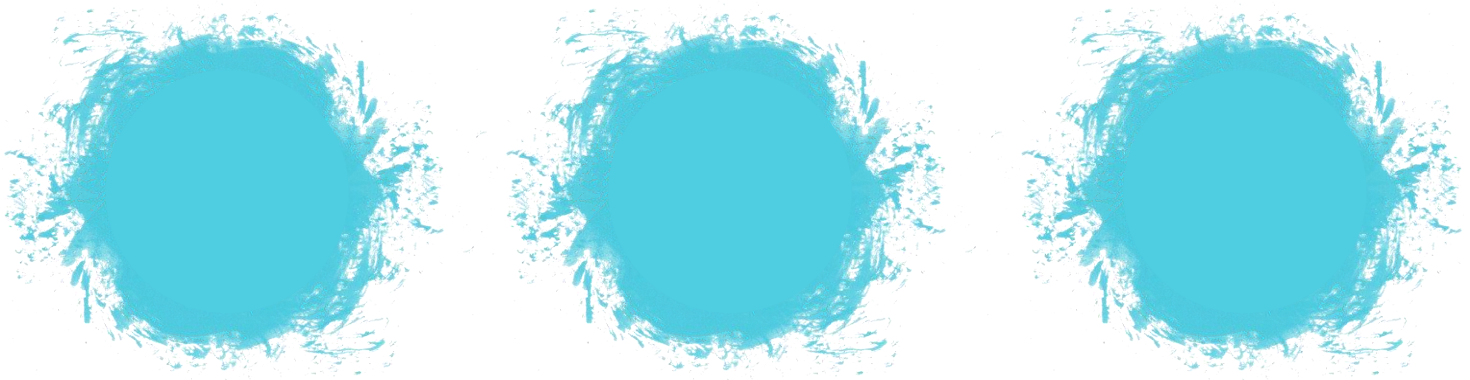
2. 전처리 과정


1
0.8
0.6
0.4
0.2
0
2016-06-01
2016-06-02
2016-06-03
2016-06-04


B
A
0.8
0.6
0.4
0.2
0
2016-06-01
2016-06-02
2016-06-03

2. 전처리 과정

- V-columns



감사합니다