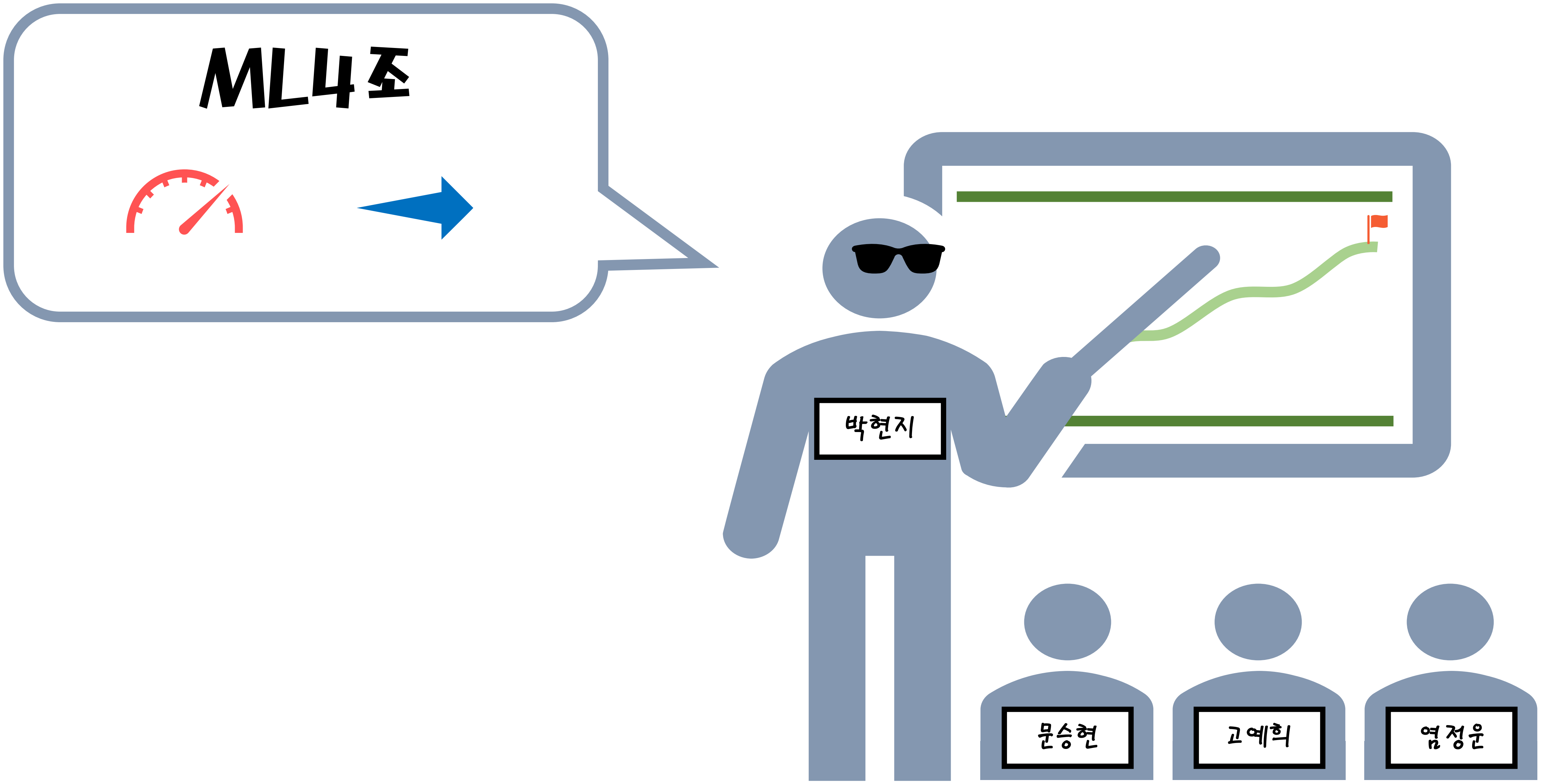
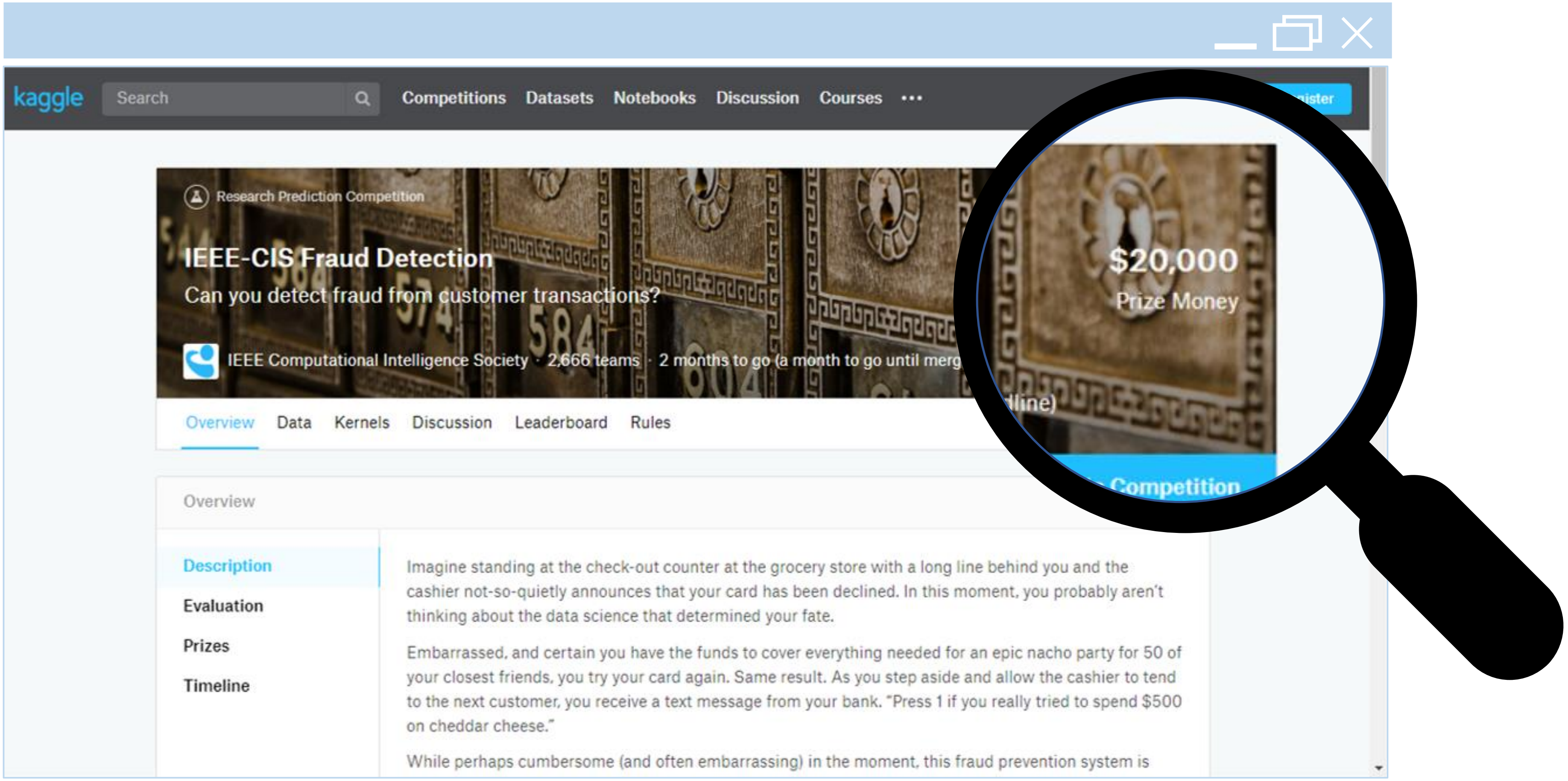

ML4조
박현지
문승현
고예희
염정운

목 차

1. 목적

- 1.1 데이터소개
- 1.2 변수소개
- 1.3 최종목표

1.1 데이터소개



거래데이터

참/거짓 판단

1.2 변수소개

train_transaction.csv

TransactionDT TransactionAMT

ProductCD card1 ~ card6

addr dist P_R_emaildomain

C1-C14 D1-D15 M1-M19 Vxxx

train_identity.csv

DeviceType

DeviceInfo

id1 - id38

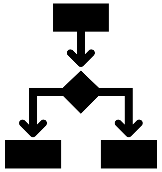
test_transaction.csv

모델 성능 테스트

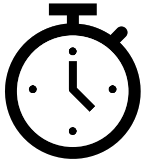
test_identity.csv

모델 성능 테스트

1.3 최종목표



모델링



빠르고

정확하게

참/거짓

2. 변수분석

2.1 변수설명

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

TransactionDT

주어진 datetime으로 부터의 timedelta

categorical variable

범주형 변수

continuous variable

연속형 변수

2.1 변수설명

train_transaction.csv

TransactionID TransactionDT

TransactionAmt ProductCD

card1 card2 card3 card4 card5 card6

Addr1 addr2 dist1 dist2

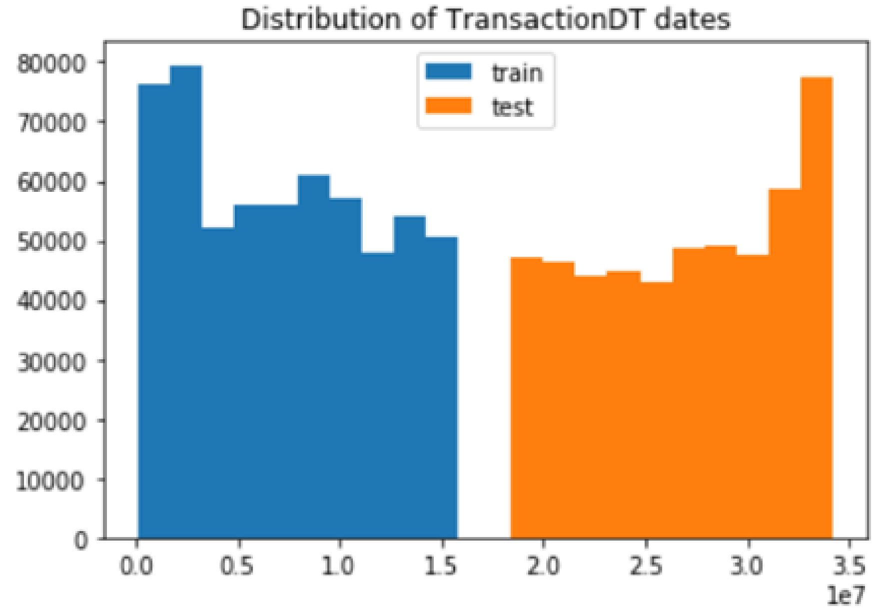
P_emaildomain R_emaildomain

C1-C14 D1-D15 M1-M19 Vxxx

categorical variable: 범주형 변수

continuous variable: 연속형 변수

Text(0.5, 1.0, 'Distribution of TransactionDT dates')



Train data와 Test data 사이의 overlapping area가 없다.

2.1 변수설명

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

TransactionAmt

거래 결제 금액(USD)

categorical variable

범주형 변수

continuous variable

연속형 변수

2.1 변수설명

train_transaction.csv

TransactionID TransactionDT

TransactionAmt ProductCD

card1 card2 card3 card4 card5 card6

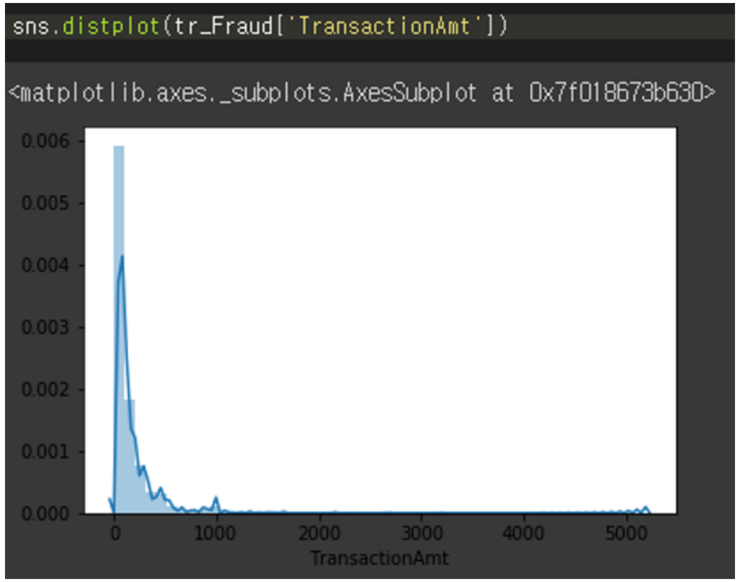
Addr1 addr2 dist1 dist2

P_emaildomain R_emaildomain

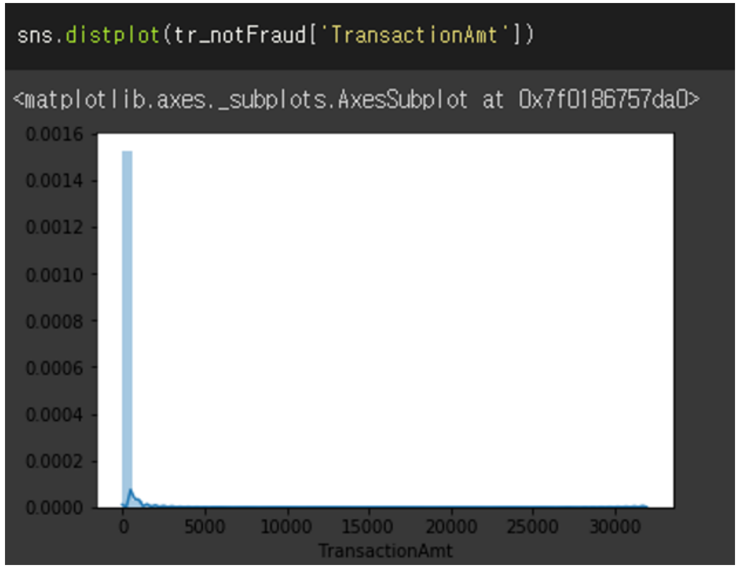
C1-C14 D1-D15 M1-M19 Vxxx

categorical variable
범주형 변수

continuous variable
연속형 변수



Fraud



Not Fraud

2.1 변수설명

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

ProductCD

제품 코드, 각 거래의 제품

categorical variable

범주형 변수

continuous variable

연속형 변수

2.1 변수설명

train_transaction.csv

TransactionID TransactionDT

TransactionAmt **ProductCD**

card1 card2 card3 card4 card5 card6

Addr1 addr2 dist1 dist2

P_emaildomain R_emaildomain

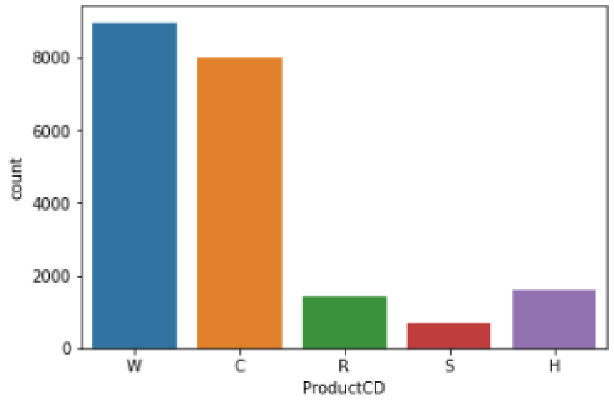
C1-C14 D1-D15 M1-M19 Vxxx

categorical variable
범주형 변수

continuous variable
연속형 변수

```
sns.countplot(x = 'ProductCD', data= tr_Fraud, order = ['W','C', 'R', 'S','H'])
# plt.title("Fraud Data의 ProductCD")
# plt.show()
```

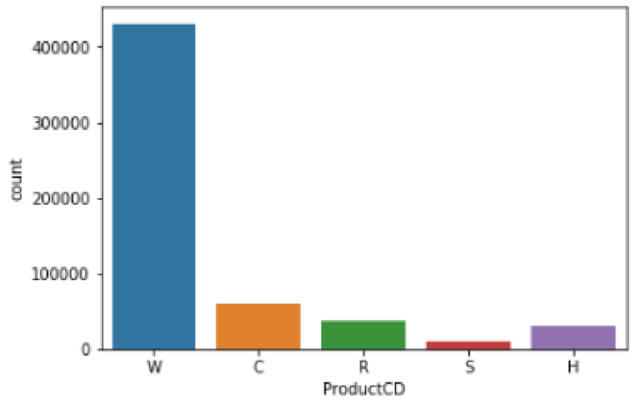
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc29a3a9f60>
```



Fraud

```
[44] sns.countplot(x = 'ProductCD', data= tr_notFraud, order = ['W','C', 'R', 'S','H'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc29a2e1b00>
```



Not Fraud

2.1 변수설명

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

categorical variable
범주형 변수

continuous variable
연속형 변수

card1
카드 관련 변수

card2
카드 관련 변수

card3
카드 관련 변수

card4
카드 종류

card5
카드 관련 변수

card6
카드 유형

2.1 변수설명

train_transaction.csv

TransactionID TransactionDT

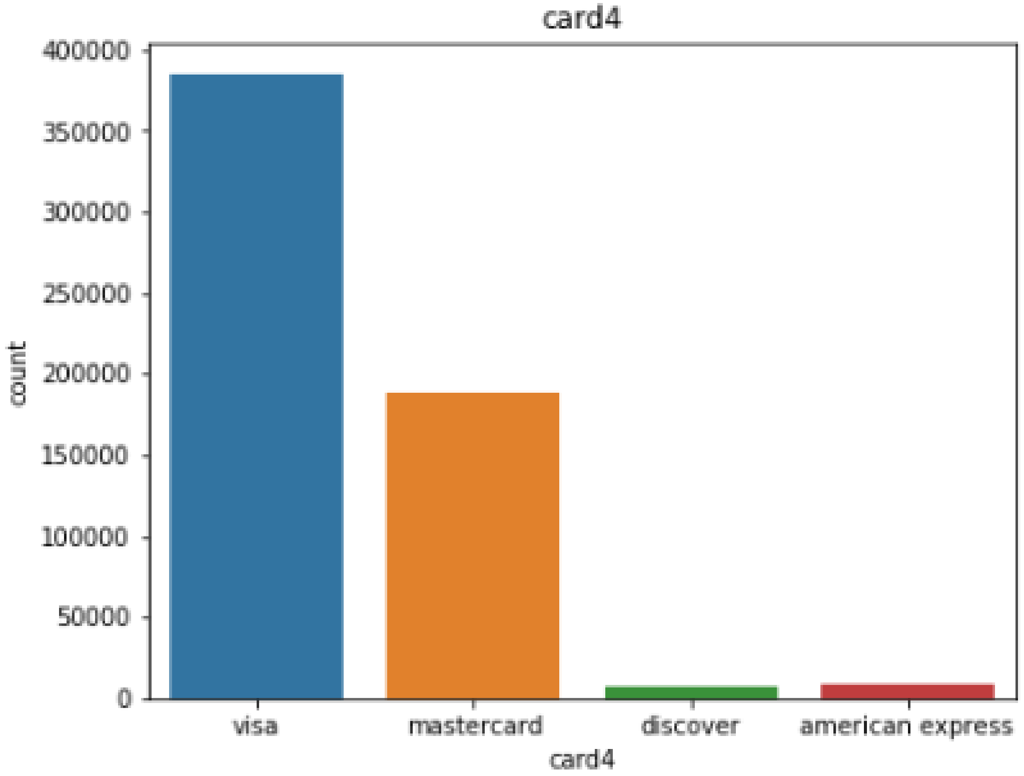
TransactionAmt ProductCD

card1 card2 card3 card4 card5 card6

Addr1 addr2 dist1 dist2

P_emaildomain R_emaildomain

C1-C14 D1-D15 M1-M19 Vxxx



categorical variable

범주형 변수

continuous variable

연속형 변수

2.1 변수설명

train_transaction.csv

TransactionID TransactionDT

TransactionAmt ProductCD

card1 card2 card3 card4 card5 card6

Addr1 addr2 dist1 dist2

P_emaildomain R_emaildomain

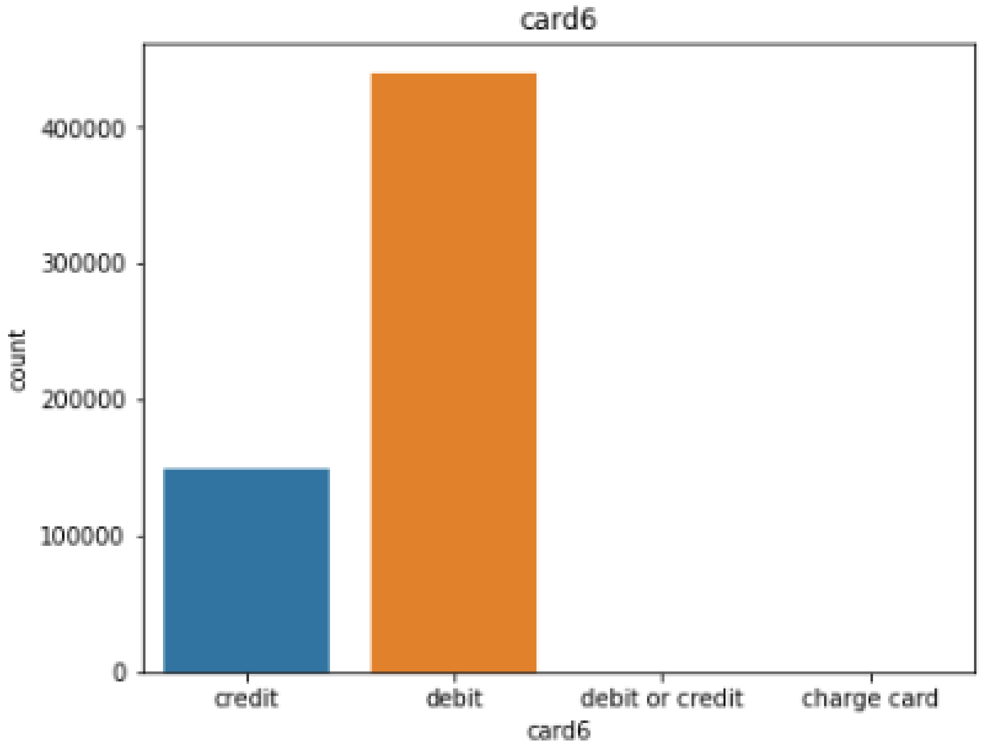
C1-C14 D1-D15 M1-M19 Vxxx

categorical variable

범주형 변수

continuous variable

연속형 변수



2.1 변수설명

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

categorical variable

범주형 변수

continuous variable

연속형 변수

addr1

도시 주소?

addr2

국가 주소?

2.1 변수설명

train_transaction.csv

TransactionID TransactionDT

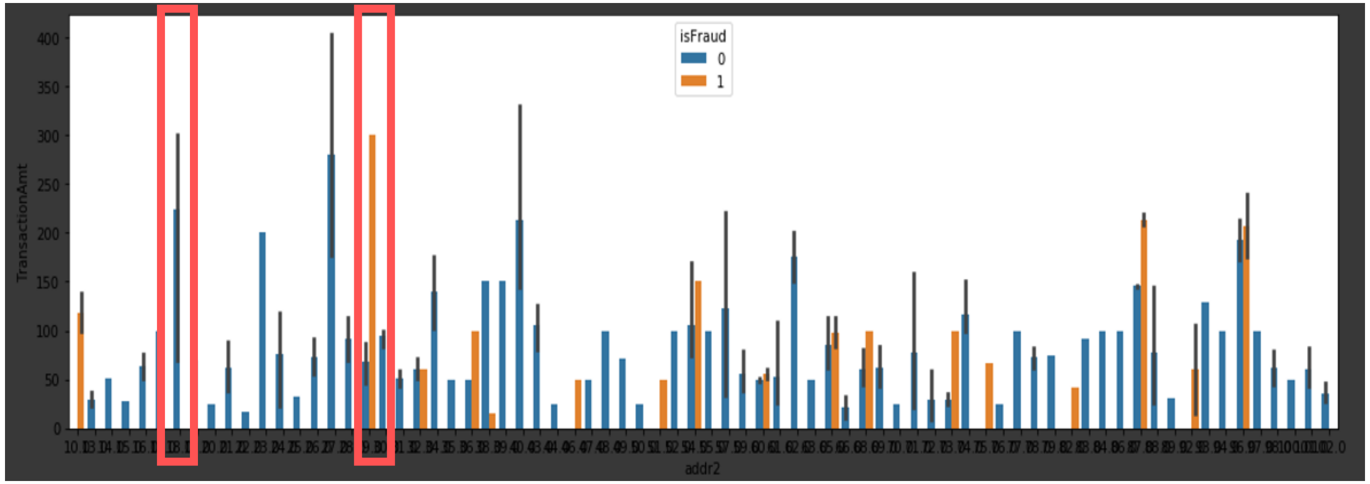
TransactionAmt ProductCD

card1 card2 card3 card4 card5 card6

Addr1 addr2 dist1 dist2

P_emaildomain R_emaildomain

C1-C14 D1-D15 M1-M19 Vxxx



categorical variable

범주형 변수

continuous variable

연속형 변수

특정값에만 데이터가 집중된다.
특정값에서 Fraud / Not Fraud 차이가 크다.

2.1 변수설명

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

categorical variable

범주형 변수

continuous variable

연속형 변수

dist1

거리 값

dist2

거리 값

2.1 변수설명

train_transaction.csv

TransactionID TransactionDT

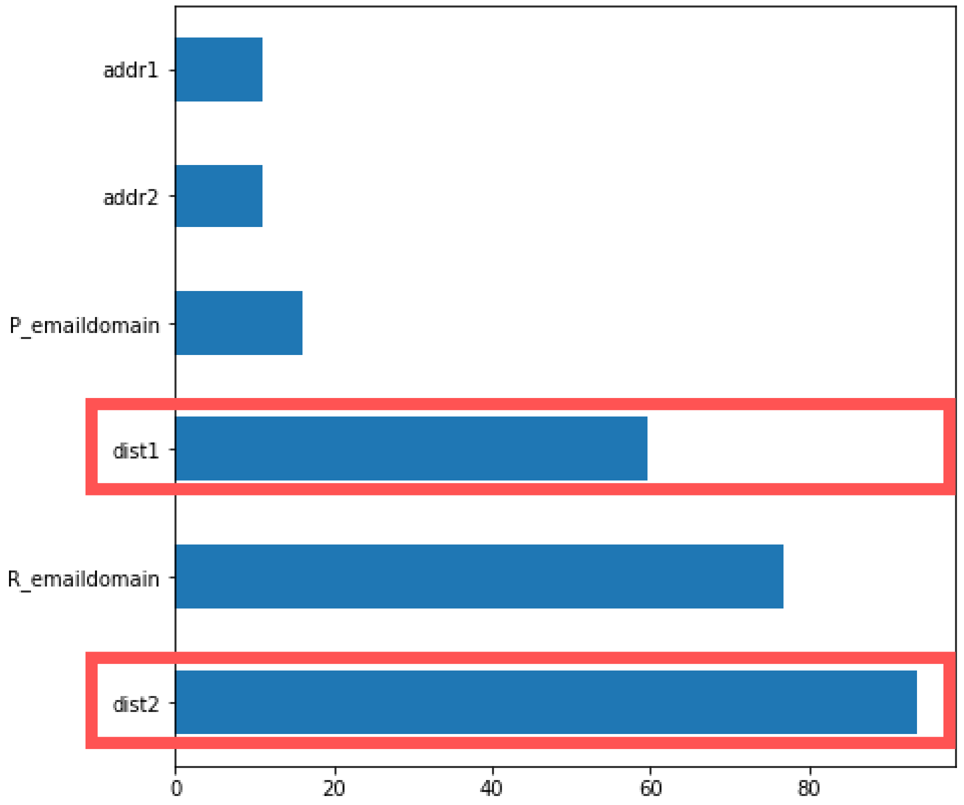
TransactionAmt ProductCD

card1 card2 card3 card4 card5 card6

Addr1 addr2 dist1 dist2

P_emaildomain R_emaildomain

C1-C14 D1-D15 M1-M19 Vxxx



categorical variable

범주형 변수

continuous variable

연속형 변수

연속형 변수지만 결측치가 너무 많아
Regression이 불가하기 때문에 DROP 하였다.

2.1 변수설명

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

P_emaildomain

구매자 이메일 도메인

categorical variable

범주형 변수

continuous variable

연속형 변수

2.1 변수설명

train_transaction.csv

TransactionID TransactionDT

TransactionAmt ProductCD

card1 card2 card3 card4 card5 card6

Addr1 addr2 dist1 dist2

P_emaildomain R_emaildomain

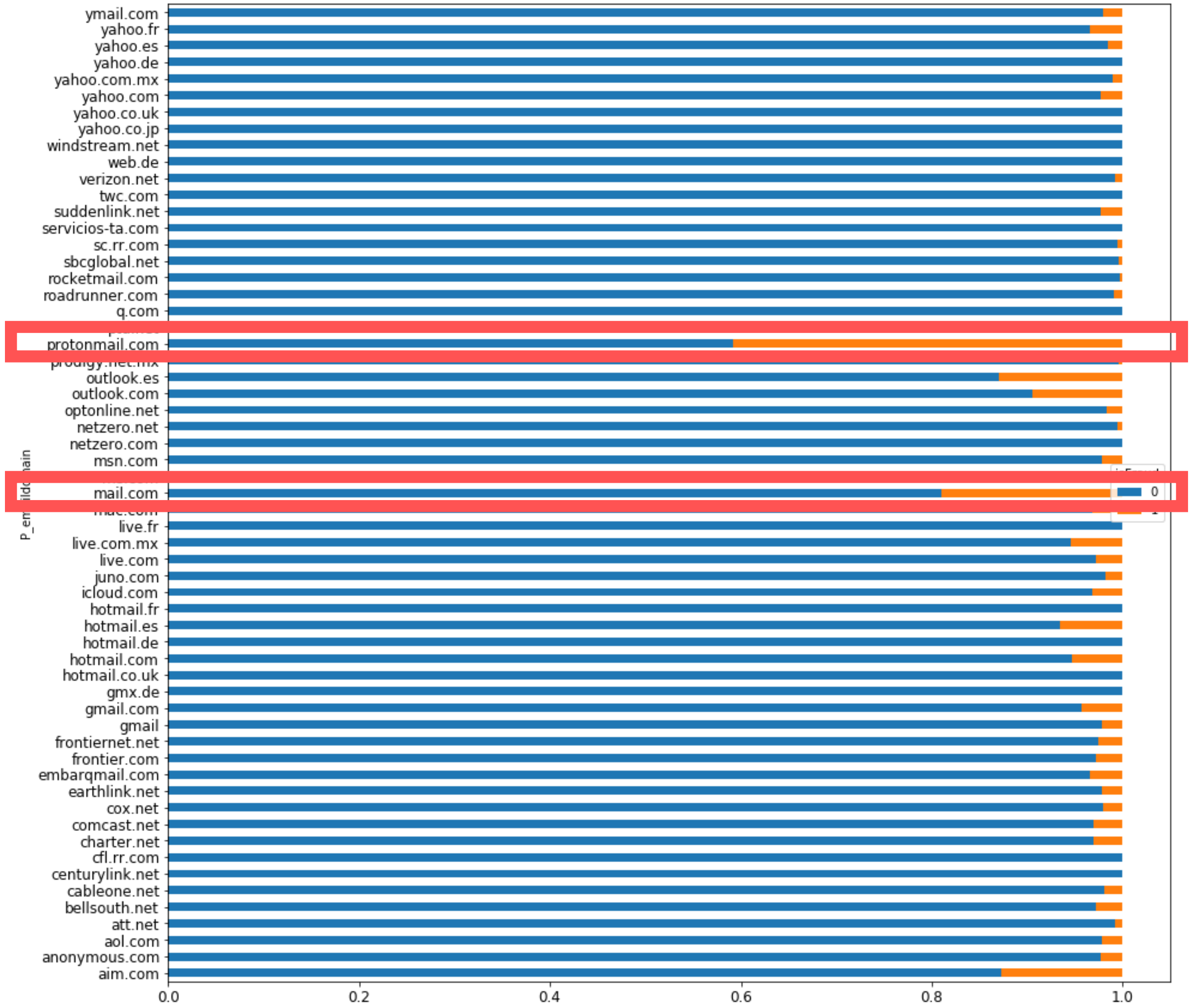
C1-C14 D1-D15 M1-M19 Vxxx

categorical variable

범주형 변수

continuous variable

연속형 변수



이메일에 따라 Fraud / Not Fraud 비율 차이가 크다.

2.1 변수설명

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

R_emaildomain

수신자 이메일 도메인

categorical variable

범주형 변수

continuous variable

연속형 변수

2.1 변수설명

train_transaction.csv

TransactionID TransactionDT

TransactionAmt ProductCD

card1 card2 card3 card4 card5 card6

Addr1 addr2 dist1 dist2

P_emaildomain R_emaildomain

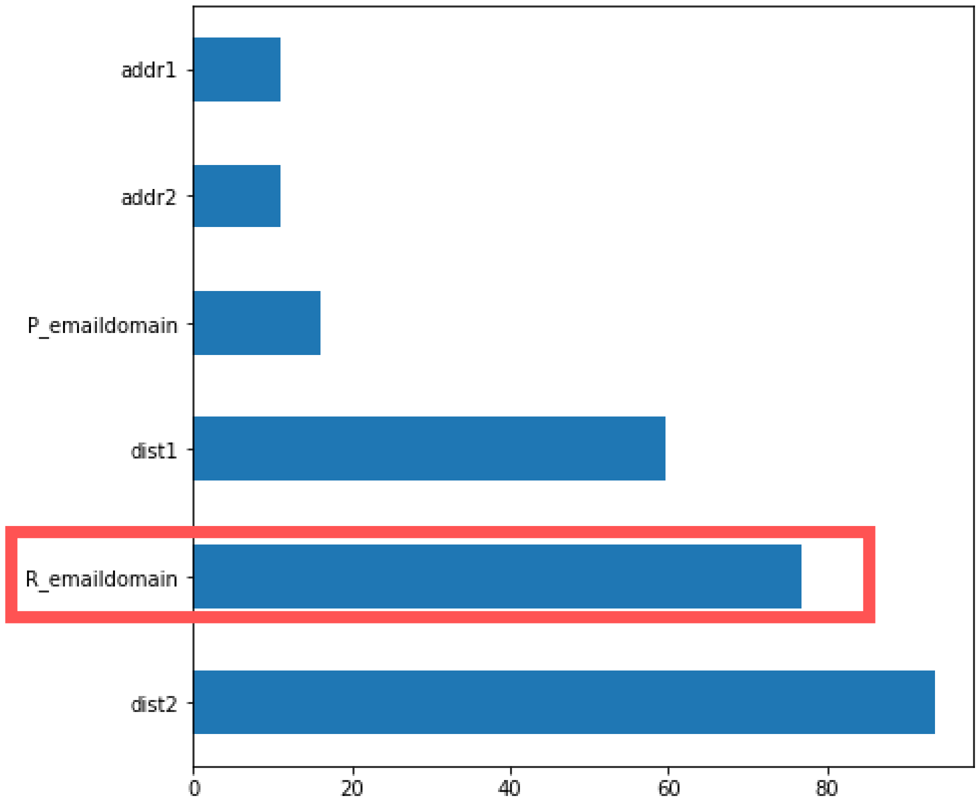
C1-C14 D1-D15 M1-M19 Vxxx

categorical variable

범주형 변수

continuous variable

연속형 변수



NA값이 76.75%로 높기 때문에 DROP 하였다.

2.1 변수설명

train_transaction.csv

TransactionID TransactionDT

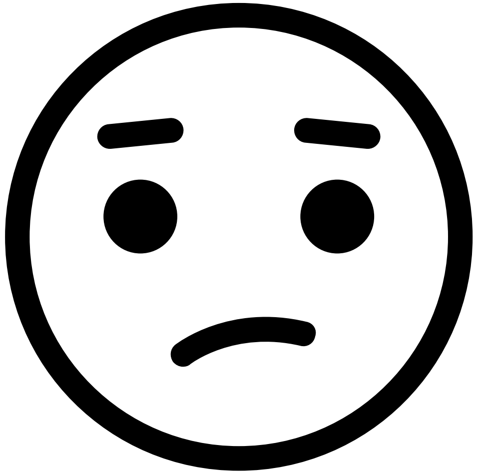
TransactionAmt ProductCD

card1 card2 card3 card4 card5 card6

Addr1 addr2 dist1 dist2

P_emaildomain R_emaildomain

C1-C14 D1-D15 M1-M19 Vxxx



categorical variable

범주형 변수

continuous variable

연속형 변수

의미를 모르는데 변수와 결측치가 너무 많다.

2.1 변수설명

```
C1       0.000000
C2       0.000000
C3       0.000000
C4       0.000000
C5       0.000000
C6       0.000000
C7       0.000000
C8       0.000000
C9       0.000000
C10      0.000000
C11      0.000000
C12      0.000000
C13      0.000000
           ...
V310     0.002032
V311     0.002032
V312     0.002032
V313     0.214888
V314     0.214888
V315     0.214888
V316     0.002032
V317     0.002032
V318     0.002032
V319     0.002032
V320     0.002032
V321     0.002032
V322    86.054967
V323    86.054967
V324    86.054967
V325    86.054967
V326    86.054967
V327    86.054967
V328    86.054967
V329    86.054967
V330    86.054967
V331    86.054967
V332    86.054967
V333    86.054967
V334    86.054967
V335    86.054967
V336    86.054967
V337    86.054967
V338    86.054967
V339    86.054967
Length: 394, dtype: float64
```

→

```python
  val = tr_trans[f'V{i}'].isna().sum()/len(tr_trans)*100
  if val not in unique_val_V:
    unique_col_V.append(f'V{i}')
    unique_val_V.append(val)

not_needed = []
print(new_tr.shape)
for i in range(1,340):
  if f'V{i}' not in unique_col_V:
    not_needed.append(f'V{i}')

new_tr = new_tr.drop(not_needed, axis = 1)
print(new_tr.shape)

unique_col_C = ['C1']
unique_val_C = []
unique_val_C.append(tr_trans['C1'].isna().sum()/len(tr_trans)*100)

for i in range(2,15):
  val = tr_trans[f'C{i}'].isna().sum()/len(tr_trans)*100
  if val not in unique_val_C:
    unique_col_C.append(f'C{i}')
    unique_val_C.append(val)

not_needed = []
print(new_tr.shape)
for i in range(1,15):
  if f'C{i}' not in unique_col_C:
    not_needed.append(f'C{i}')

new_tr = new_tr.drop(not_needed, axis = 1)
print(new_tr.shape)

unique_col_D = ['D1']
unique_val_D = []
unique_val_D.append(tr_trans['D1'].isna().sum()/len(tr_trans)*100)

for i in range(2,16):
  val = tr_trans[f'D{i}'].isna().sum()/len(tr_trans)*100
  if val not in unique_val_D:
    unique_col_D.append(f'D{i}')
    unique_val_D.append(val)

not_needed = []
print(new_tr.shape)
for i in range(1,16):
  if f'D{i}' not in unique_col_D:
    not_needed.append(f'D{i}')

new_tr = new_tr.drop(not_needed, axis = 1)
print(new_tr.shape)

unique_col_M = ['M1']
unique_val_M = []
unique_val_M.append(tr_trans['M1'].isna().sum()/len(tr_trans)*100)

for i in range(2,10):
  val = tr_trans[f'M{i}'].isna().sum()/len(tr_trans)*100
  if val not in unique_val_M:
    unique_col_M.append(f'M{i}')
    unique_val_M.append(val)
```

→

```
card3          0.000000
card4          0.000000
card5          0.000000
card6          0.000000
addr2          0.000000
C1             0.000000
D1             0.214888
D2            47.549192
D3            44.514851
D4            28.604667
D5            52.467403
D10           12.873302
D11           47.293494
D15           15.090087
M1            45.907136
M4            47.658753
M5            59.349409
M6            28.678836
M7            58.635317
M8            58.633115
V1            47.293494
V12           12.881939
V35           28.612626
V53           13.055170
V75           15.098723
V95            0.053172
V167          76.355370
V169          76.323534
V217          77.913435
V220          76.053104
V279           0.002032
V281           0.214888
exist          0.000000
DeviceInfo     0.000000
dt_month       0.000000
dt_hour        0.000000
dtype: float64
```

→

```
print(added.isna().sum()/len(ad

TransactionID      0.000000
isFraud            0.000000
TransactionDT      0.000000
TransactionAmt     0.000000
ProductCD          0.000000
card1              0.000000
card2              0.000000
card3              0.000000
card4              0.000000
card5              0.000000
card6              0.000000
addr2              0.000000
C1                 0.000000
D1                 0.214888
D10               12.873302
D15               15.090087
V12               12.881939
V53               13.055170
V75               15.098723
V95                0.053172
V279               0.002032
V281               0.214888
exist              0.000000
DeviceInfo         0.000000
dt_month           0.000000
dt_hour            0.000000
dtype: float64
```

동일 비율 결측치

변수 통합

변수 별 결측치 파악

변수 축약

2.1 변수설명

train_identity.csv

DeviceType

DeviceInfo

id1 - id38

categorical variable
범주형 변수

continuous variable
연속형 변수

DeviceType

네트워크 연결 정보(mobile, desktop)

DeviceInfo

네트워크 연결 정보(IOS, Windows)

id1-id38

그 외 네트워크 연결 정보(Browser, IP etc)

2.1 변수설명

train_identity.csv

DeviceType

DeviceInfo

id1 - id38

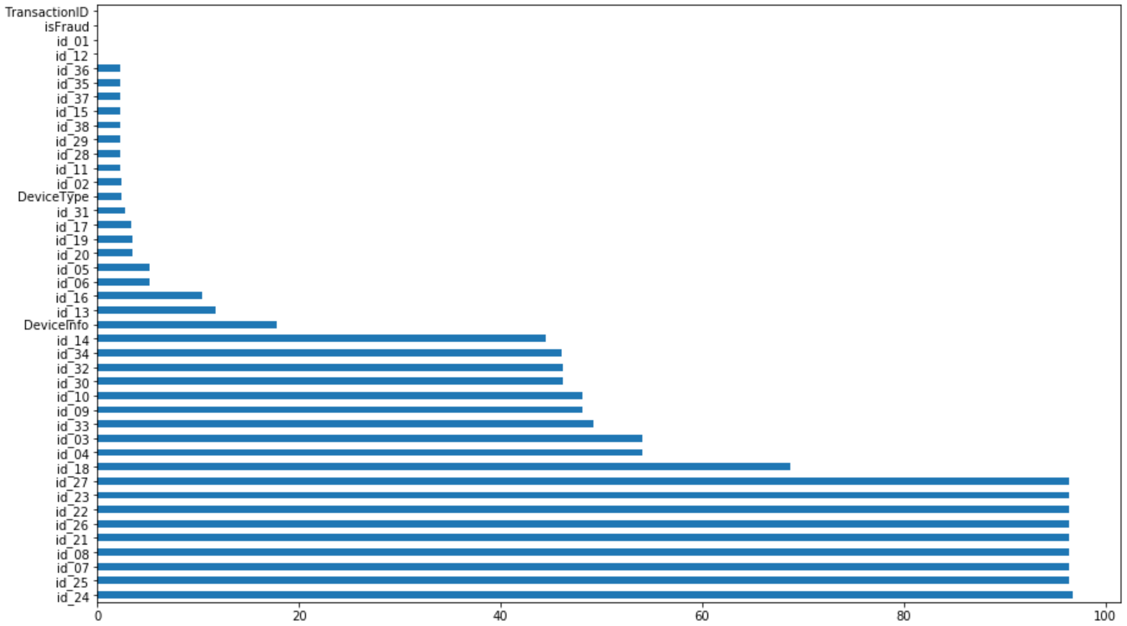
categorical variable

범주형 변수

continuous variable

연속형 변수

Id의 NA값 분포 그래프



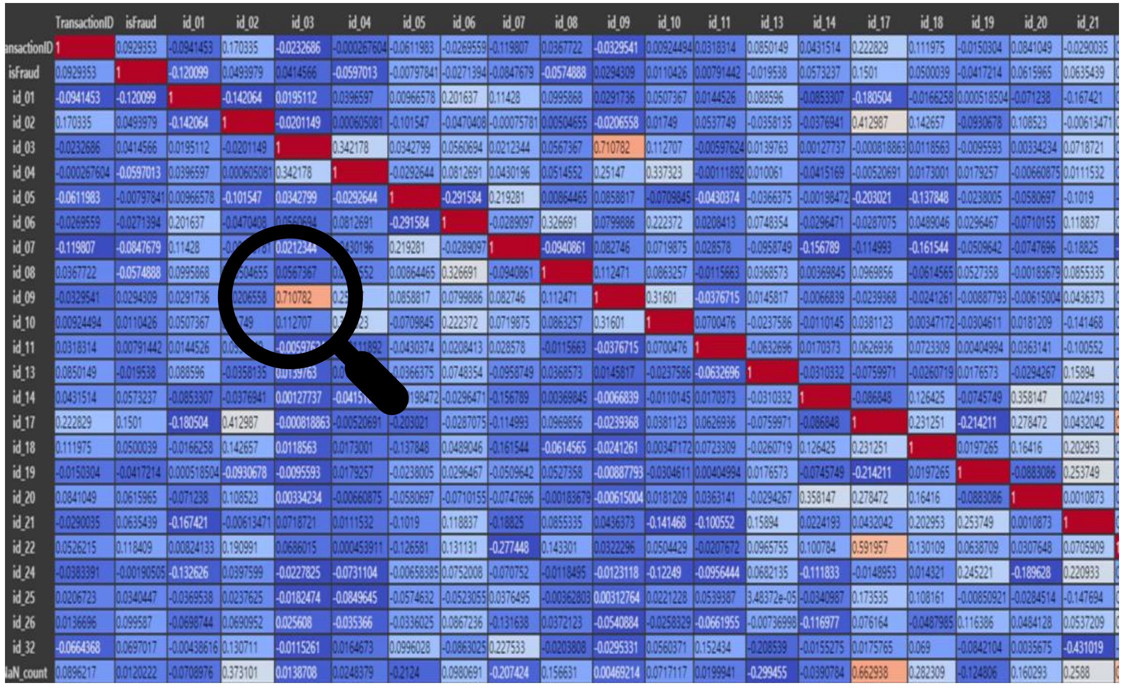
2.1 변수설명

train_identity.csv

DeviceType

DeviceInfo

id1 - id38



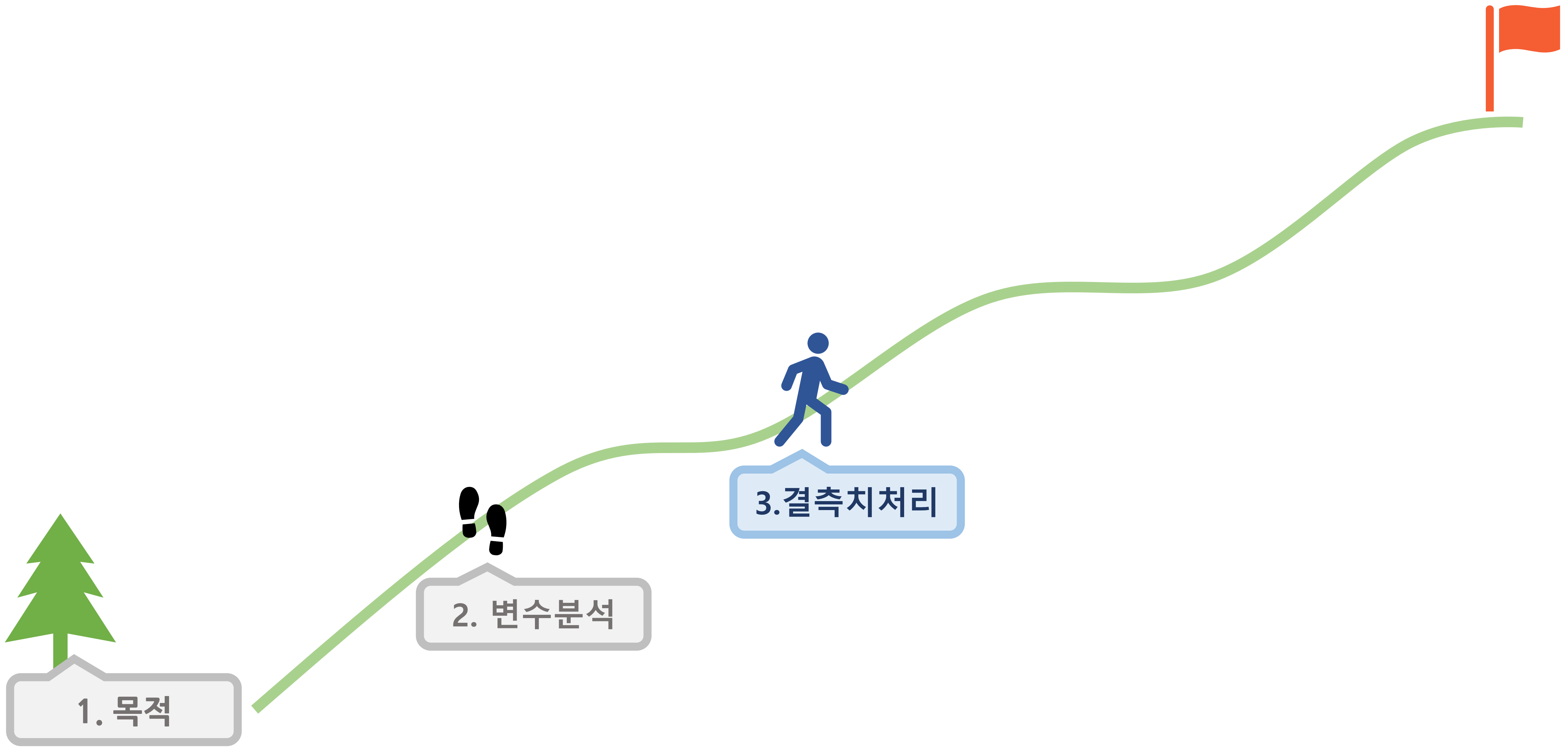
categorical variable

범주형 변수

continuous variable

연속형 변수

id3과 id9의 상관관계가 높다.


1. 목적
2. 변수분석
3.결측치처리

3.결측치 처리

- 3.1 방법소개
- 3.2 결측치비율
- 3.3 결측치처리

결측값처리

고려사항

- 변수간 관계를 파악
- 왜곡을 적게 만들어 모델의 정확성 높임
- 결측값이 무작위인가? 관계가 있는가?

A. 삭제
결측치가 생긴부분을 삭제

장점 : 빠르고 간편
단점 : 모델의 유효성이 낮아짐 무작위발생이 아닐 시 왜곡된 모델 생성

B. 대체 (평균, 최빈값, 중간값)
일괄 대체 / 범주형 변수로 유사유형 평균값으로 대체

장점 : 빠르고 간편
단점 : 모델의 유효성이 낮아짐, 유사유형 선택 시 왜곡된 모델 생성

C. 예측값 삽입
결측값이 없는 관측치로 예측 모델 생성 (Regression, Logistic regression 등)

장점 : 자의적인 판단이 적음
단점 : 다양한 변수에서 결측치가 발생하거나 결측치가 많은 경우 사용불가

3.2 결측치비율

	전체	결측치 수	결측치 비율(%)
TransactionID	590541	0	0
isFraud	590541	0	0
TransactionDT	590541	0	0
TransactionAmt	590541	0	0
ProductCD	590541	0	0
card1	590541	0	0
card2	590541	8933	1.51
card3	590541	1565	0.26
card4	590541	1577	0.27
card5	590541	4259	0.72
card6	590541	1571	0.27
addr1	590541	65706	11.1
addr2	590541	65706	11.1
dist1	590541	352271	59.65
dist2	590541	552913	93.62
P_emaildomain	590541	94456	15.99
R_emaildomain	590541	453249	76.75

train_transaction.csv

TransactionDT TransactionAMT

ProductCD card1 ~ card6

addr dist P_R_emaildomain

C1-C14 D1-D15 M1-M19 Vxxx

3.2 결측치비율

dist1

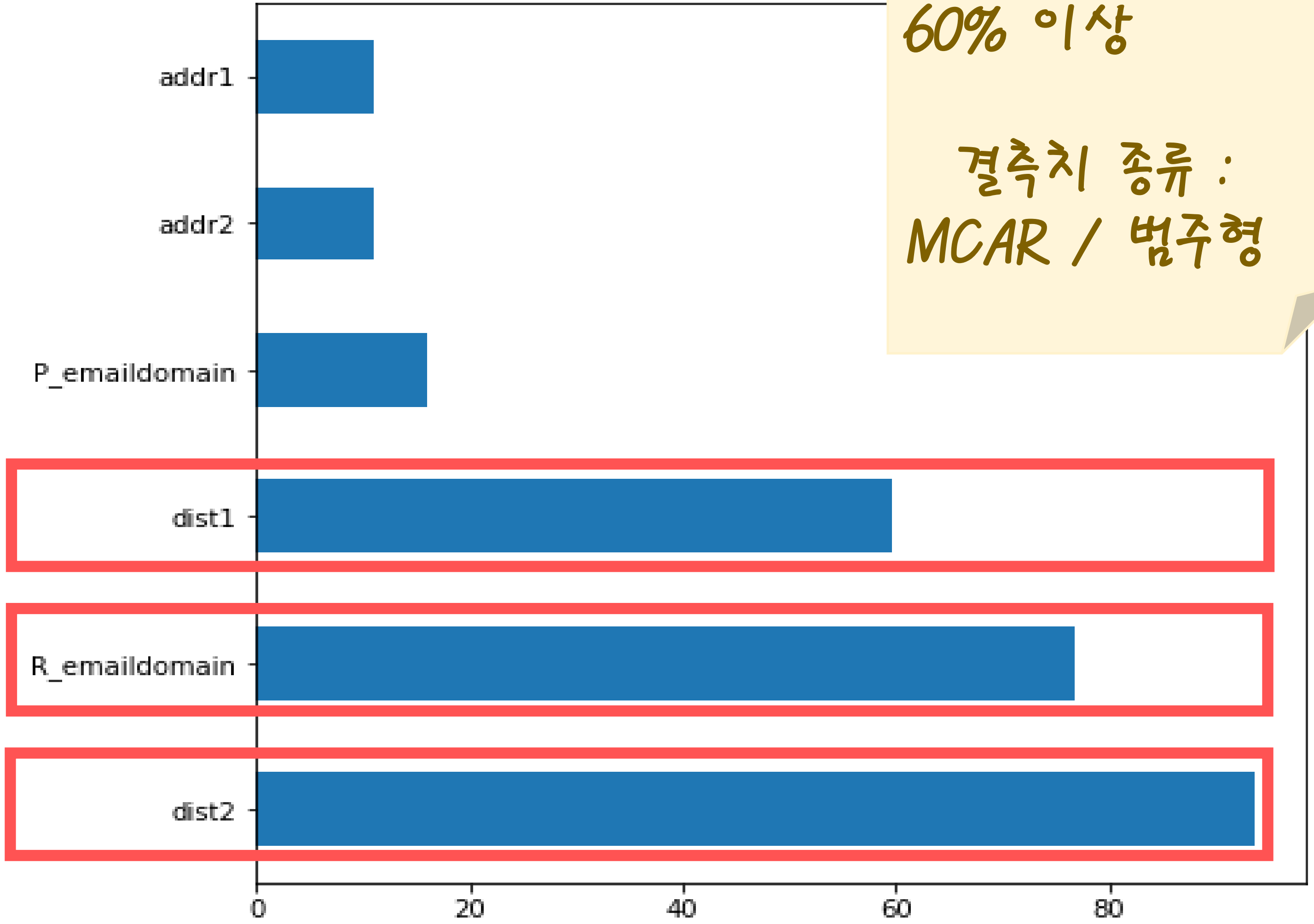
결측치 59.652%

dist2

결측치 93.628%

R_emaildomain

결측치 76.751%



결측치 비율 :
60% 이상

결측치 종류 :
MCAR / 범주형

addr1
도시

?

addr2
국가

?

addr1 & addr2

결측치가 완벽하게
일치하므로 하나로 통합

3.2 결측치비율

1	Transac	isFraud	Fraud	Transac	Transac	Product	card1	card2	card3	card4	card5	card6	addr1	addr2	dist1	dist2	P_email	R_e
65	2987063	0	notFraud	87604	80	W	17967	285	150	visa	226	debit	184	87			yahoo.com	
66	2987064	0	notFraud	87611	250	W	3278	453	150	visa	226	debit	122	87			gmail.com	
67	2987065	0	notFraud	87650	114.95	W	17188	321	150	visa	226	debit	299	87	1			
68	2987066	0	notFraud	87660	300	H	15333	562	150	visa	226	credit	315	87			anonymous.co	
69	2987067	0	notFraud	87664	445	W	1240	302	150	visa	195	credit	485	87	24			
70	2987068	0	notFraud	87667	3.081	C	14076	545	185	visa	147	credit					hotmail.cc	hot
71	2987069	0	notFraud	87725	20	S	12866	303	150	visa	226	debit	330	87		84		hot
72	2987070	0	notFraud	87735	100	H	3682	264	150	visa	162	credit	325	87			anonymous.co	
73	2987071	0	notFraud	87736	59	W	1662	111	150	visa	195	debit	472	87			gmail.com	
74	2987072	0	notFraud	87752	6.767	C	13832	375	185	mastercard	224	debit					outlook.cc	outl
75	2987073	0	notFraud	87759	554	W	1955	383	150	visa	226	debit	315	87	5		yahoo.com	
76	2987074	0	notFraud	87775	27.793	C	15885	545	185	visa	138	debit				100	gmail.com	gma
77	2987075	0	notFraud	87779	68.5	W	4806	490	150	visa	226	debit	315	87			gmail.com	
78	2987076	0	notFraud	87787	36.95	W	18132	567	150	mastercard	117	debit	441	87	1			
79	2987077	0	notFraud	87793	280	W	3278	453	150	visa	226	debit	122	87			gmail.com	
80	2987078	0	notFraud	87825	300	W	14858	558	150	visa	226	debit	220	87	2239			
81	2987079	0	notFraud	87839	28.699	C	4504	500	185	mastercard	219	credit					hotmail.cc	hot
82	2987080	0	notFraud	87856	60	W	7207	111	150	visa	226	debit	204	87			comcast.net	
83	2987081	0	notFraud	87868	104.95	W	17188	321	150	visa	226	debit	299	87	1		gmail.com	
84	2987082	0	notFraud	87899	280	W	15066	170	150	mastercard	102	credit	325	87			optonline.net	
85	2987083	0	notFraud	87924	411.95	W	14695	396	150	mastercard	224	credit	315	87	22		gmail.com	
86	2987084	0	notFraud	87928	125.674	C	5583	103	185	visa	226	credit				744	anonymou	ano
87	2987085	0	notFraud	87935	42.596	C	15885	545	185	visa	138	debit					anonymou	ano
88	2987086	0	notFraud	87950	44.5	W	11815	206	150	mastercard	166	debit	476	87	6		gmail.com	
89	2987087	0	notFraud	88004	88.95	W	1549	143	150	visa	226	debit	205	87	9		yahoo.com	
90	2987088	0	notFraud	88021	140	W	18095	243	150	visa	226	debit	387	87			gmail.com	
91	2987089	0	notFraud	88042	318.95	W	17188	321	150	visa	226	debit	299	87	5		gmail.com	
92	2987090	0	notFraud	88046	77.821	C	15885	545	185	visa	138	debit					gmail.com	gma
93	2987091	0	notFraud	88053	107	W	14165	111	150	mastercard	224	debit	181	87	10			
94	2987092	0	notFraud	88054	117	W	6481	111	150	visa	226	debit	337	87	327		cox.net	
95	2987093	0	notFraud	88070	50	H	5220	360	150	visa	226	credit	231	87			charter.ne	cha
96	2987094	0	notFraud	88107	527	W	5409	170	150	visa	226	credit	315	87	5		yahoo.com	
97	2987095	0	notFraud	88120	59	W	2538	476	150	visa	166	debit	330	87			anonymous.co	

3.2 결측치비율

	전체	결측치 수	결측치 비율(%)
TransactionID	590541	0	0
isFraud	590541	0	0
TransactionDT	590541	0	0
TransactionAmt	590541	0	0
ProductCD	590541	0	0
card1	590541	0	0
card2	590541	8933	1.51
card3	590541	1565	0.26
card4	590541	1577	0.27
card5	590541	4259	0.72
card6	590541	1571	0.27
addr1	590541	65706	11.1
addr2	590541	65706	11.1
dist1	590541	352271	59.65
dist2	590541	552913	93.62
P_emaildomain	590541	94456	15.99
R_emaildomain	590541	453249	76.75

	전체	결측치 수	결측치 비율(%)
TransactionID	590541	0	0
isFraud	590541	0	0
TransactionDT	590541	0	0
TransactionAmt	590541	0	0
ProductCD	590541	0	0
card1	590541	0	0
card2	590541	8933	1.51
card3	590541	1565	1.51
card4	590541	1577	0.26
card5	590541	4259	0.27
card6	590541	1571	0.72
addr2	590541	65706	11.1
P_emaildomain	590541	94456	15.99

3.3 결측치처리

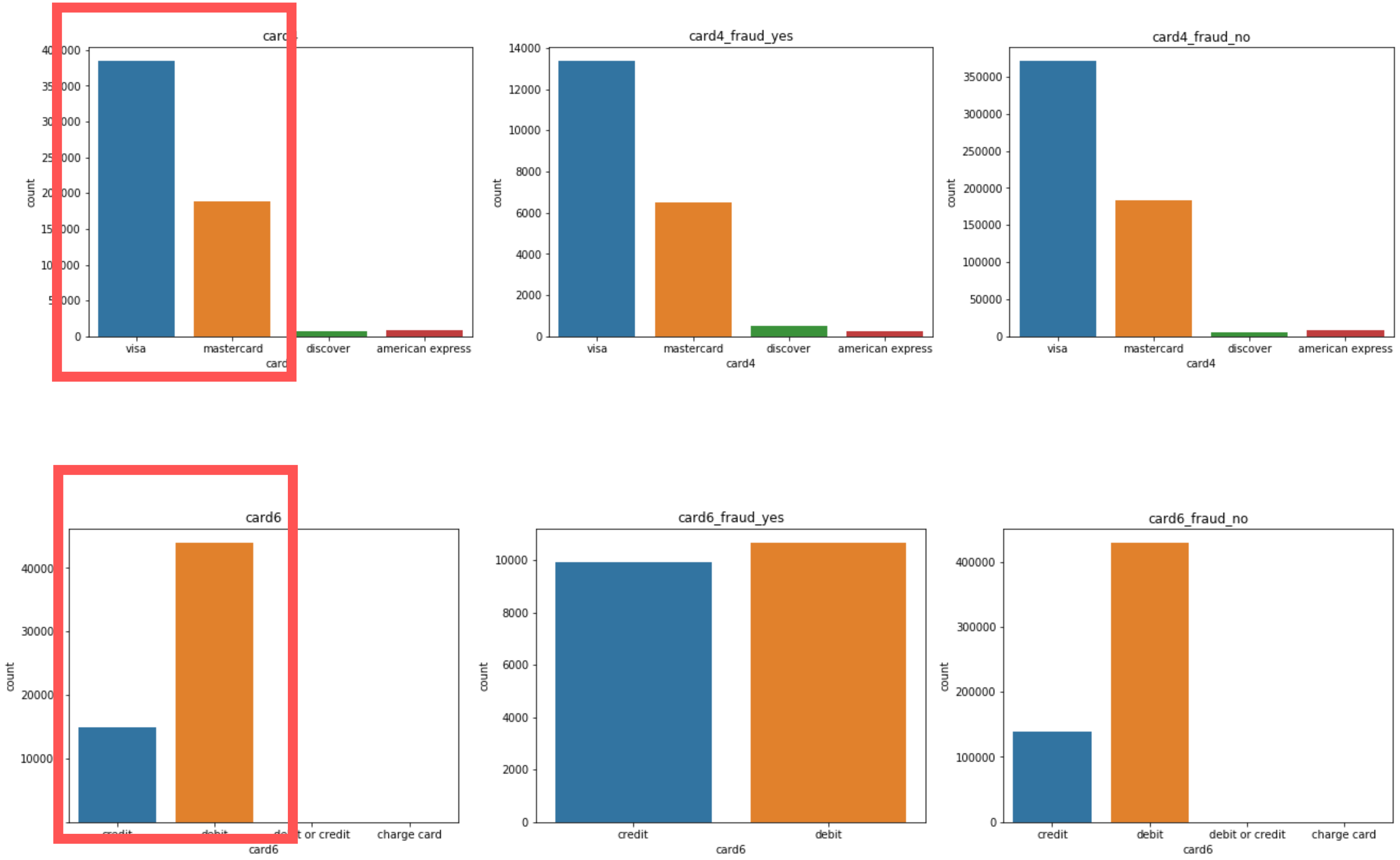
Card4

Card6

최빈값으로 대체

결측치 비율이 낮으며, 범주가 많지 않아 최빈값 대체가능

3.3 결측치처리


card4
card4_fraud_yes
card4_fraud_no
count
visa
mastercard
discover
american express
card4
card6
card6_fraud_yes
card6_fraud_no
credit
debit
debit or credit
charge card
card6

3.3 결측치처리

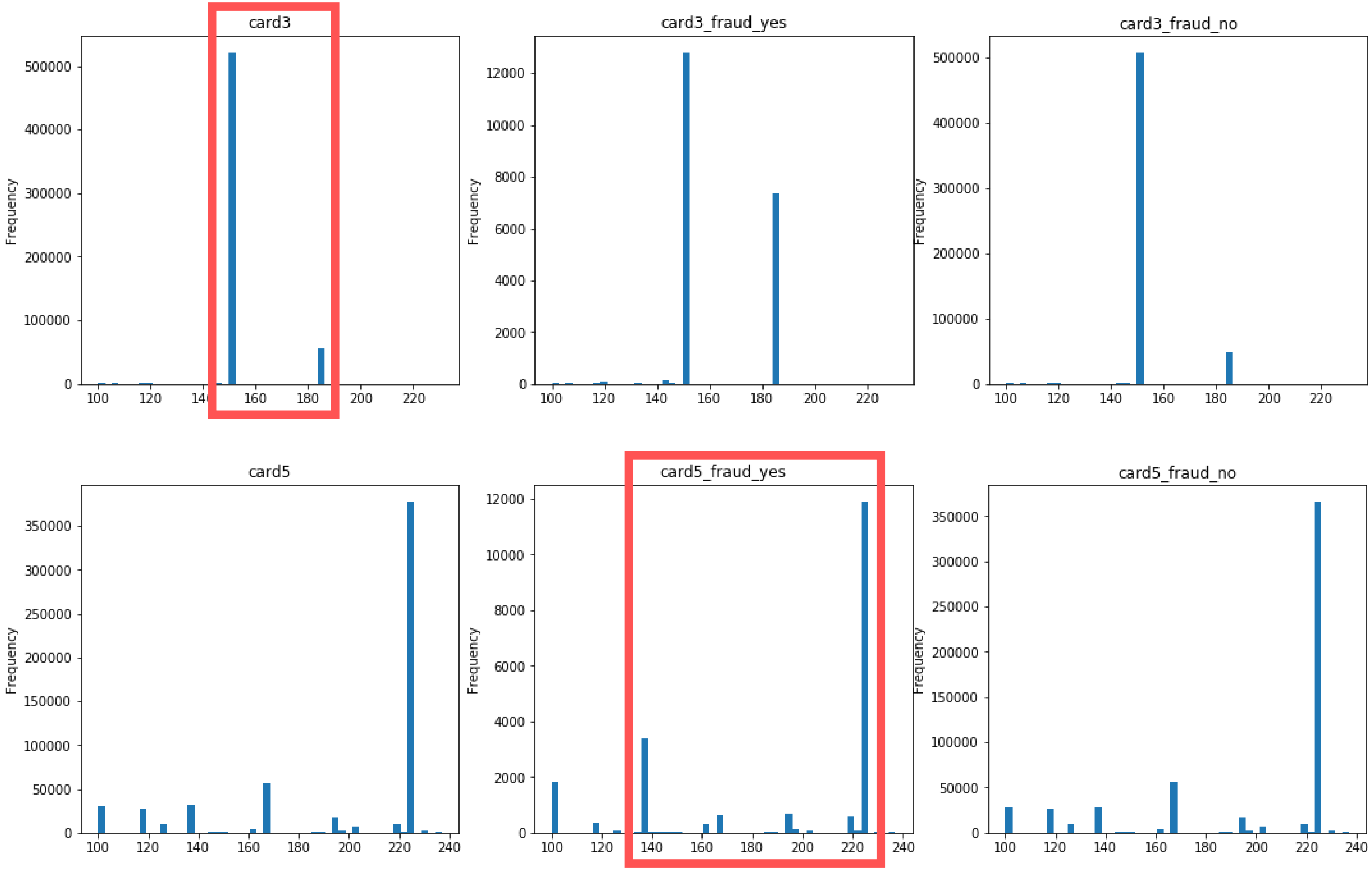
Card3

Card5

최빈값으로 대체

그래프에서 적은 수의 변수만 유의미한 수치로 나옴

3.3 결측치처리


card3
card3_fraud_yes
card3_fraud_no
card5
card5_fraud_yes
card5_fraud_no
Frequency

3.3 결측치처리

DeviceInfo

Unknown 처리

현실적으로 기기의 기종을 찾기 불가 (MCAR)

특히, 기기의 기종, OS이름, 브라우저 엔진 등

여러 다른 변수가 섞여있는 Column

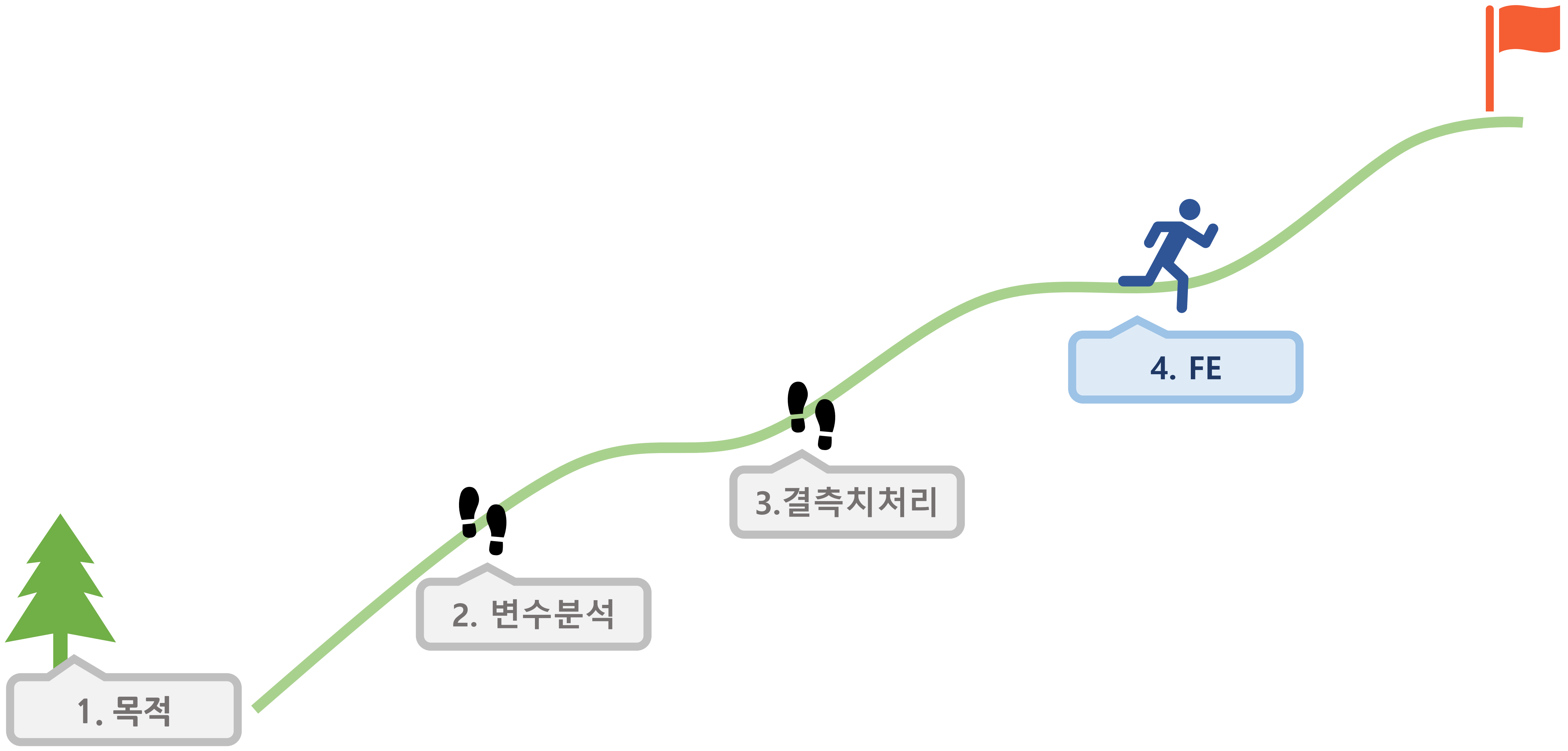
Outlier 처리?

이상치(Outlier)의 경우

Fraud Detection에서는

주요 요인이 될 수 있음

이상치 처리에 영향을
크게 받지 않는
Decision Tree기반 모델 이용 예정


1. 목적
2. 변수분석
3.결측치처리
4. FE

4. FE

4.1 FE소개

4.2 FE결과

Feature Engineering

기존 변수 사용 -> 데이터 추가

관측치 / 변수 추가 없이 기존 데이터 강화

- Scaling ->log / root
- Binning -> 연속형을 범주형으로
- Transform -> 새로운 변수 추가 및 결합
- Dummy -> One-Hot Encoding

From Ybigta 세션자료

The machine learning models affected by the magnitude of the feature are:

- Linear and Logistic Regression
- Neural Networks
- Support Vector Machines
- KNN
- K-means clustering
- Linear Discriminant Analysis (LDA)
- Principal Component Analysis (PCA)

Machine learning models insensitive to feature magnitude are the ones based on Trees:

- Classification and Regression Trees
- Random Forests
- Gradient Boosted Trees

Gradient Boost계열 모델 이용 예정

Magnitude feature 재조정 필요 없음

Feature Engineering

기존

관측치 / 변수 추가 없이 기존 데이터 강화

- Scaling ->
- Binning -> 연속형을 범주형으로
- Transform -> 새로운 변수 추가 및 결합
- Dummy -> One-Hot Encoding

From Ybigta 세션자료

The machine learning models affected by the magnitude of the feature are:

- *…d Logistic Regression*
- *…etworks*
- *…ector Machines*
- *KNN*
- *K-means clustering*
- *Linear Discriminant Analysis (LDA)*
- *Principal Component Analysis (PCA)*

Machine learning models insensitive to feature magnitude are the ones based on Trees:

- *Classification and Regression Trees*
- *Random Forests*
- *Gradient Boosted Trees*

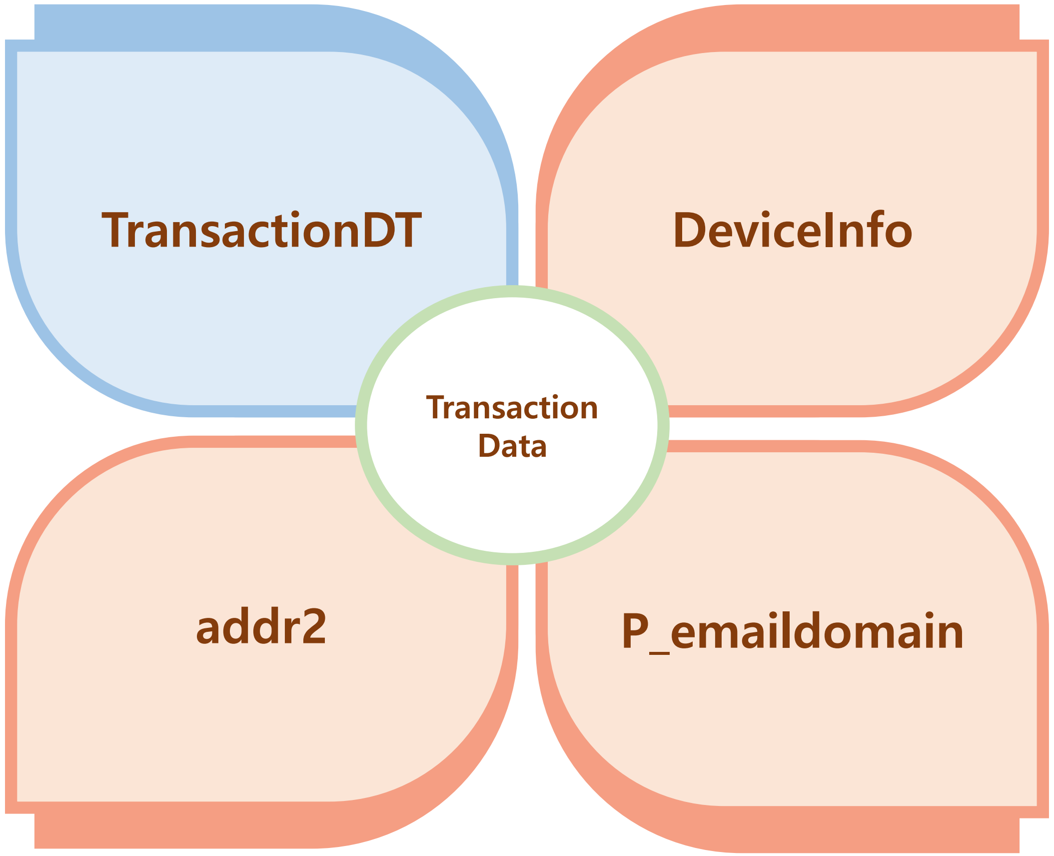
4.2 FE결과

YYYY/MM/DD/HH

변환

Cramer's V

이용하여 merge



P_emaildomain과 merge

addr2와 merge

4.2 FE결과

TransactionDT

```
ain_tr_time["train_tr_year"] = train_tr_time["TransactionDT_cleaned"].astype(str).str[:4]
ain_tr_time["train_tr_month"] = train_tr_time["TransactionDT_cleaned"].astype(str).str[5:7]
ain_tr_time["train_tr_day"] = train_tr_time["TransactionDT_cleaned"].astype(str).str[8:10]
ain_tr_time["train_tr_hour"] = train_tr_time["TransactionDT_cleaned"].astype(str).str[11:13]
ain_tr_time["train_tr_minute"] = train_tr_time["TransactionDT_cleaned"].astype(str).str[14:16]
ain_tr_time["train_tr_second"] = train_tr_time["TransactionDT_cleaned"].astype(str).str[17:]

ain_tr_time["train_tr_dayofweek"] = train_tr_time["TransactionDT_cleaned"].dt.dayofweek

int(train_tr_time.shape)

ain_tr_time[["train_tr_year", "train_tr_month", "train_tr_day", "train_tr_hour", "train_tr_minute", "train_tr_second", "train_tr_dayofweek"]].head()
```

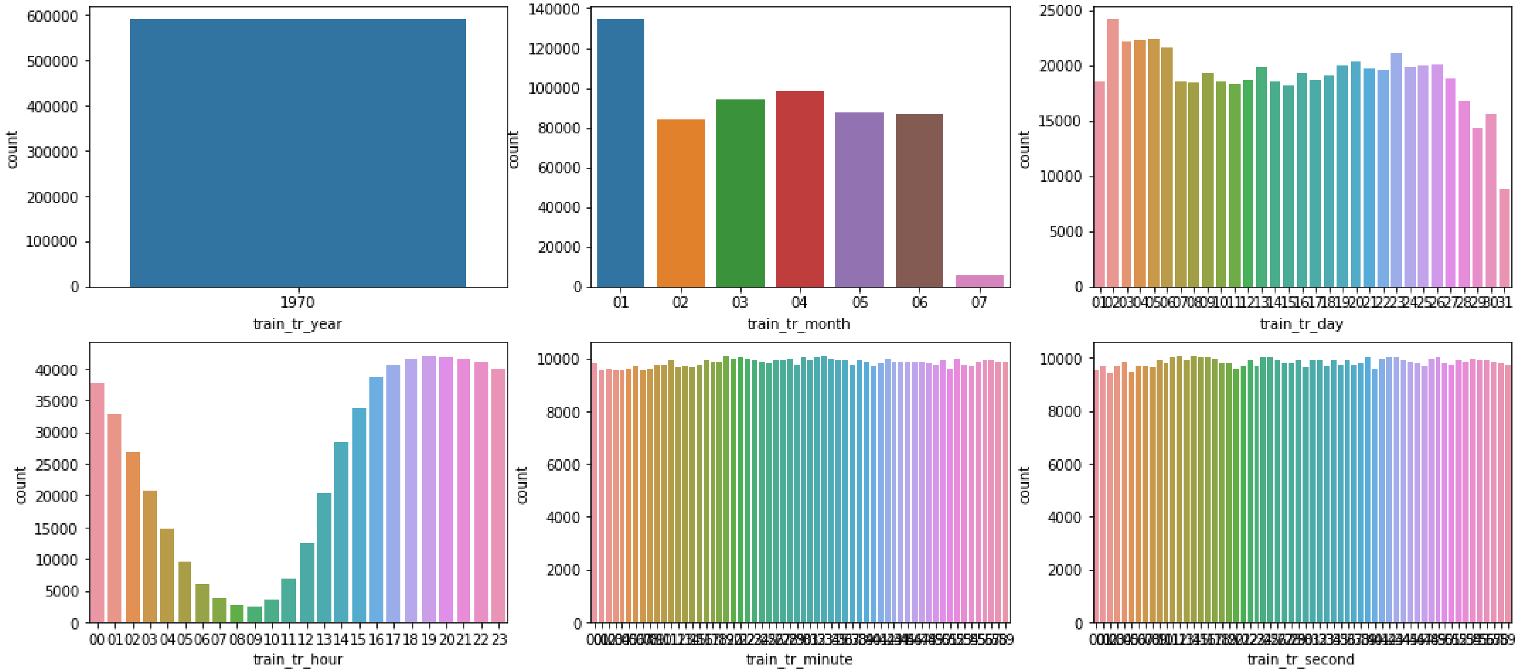
0540, 402)

train_tr_year	train_tr_month	train_tr_day	train_tr_hour	train_tr_minute	train_tr_second	train_tr_dayofweek
1970	01	02	00	00	00	4
1970	01	02	00	00	01	4
1970	01	02	00	01	09	4
1970	01	02	00	01	39	4
1970	01	02	00	01	46	4

```
ain_tr_time.loc[train_tr_time["train_tr_dayofweek"] == 0, "train_tr_dayofweek(humanized)"] = "Monday"
ain_tr_time.loc[train_tr_time["train_tr_dayofweek"] == 1, "train_tr_dayofweek(humanized)"] = "Tuesday"
ain_tr_time.loc[train_tr_time["train_tr_dayofweek"] == 2, "train_tr_dayofweek(humanized)"] = "Wednesday"
ain_tr_time.loc[train_tr_time["train_tr_dayofweek"] == 3, "train_tr_dayofweek(humanized)"] = "Thursday"
ain_tr_time.loc[train_tr_time["train_tr_dayofweek"] == 4, "train_tr_dayofweek(humanized)"] = "Friday"
ain_tr_time.loc[train_tr_time["train_tr_dayofweek"] == 5, "train_tr_dayofweek(humanized)"] = "Saturday"
ain_tr_time.loc[train_tr_time["train_tr_dayofweek"] == 6, "train_tr_dayofweek(humanized)"] = "Sunday"

int(train_tr_time.shape)

ain_tr_time[["TransactionDT_cleaned", "train_tr_dayofweek", "train_tr_dayofweek(humanized)"]].head()
```

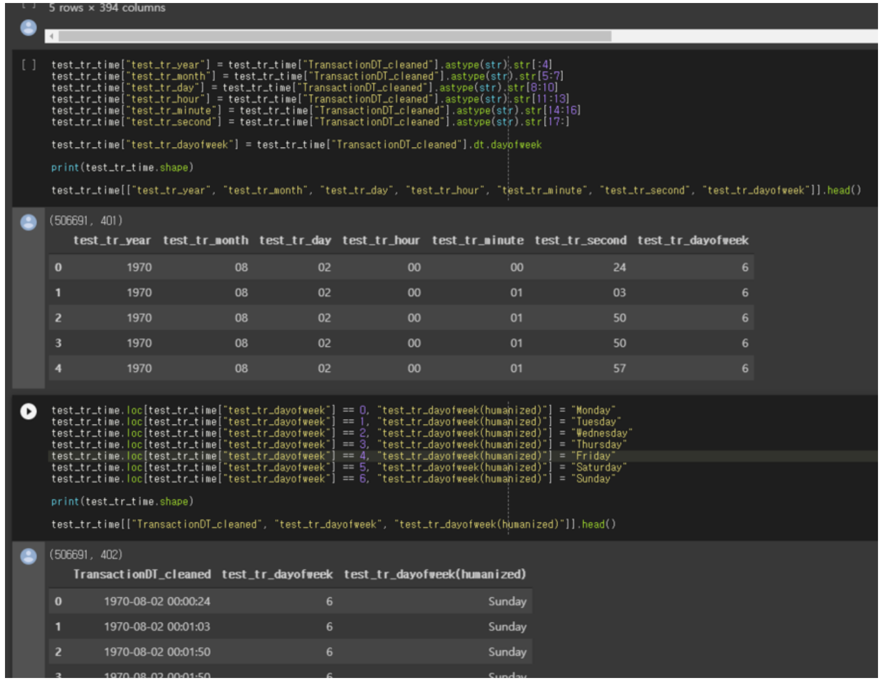
0540, 403)

TransactionDT_cleaned	train_tr_dayofweek	train_tr_dayofweek(humanized)
1970-01-02 00:00:00	4	Friday
1970-01-02 00:00:01	4	Friday
1970-01-02 00:01:09	4	Friday
1970-01-02 00:01:39	4	Friday
1970-01-02 00:01:46	4	Friday



4.2 FE결과

TransactionDT



5 rows × 394 columns

```
test_tr_time["test_tr_year"] = test_tr_time["TransactionDT_cleaned"].astype(str).str[:4]
test_tr_time["test_tr_month"] = test_tr_time["TransactionDT_cleaned"].astype(str).str[5:7]
test_tr_time["test_tr_day"] = test_tr_time["TransactionDT_cleaned"].astype(str).str[8:10]
test_tr_time["test_tr_hour"] = test_tr_time["TransactionDT_cleaned"].astype(str).str[11:13]
test_tr_time["test_tr_minute"] = test_tr_time["TransactionDT_cleaned"].astype(str).str[14:16]
test_tr_time["test_tr_second"] = test_tr_time["TransactionDT_cleaned"].astype(str).str[17:]

test_tr_time["test_tr_dayofweek"] = test_tr_time["TransactionDT_cleaned"].dt.dayofweek

print(test_tr_time.shape)

test_tr_time[["test_tr_year", "test_tr_month", "test_tr_day", "test_tr_hour", "test_tr_minute", "test_tr_second", "test_tr_dayofweek"]].head()
```

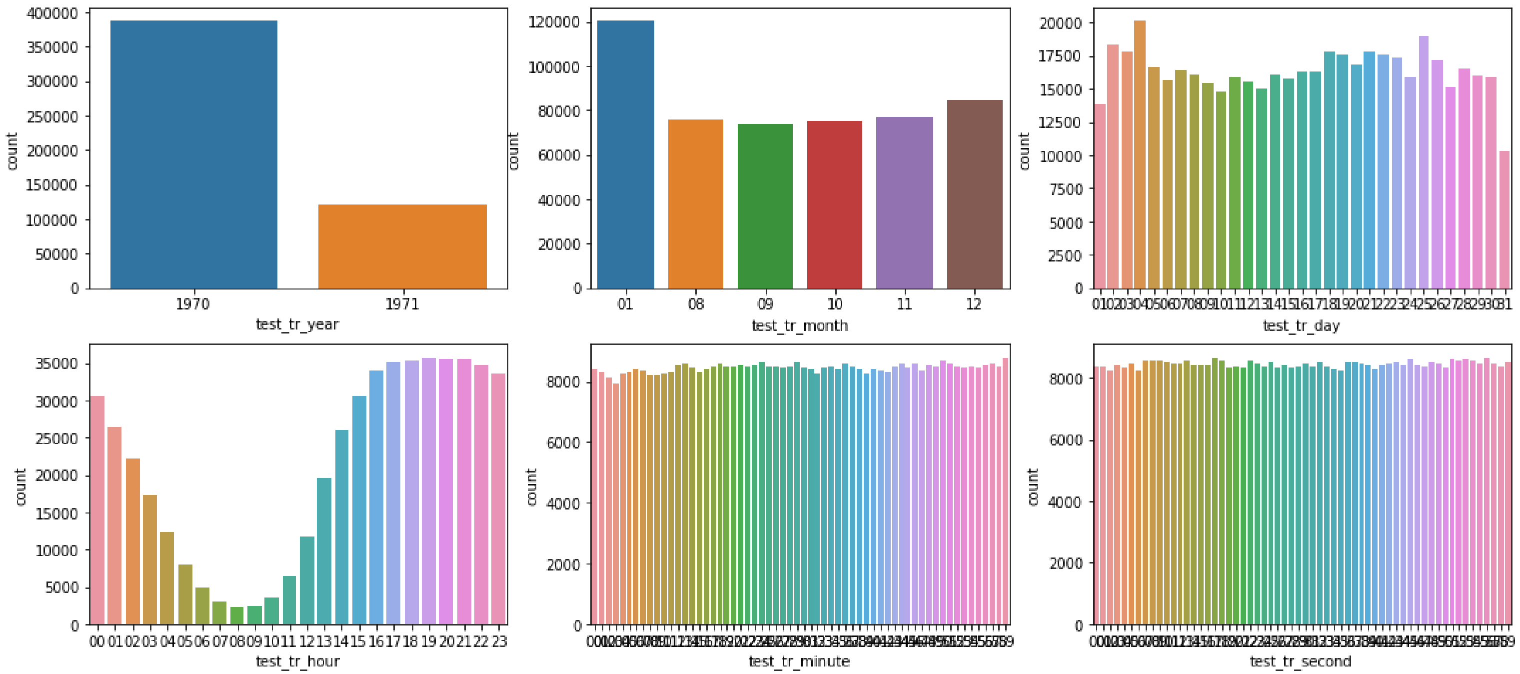
(506691, 401)

	test_tr_year	test_tr_month	test_tr_day	test_tr_hour	test_tr_minute	test_tr_second	test_tr_dayofweek
0	1970	08	02	00	00	24	6
1	1970	08	02	00	01	03	6
2	1970	08	02	00	01	50	6
3	1970	08	02	00	01	50	6
4	1970	08	02	00	01	57	6

```
test_tr_time.loc[test_tr_time["test_tr_dayofweek"] == 0, "test_tr_dayofweek(humanized)"] = "Monday"
test_tr_time.loc[test_tr_time["test_tr_dayofweek"] == 1, "test_tr_dayofweek(humanized)"] = "Tuesday"
test_tr_time.loc[test_tr_time["test_tr_dayofweek"] == 2, "test_tr_dayofweek(humanized)"] = "Wednesday"
test_tr_time.loc[test_tr_time["test_tr_dayofweek"] == 3, "test_tr_dayofweek(humanized)"] = "Thursday"
test_tr_time.loc[test_tr_time["test_tr_dayofweek"] == 4, "test_tr_dayofweek(humanized)"] = "Friday"
test_tr_time.loc[test_tr_time["test_tr_dayofweek"] == 5, "test_tr_dayofweek(humanized)"] = "Saturday"
test_tr_time.loc[test_tr_time["test_tr_dayofweek"] == 6, "test_tr_dayofweek(humanized)"] = "Sunday"

print(test_tr_time.shape)

test_tr_time[["TransactionDT_cleaned", "test_tr_dayofweek", "test_tr_dayofweek(humanized)"]].head()
```

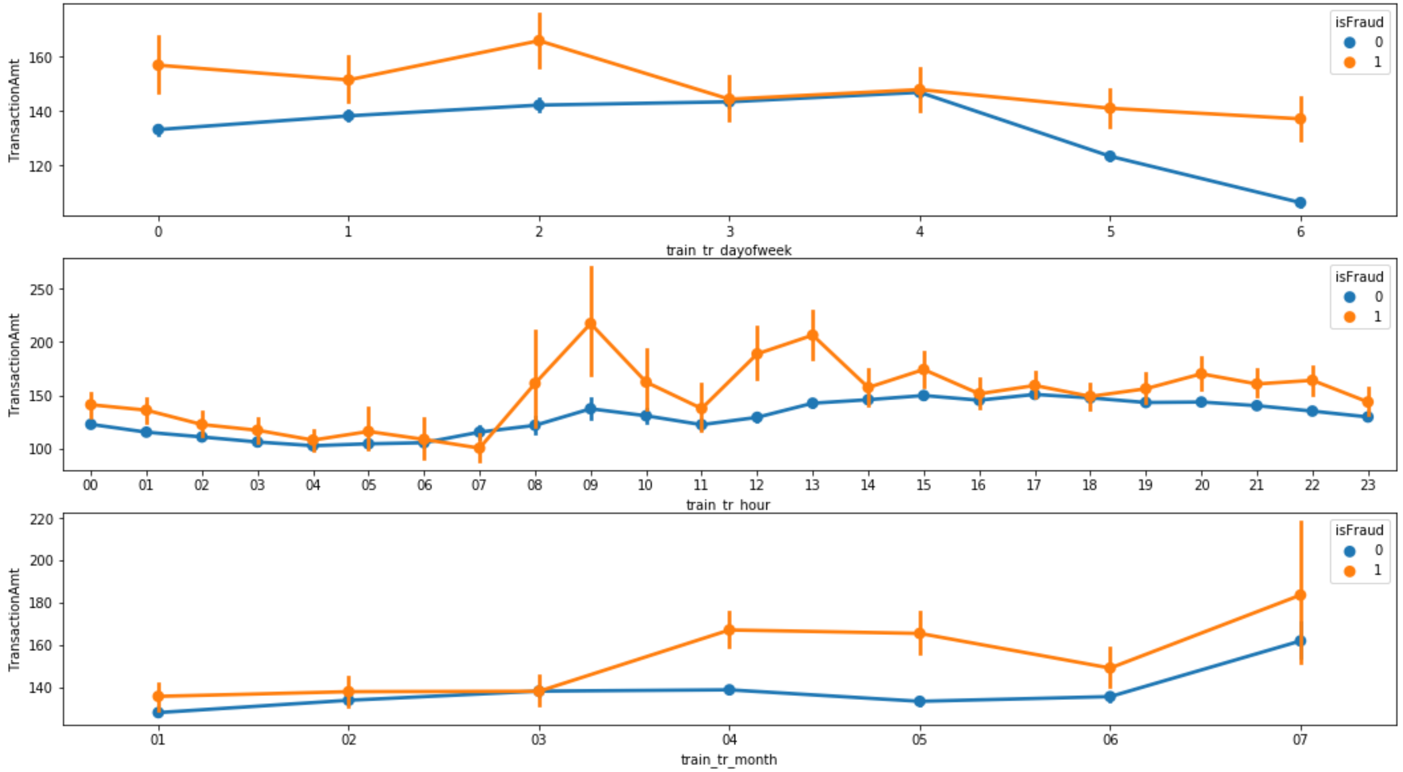
(506691, 402)

	TransactionDT_cleaned	test_tr_dayofweek	test_tr_dayofweek(humanized)
0	1970-08-02 00:00:24	6	Sunday
1	1970-08-02 00:01:03	6	Sunday
2	1970-08-02 00:01:50	6	Sunday



4.2 FE결과

TransactionDT



변형한 TransactionDT 변수들 중

유의미한 fraud차이가 있는 변수는

day of week, hour, month

4.2 FE결과

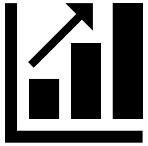
addr2 & P_emaildomain

Name	TRUE	FALSE	Overall	Percent	Rank	Service info	Region	Country_cod
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

P_Emaildomain에서

얻어낸

확실한 국가데이터값



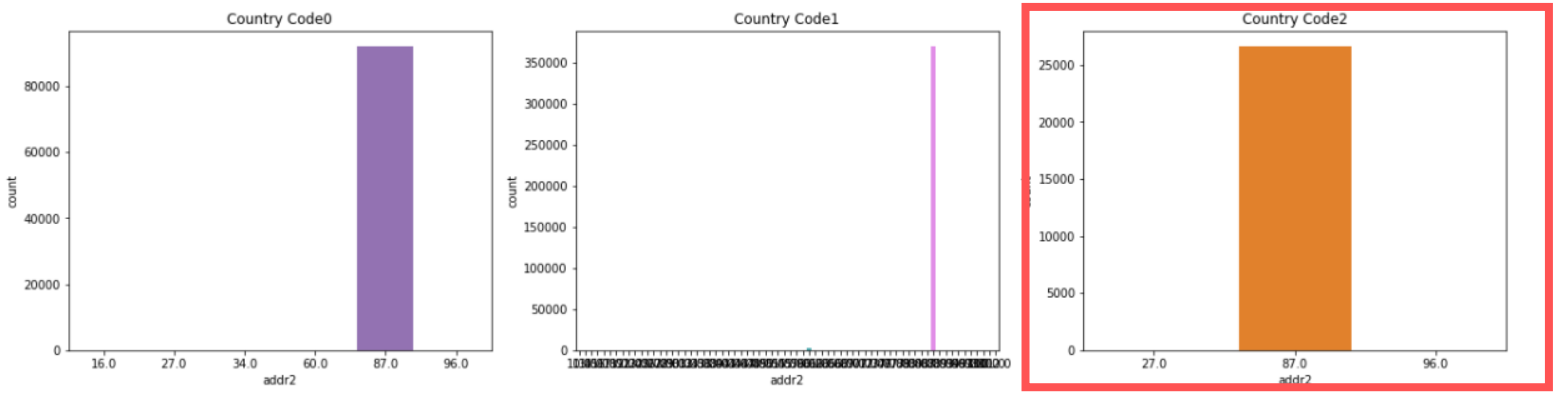
국가코드라

추측되는

Addr2값

addr2 & P_emaildomain

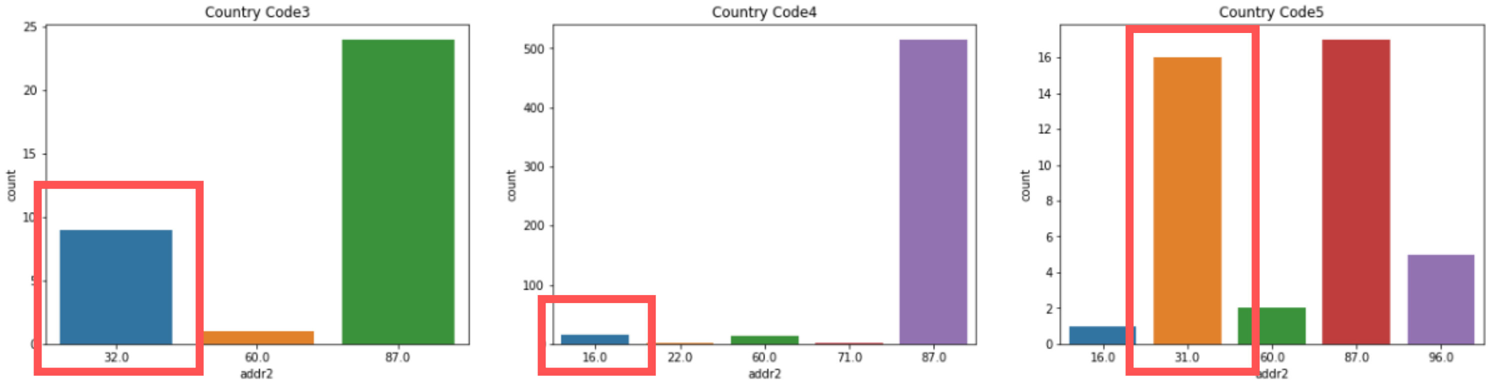
Region from email	Code
Null	0
Global	1
USA	**2**
UK	3
Germany	4
France	5
Spain	6
Mexico	7
Japan	8
anonymous	9



4.2 FE결과

addr2 & 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



4.2 FE결과

DeviceInfo

train_identity.csv

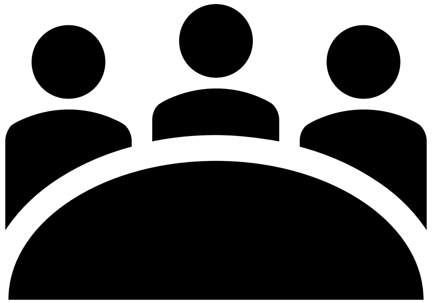
DeviceType

DeviceInfo

id1 - id38

Transaction & Identity

데이터의 통합이 필요함



4.2 FE결과

DeviceInfo

Article Talk Read Edit View history Search Wikipedia

Cramér's V

From Wikipedia, the free encyclopedia

In statistics, **Cramér's V** (sometimes referred to as **Cramér's phi** and denoted as φ_c) is a measure of association between two nominal variables, giving a value between 0 and +1 (inclusive). It is based on Pearson's chi-squared statistic and was published by Harald Cramér in 1946.[1]

Contents [hide]

Usage and interpretation [edit]

φ_c is the intercorrelation of two discrete variables[2] and may be used with variables having two or more levels. φ_c is a symmetrical measure, it does not matter which variable we place in the columns and which in the rows. Also, the order of rows/columns doesn't matter, so φ_c may be used with nominal data types or higher (notably ordered or numerical).

Cramér's V may also be applied to goodness of fit chi-squared models when there is a $1 \times k$ table (in this case $r = 1$). In this case k is taken as the number of optional outcomes and it functions as a measure of tendency towards a single outcome.[citation needed]

Cramér's V varies from 0 (corresponding to no association between the variables) to 1 (complete association) and can reach 1 only when the two variables are equal to each other.

isFraud(이분형)데이터와

train_Identity.csv 변수들(범주형)

연관성 파악

Cramer's V?

Calculation [edit]

Let a sample of size n of the simultaneously distributed variables A and B for $i = 1, \ldots, r; j = 1, \ldots, k$

$n_{ij} =$ number of times the values (A_i, B_j) were observed.

The chi-squared statistic then is:

$$\chi^2 = \sum_{i,j} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

Cramér's V is computed by taking the square root of the chi-squared statistic divided by the sample size

$$V = \sqrt{\frac{\varphi^2}{\min(k-1, r-1)}} = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}}$$

where:

- φ is the phi coefficient.
- χ^2 is derived from Pearson's chi-squared test
- n is the grand total of observations and
- k being the number of columns.
- r being the number of rows.

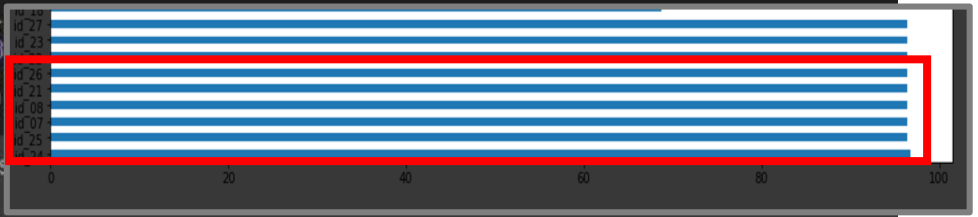
The p-value for the significance of V is the same one that is calculated using the Pearson's chi-squared te

The formula for the variance of $V = \varphi_c$ is known.[3]

In R, the function `cramerV()` from the package `rcompanion` [4] calculates V using the chisq.test function `cramersV()` from the `lsr` [5] package, `cramerV()` also offers an option to correct for bias. It applies th

```
[ ]  def cramers_v(x, y):
         confusion_matrix = pd.crosstab(x,y)
         chi2 = ss.chi2_contingency(confusion_matrix)[0]
         n = confusion_matrix.sum().sum()
         phi2 = chi2/n
         r,k = confusion_matrix.shape
         phi2corr = max(0, phi2-((k-1)*(r-1))/(n-1))
         rcorr = r-((r-1)**2)/(n-1)
         kcorr = k-((k-1)**2)/(n-1)
         return np.sqrt(phi2corr/min((kcorr-1),(rcorr-1)))


     x = id_new['isFraud']

     for i in id_new.columns:
       if i == 'isFraud' or i == 'TransactionID':
         print(i + ' pass')
         continue

       y = id_new[i]

       score = cramer
       if score > 0.3
         print(i + "
         print(score)
```

```
TransactionID pas
id_21 score:
0.42701968880673
id_25 score:
0.4985777900494445
DeviceInfo score:
0.38344012723979387
isFraud pass
```




1. 목적
2. 변수분석
3.결측치처리
4. FE
5. 결과

5. 결과

5. 전처리결과

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

끝