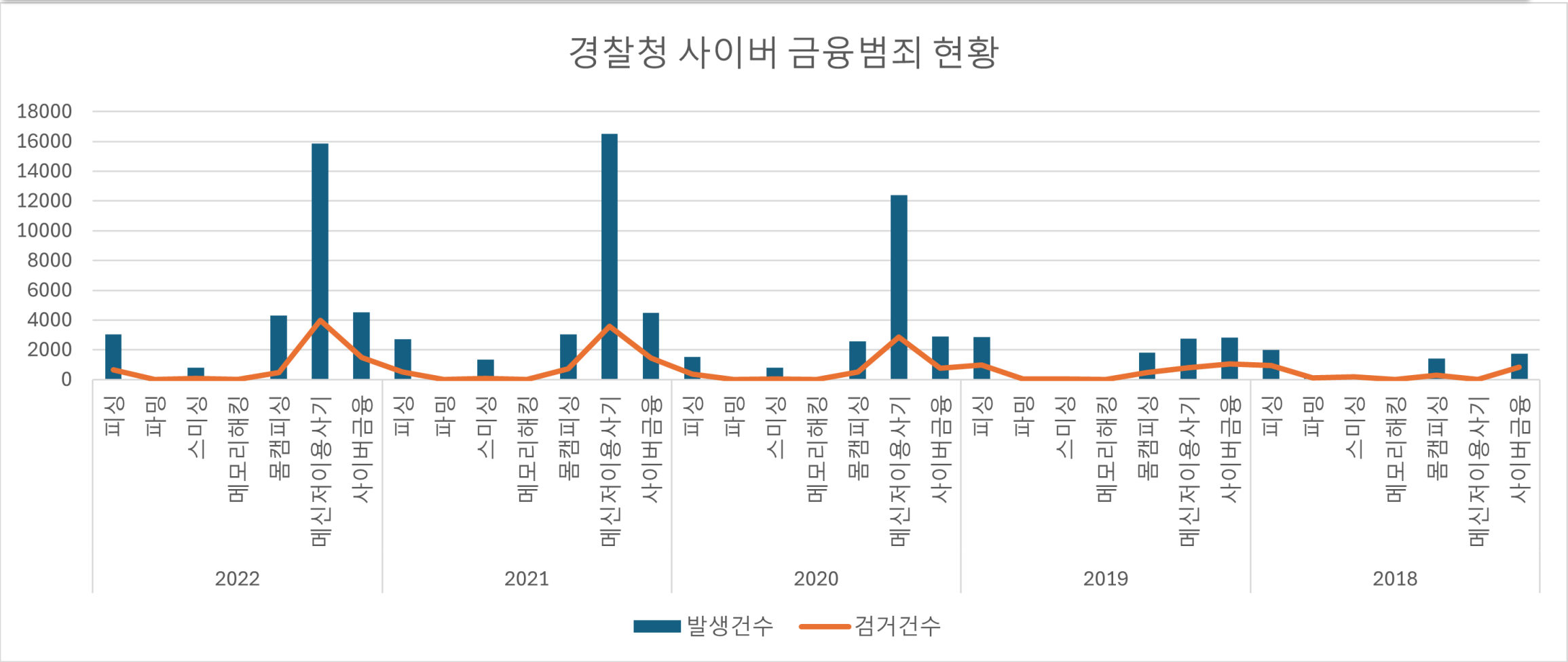
거래 빅데이터를 이용한
이상 거래 탐지

- 데이터 분석과 탐지 방법론을 중심으로

가천대학교
Gachon University

1. 프로젝트 필요성



- 사이버 금융범죄 증가와 검거 현황

  -발생건수, 검거건수

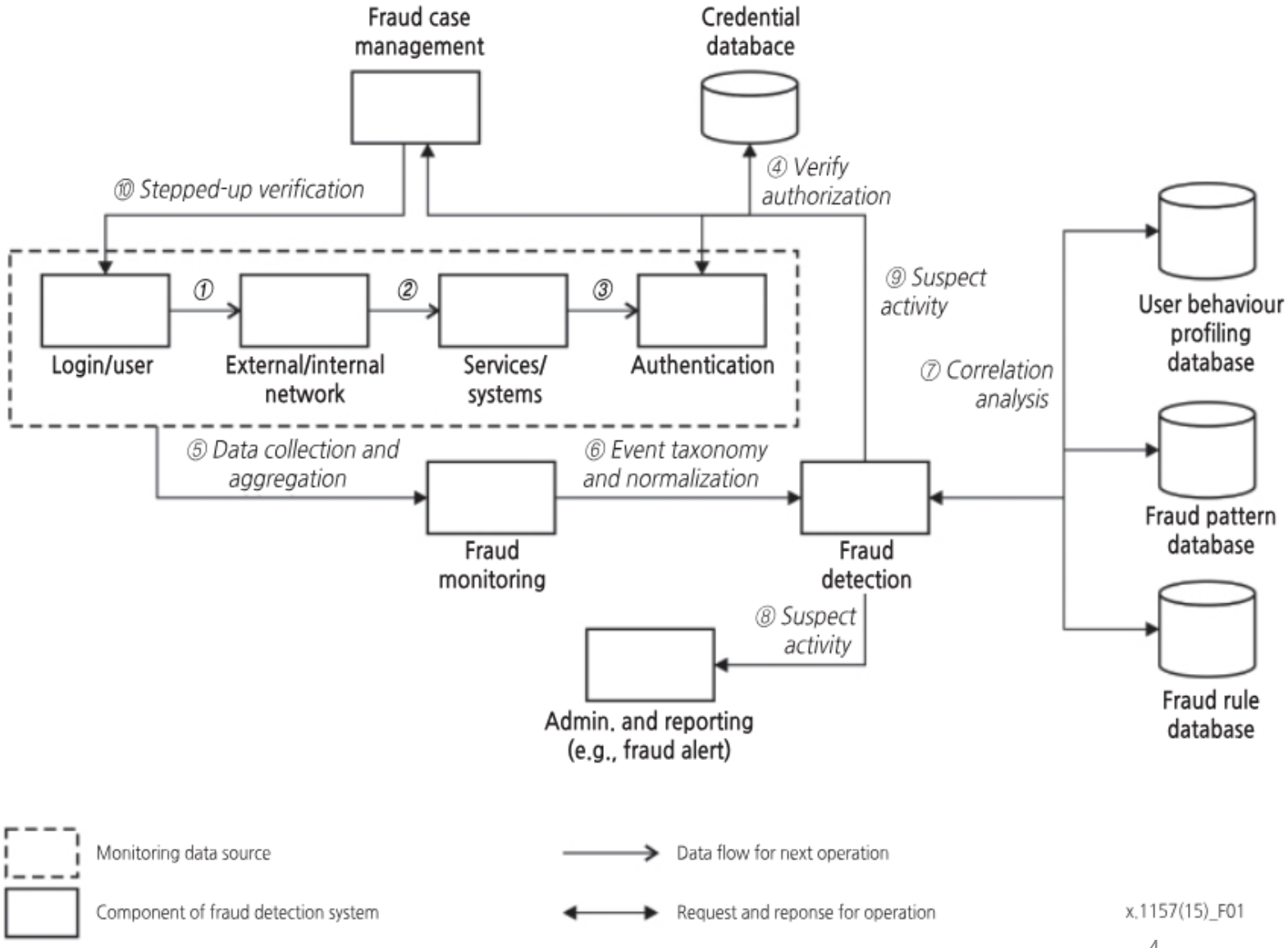
1. 프로젝트 필요성

금감원, 은행 및 증권회사의 탐지시스템 운영 현황 분석

▲ 이상금융거래 탐지시스템(FDS·Fraud DetectionSystem)이 지난해 3천6백여건의 금융 사고를 예방한 것으로 나타났다. 자료/금융감독원

2. 프로젝트 주제 및 목적

- 주제: 카드거래 데이터 기반한 FDS 구축
- 목적: FDS 통해 사기거래 감소시키는 모델 생성
- 데이터 분석 후 FDS를 3가지 방법으로 모델 생성 후 평가



<우> 실제 FDS작동원리

3. 데이터 선정

Credit Card Fraud

Crack the model from credit card fraudster dataset.

Val

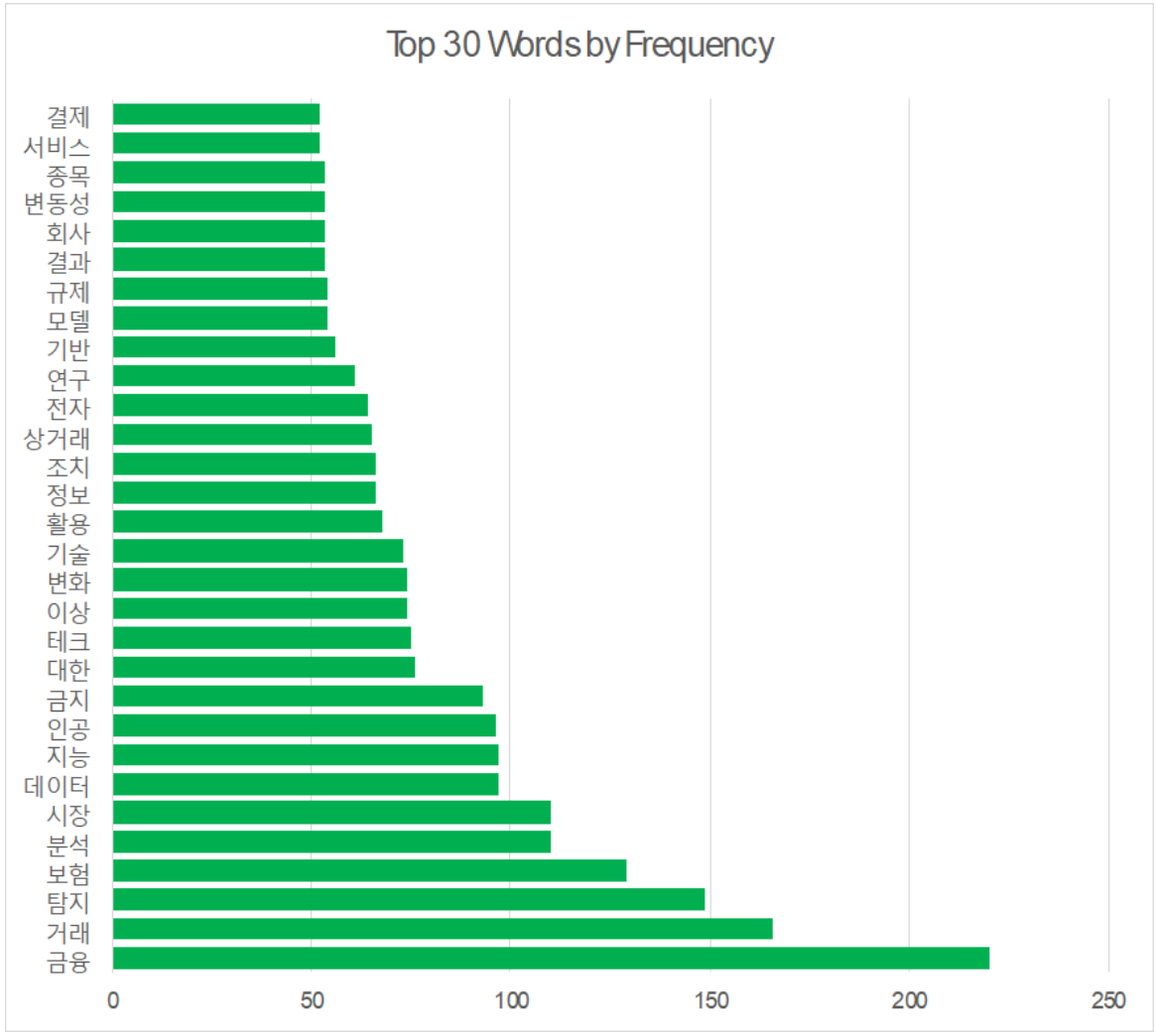
- distance_from_home: 주거지와 거래 장소와의 거리
- distance_from_last_transaction: 마지막 거래장소와의 차이
- ratio_to_median_purchase_price: 평균 결재 금액과의 비

Bool

- repeat_retailer: 거래가 동일한 소매업체에서 이루어졌는지
- used_chip: 신용카드 칩을 사용한 거래인지
- used_pin_number: PIN 번호를 사용하여 거래가 이루어졌는지
- online_order: 거래가 온라인 주문이었는지
- fraud(label): 거래가 사기였는지, 정상 거래였는지

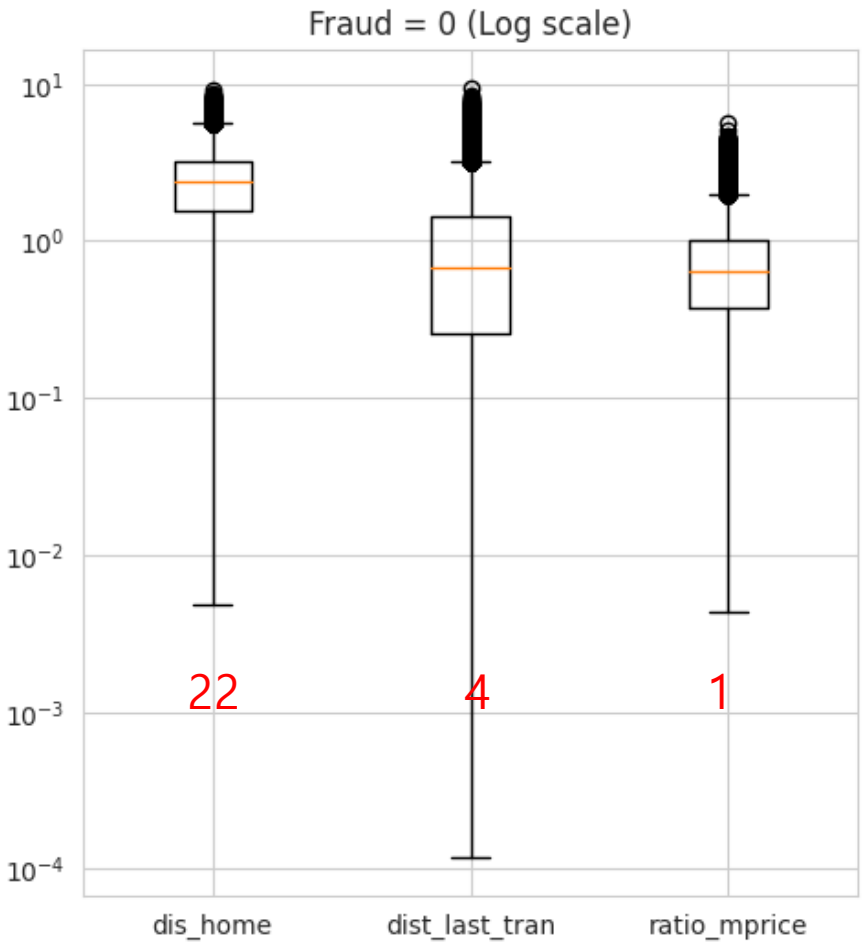
4. 선행 연구 분석 결과

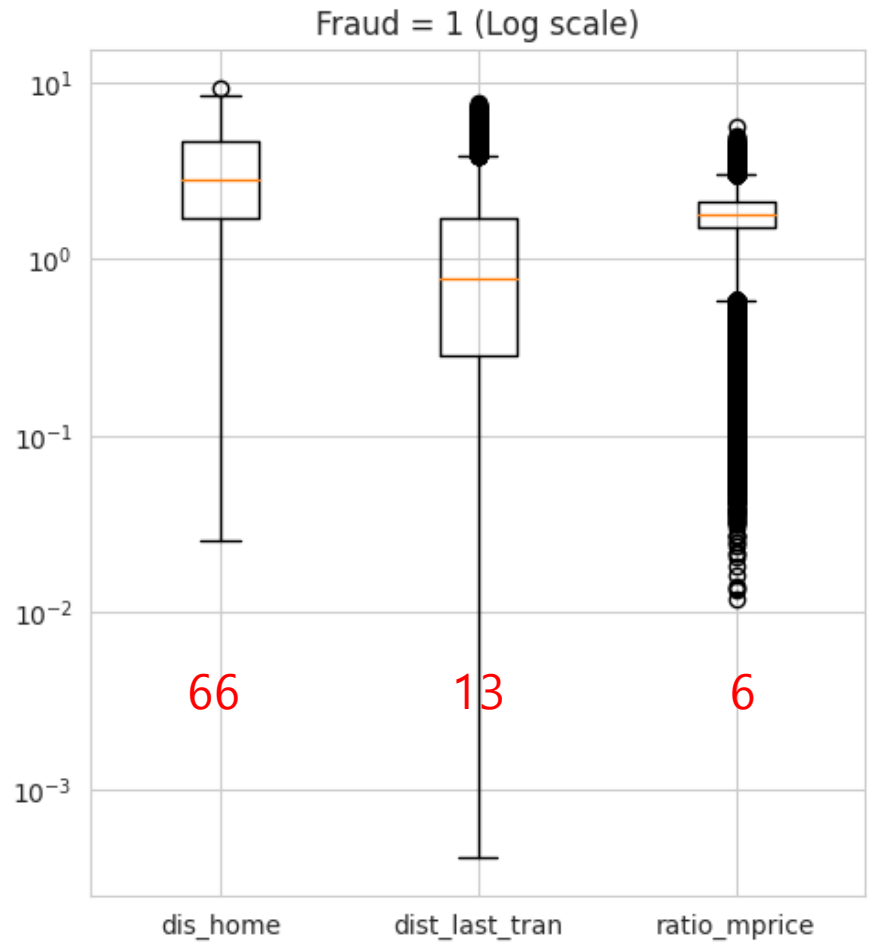
- **RISS 논문 초록 웹 크롤링 후 빈도 시각화, 연구 동향**



5.1 탐색적 데이터 분석(EDA)

- 양적 변수 3개의 기초통계량 분석 *로그스케일 적용, 하단 숫자는 평균값





- 주거지와 거래 장소와의 거리

  -평균 기준으로 사기거래가 정상 거래의 약 3배 원거리에서 발생

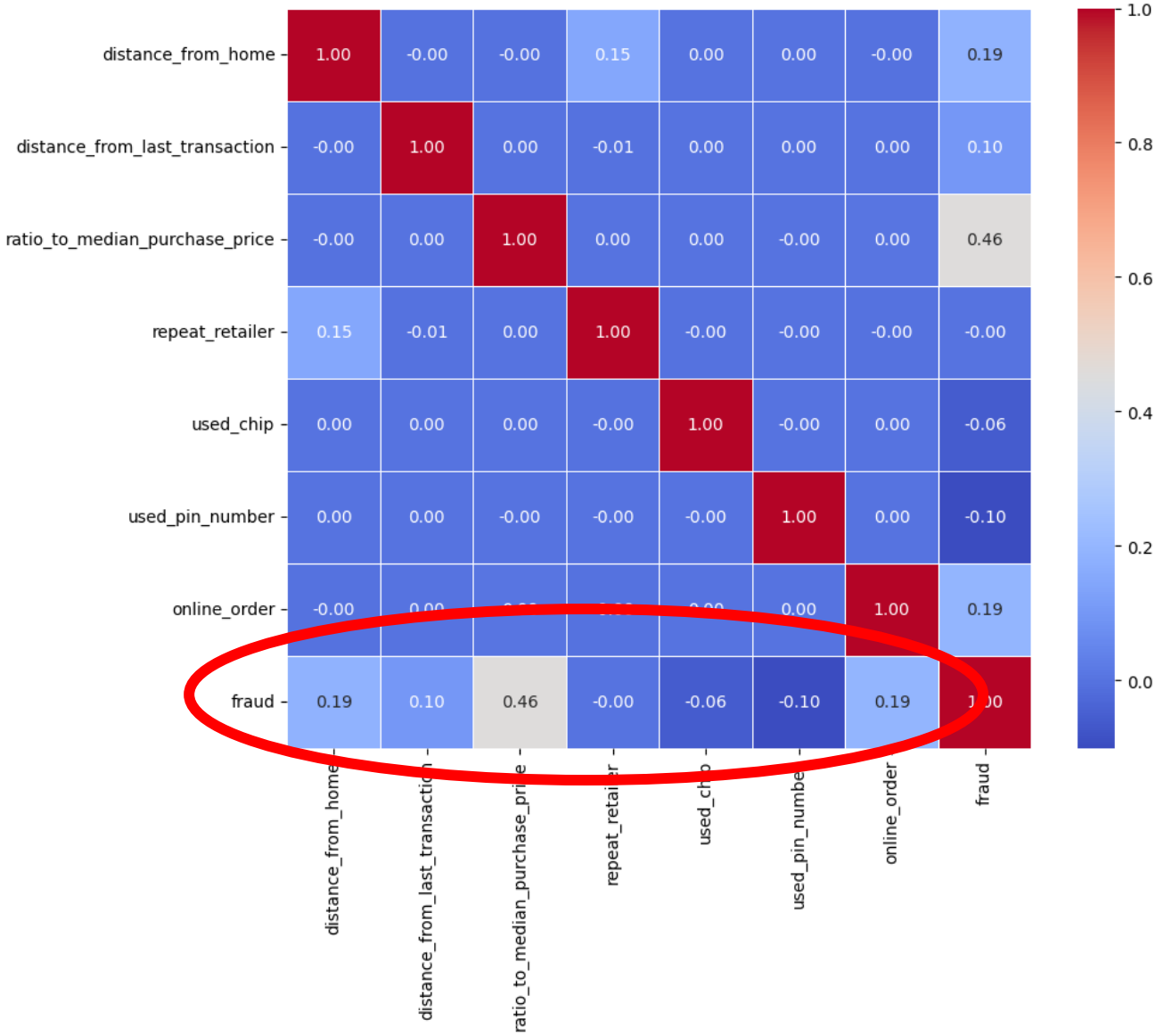
- 마지막 거래장소와의 거리

  -평균 기준으로 사기거래가 정상 거래의 약 3배 원거리에서 발생

- 평소 결제 금액 중앙값과의 비

  -평균 기준으로 사기거래가 정상 거래의 약 6배 많은 금액 결제

5.1 탐색적 데이터 분석(EDA)



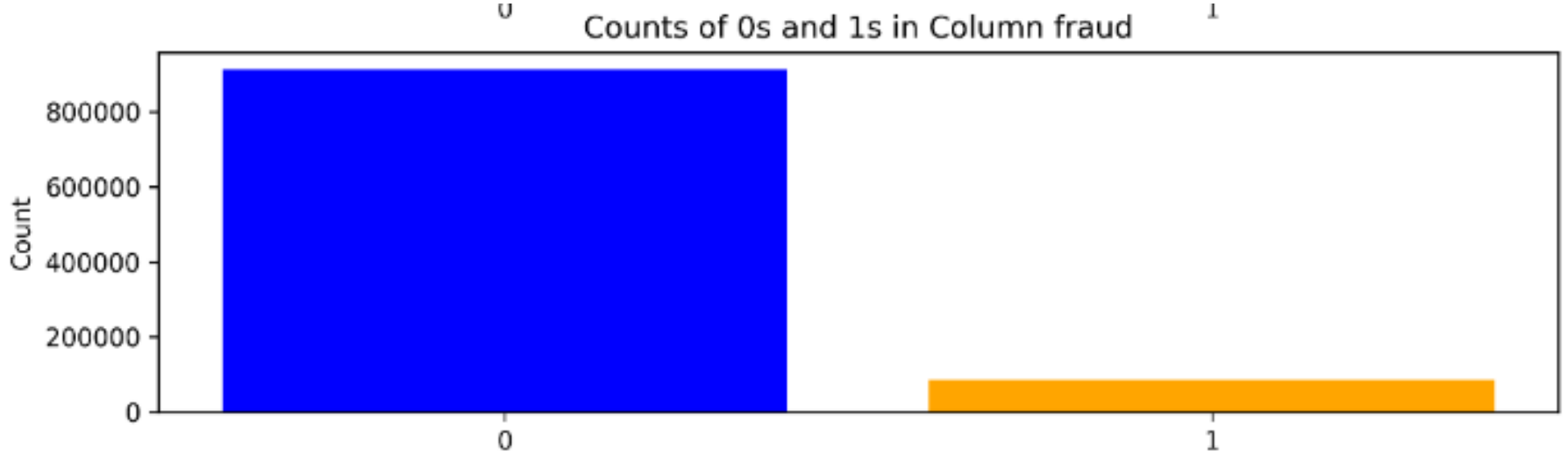
- 붉은 계열: 상관관계 높다
- 청색 계열: 상관관계 낮다

- 양적 변수를 주목해 봤을 때도, 상관계수가 0.19, 0.10, 0.46

  =선형성이 비교적 낮아 선형 모델 적용X

5.1 탐색적 데이터 분석(EDA)

- 데이터셋의 정상거래와 사기거래 수

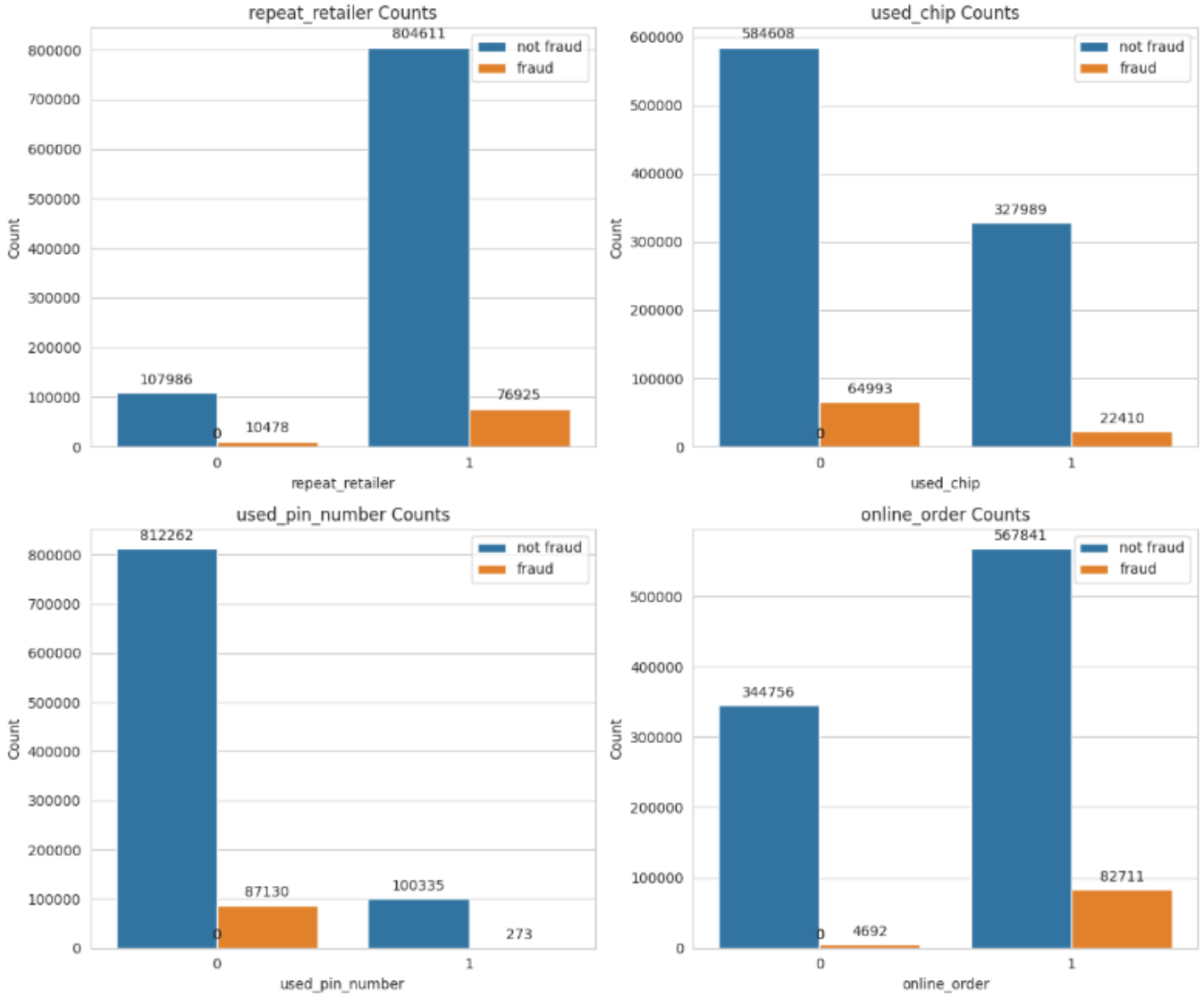


- 불균형 데이터: SMOTE적용하여 오버 샘플링

  - 소수의 클래스의 데이터를 복제하거나 합성하여 데이터셋의 클래스 분포를 균형 있게 만든다.

5.1 탐색적 데이터 분석(EDA)

- 범주형 데이터 값에 따라 사기/정상거래 수 측정



- 카이 제곱 검정 사용
- 각 범주형 변수와 사기거래의 연관성 검정 결과:
  online_order>used_chip>used_pin_number>repeat_retailer
- repeat_retailer의 P-value=0.4908: Fraud 변수와의 유의미한 상관관계가 없음

  =>해당 변수를 데이터셋에서 제외

5.2 방법론0: 사용 가능한 방법론 조건 및 분석 과정

- PCA와 K-means 기각 이유

  -해당 데이터셋의 연속형과 범주형 개수는 retail변수 기각 후 각각 3개

  -사용할 알고리즘은 범주형 변수 처리에 적합해야, 라벨有=지도학습 가능

  -PCA경우 모든 주성분의 분산이 거의 동일, k-mean는 모델 성능 매우 저조

  = 둘 다 연속형 범주에 최적화된 알고리즘이기 때문

- 데이터 전처리

  -연속형 변수는 정규화

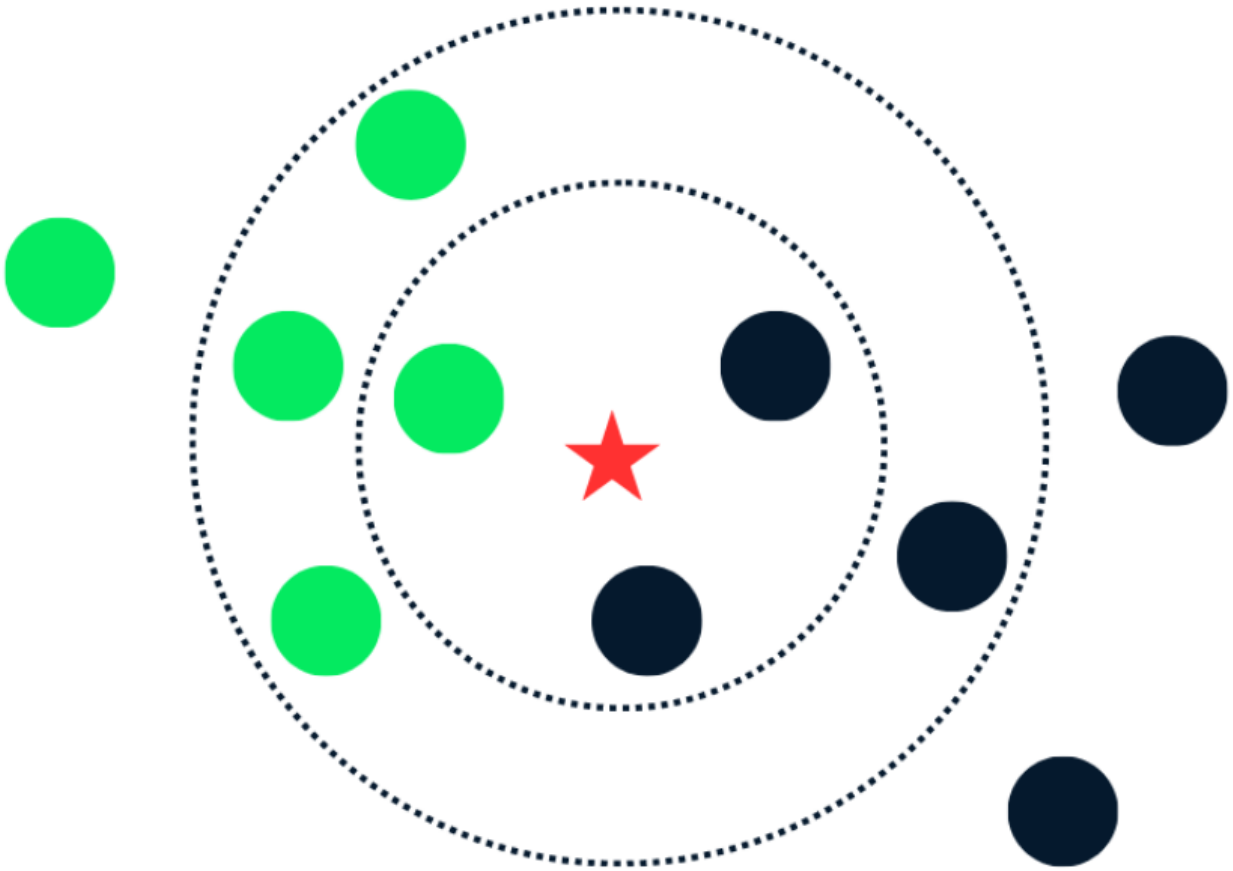
  -오버 샘플링을 통해 정상거래와 사기거래 클래스 간의 비율을 맞춤

- 성능 평가 지표

  -F1-Score

  -원 데이터셋(10) / 트레인셋(8): 모델 구축 / 테스트셋(2): f1-score로 성능 평가

5.2 방법론1: K- Nearest Neighbors



- 지도학습, 분류 알고리즘

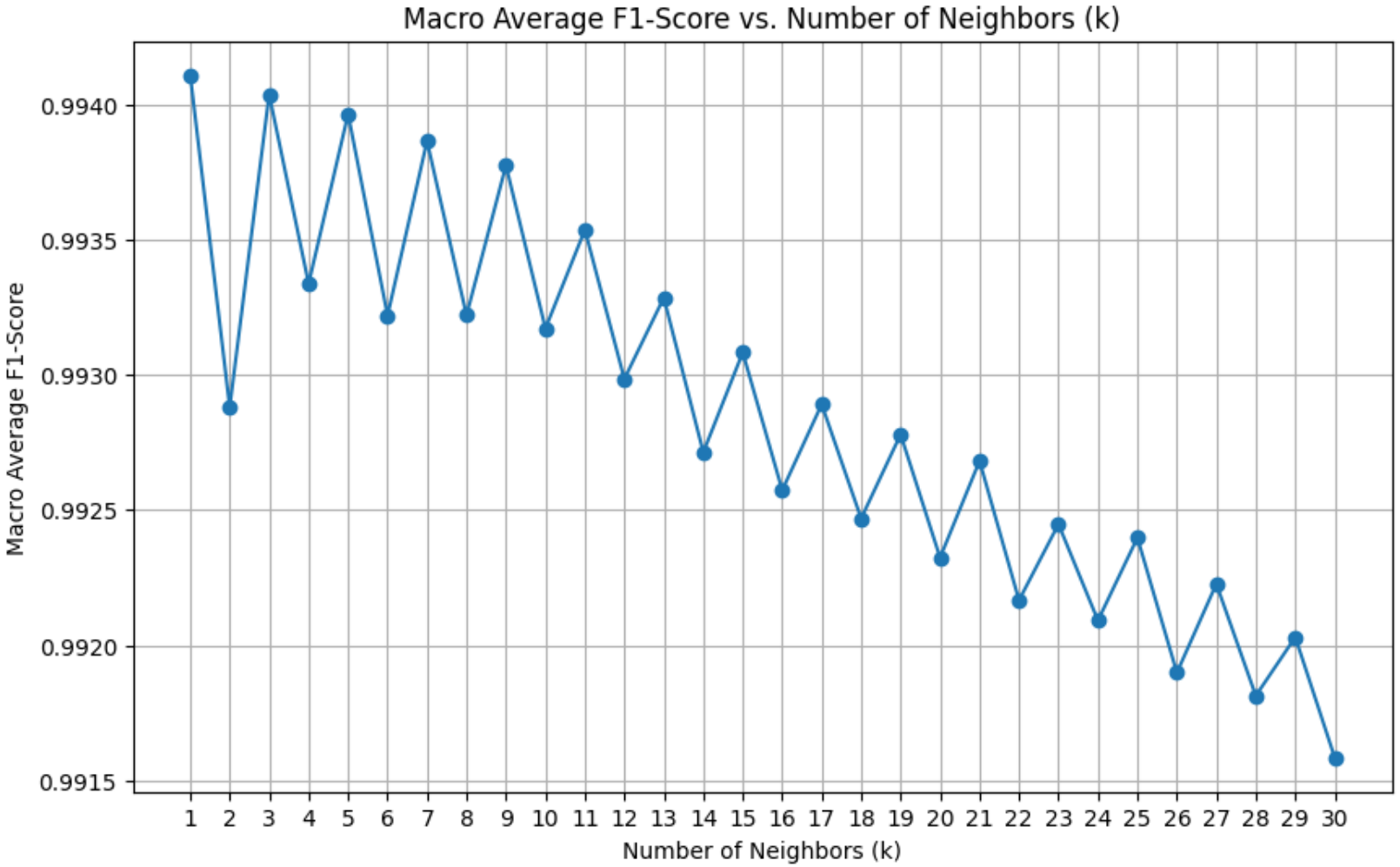
  (1) 분류하려는 새로운 데이터 포인트와 기존 데이터 포인트들 간의 거리 측정

  (2)거리를 바탕으로 가장 가까운 K개의 이웃을 선정

  (3) 분류 작업에서는 K개의 가장 가까운 이웃 중 다수가 속한 클래스를 새로운 데이터의 클래스로 예측

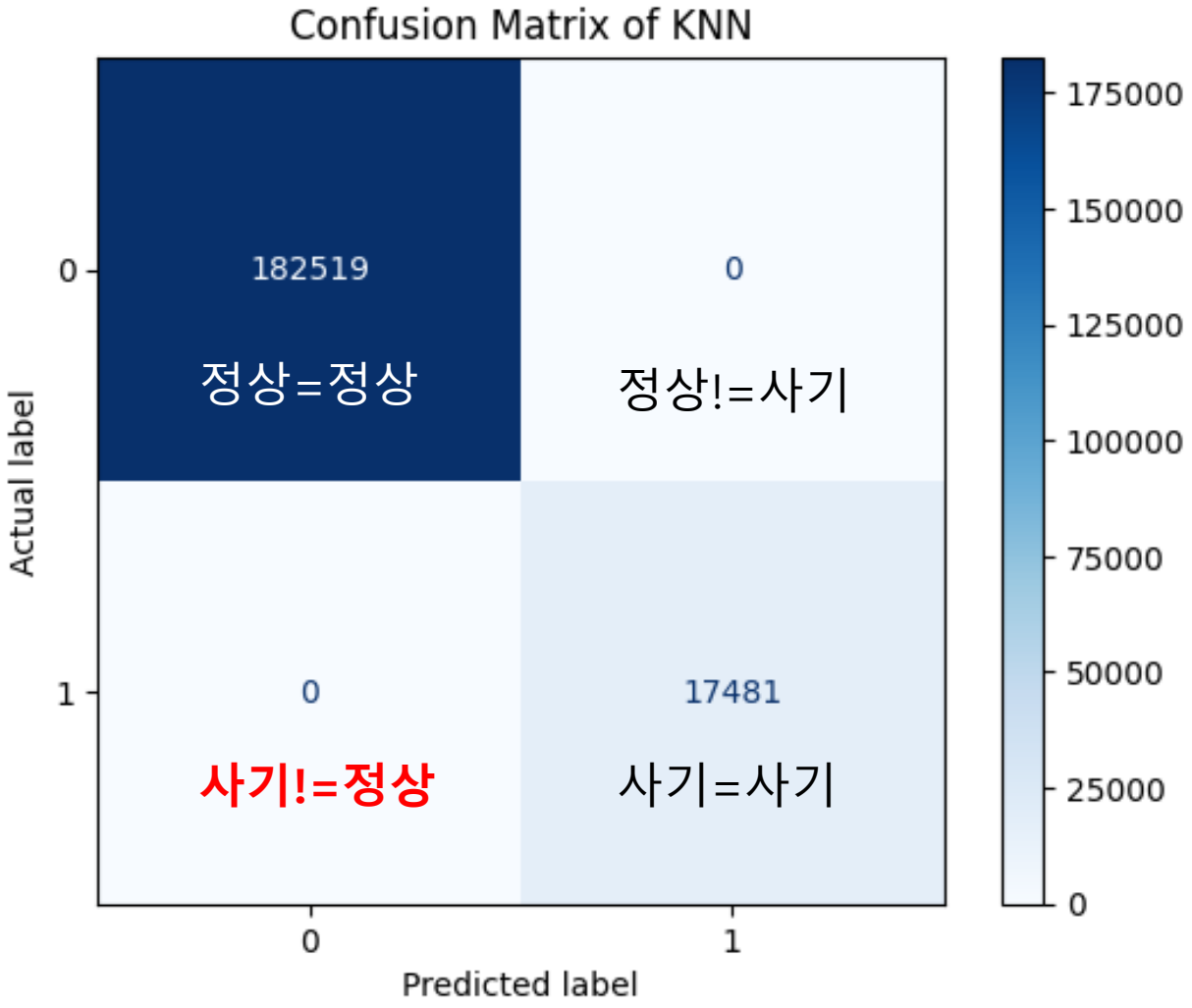
- 근접 이웃을 몇 명 참고할 것인가?
  각 k개 이웃별로 성능을 검증

5.2 방법론1: K- Nearest Neighbors



- K=1, 근접 이웃 1개 참고 했을 때 성능 가장 좋음
- 짝수에서 성능이 낮게 나오는 이유?
  - 참조한 이웃의 클래스가 정상과 사기의 비율이 1:1로 나눠졌을 때 제대로 분류X
- 5-Fold Cross-Validation 사용
  - 모델의 일반화 능력 더 정확하게 평가 가능

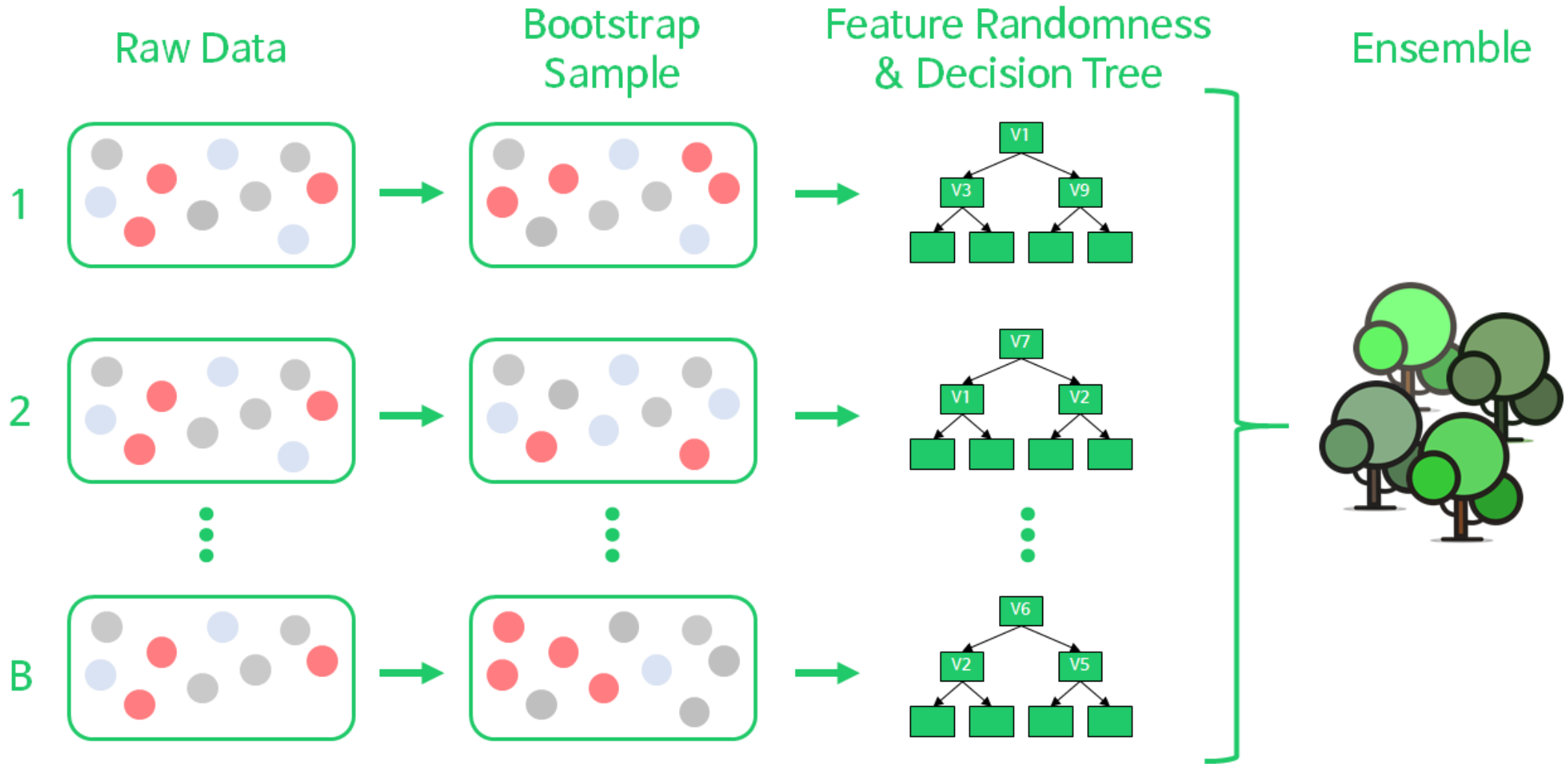
5.2 K- Nearest Neighbors 평가



KNN의 F1-Score: 1.0000000000000000

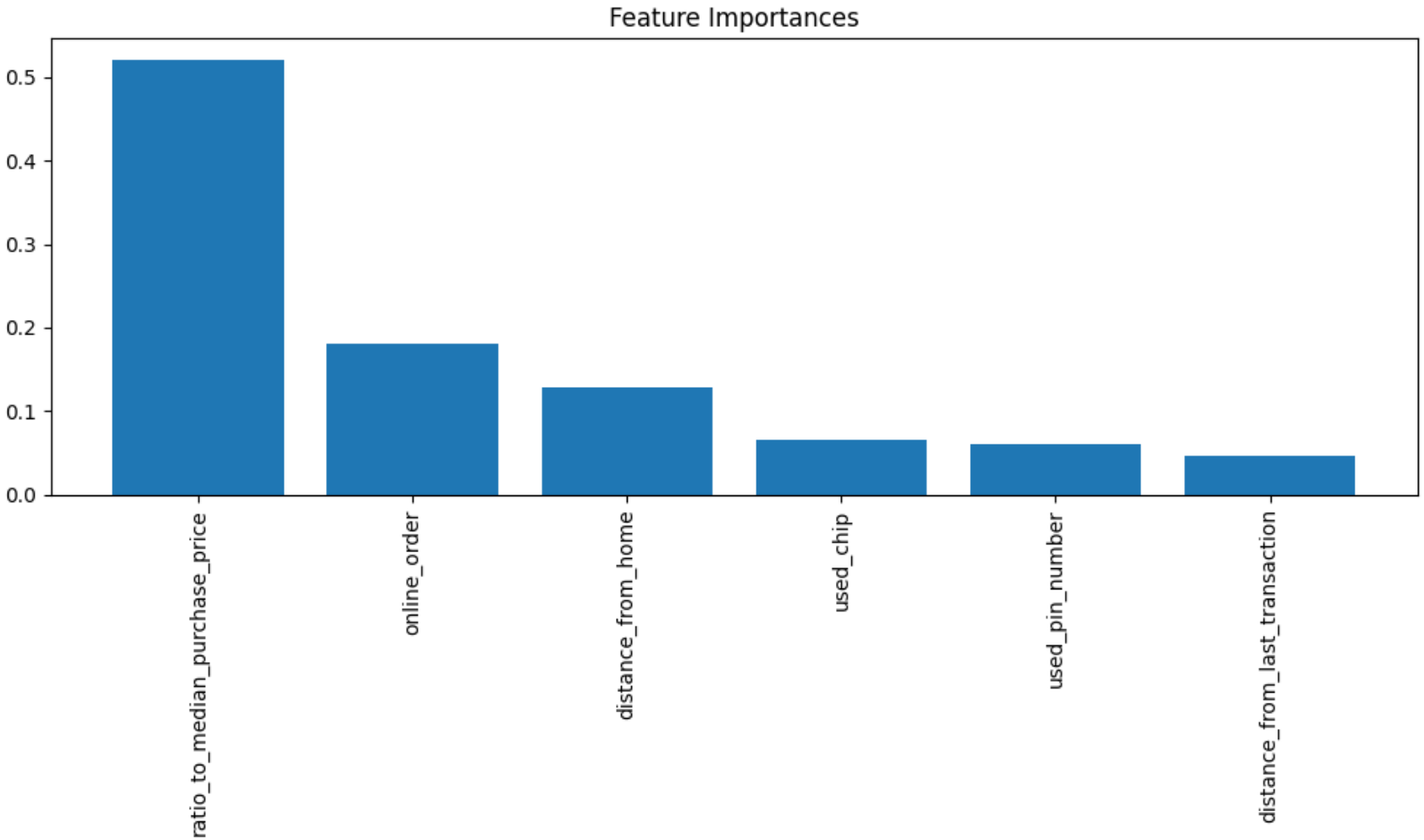
- 모델이 테스트셋에서 정상과 사기 거래를 완벽하게 분류했음을 알 수 있다.
- 컨퓨전 매트릭스?

5.2 방법론2: 랜덤 포레스트



5.2 방법론2: 랜덤 포레스트

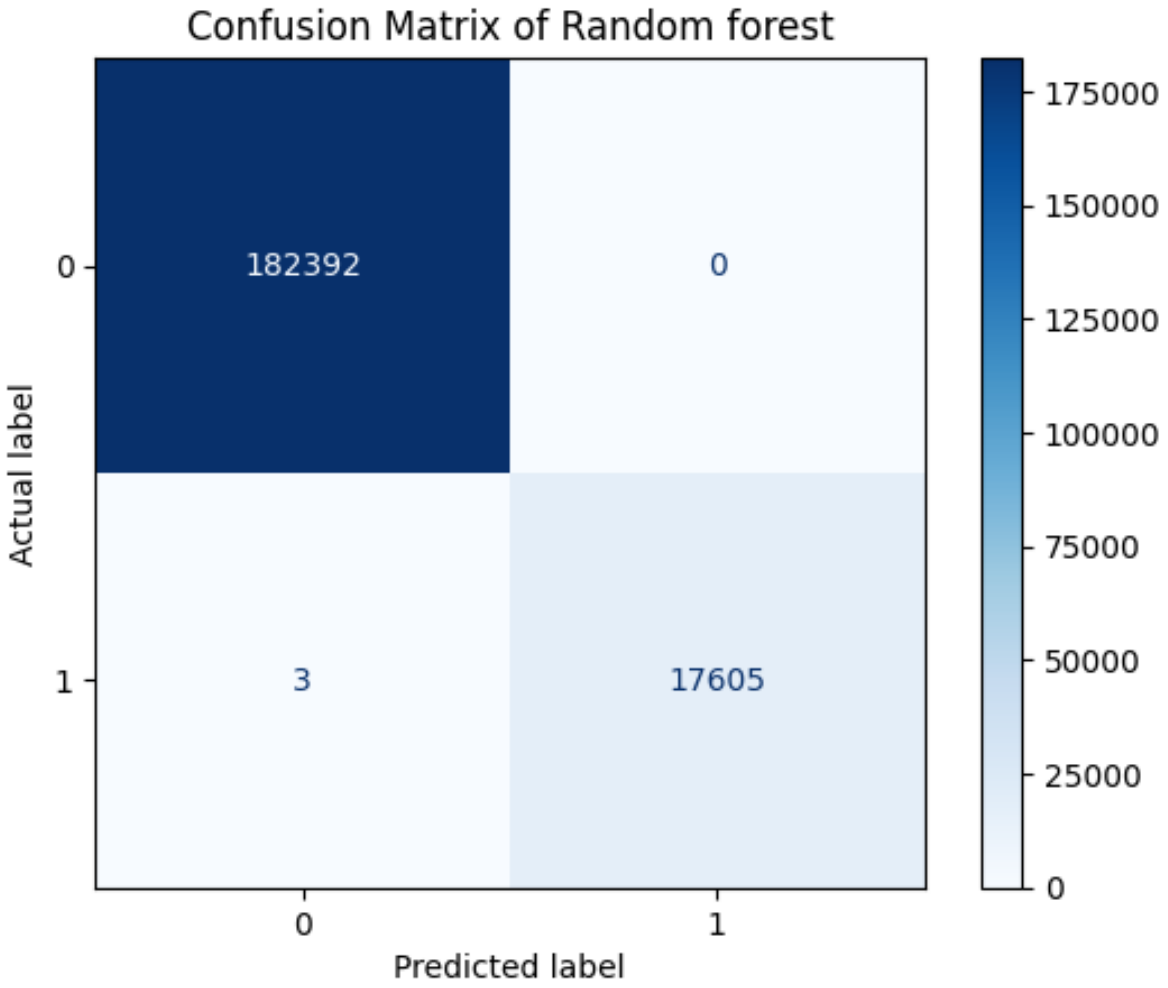
- 예측에 중요한 역할을 한 변수



<변수 중요도>

1. 이전 결재 금액 중앙값과의 비
2. 온라인 주문이었는지
3. 주거지와의 거리
4. 신용카드 칩을 사용했는지
5. 핀 넘버를 사용했는지
6. 마지막 거래 장소와의 거리

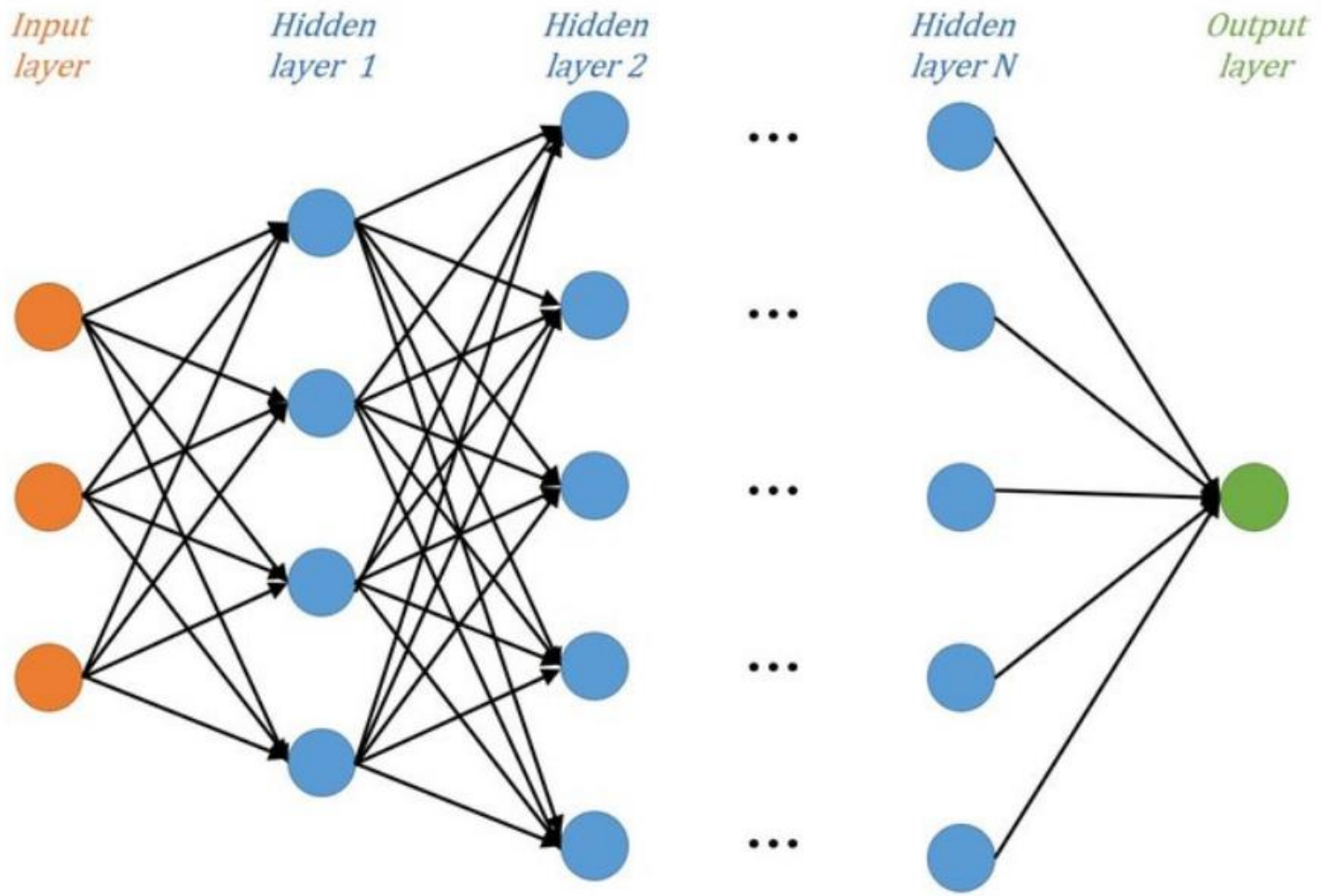
5.2 랜덤 포레스트 평가



RandomForest의 F1-score: 0.9994320440733 8

- 정상 거래를 사기 거래로 잘못 예측한 경우는 0번
- 사기 거래를 정상 거래로 잘못 예측한 횟수는 3번

5.2 방법론3: 인공신경망(딥러닝)



- 이 데이터셋의 로지스틱 회귀모형의 f1-score는 0.7

  = 범주형 데이터의 수가 너무 많기 때문이라고 예상됨

- 인공신경망: 사용 함수 sigmoid로 동일하지만 오차 역전파를 통해 최적 기울기를 찾아내는 능력
- 완전한 블랙박스, 결과를 설명 불가

5.2 방법론3: 인공신경망(딥러닝)

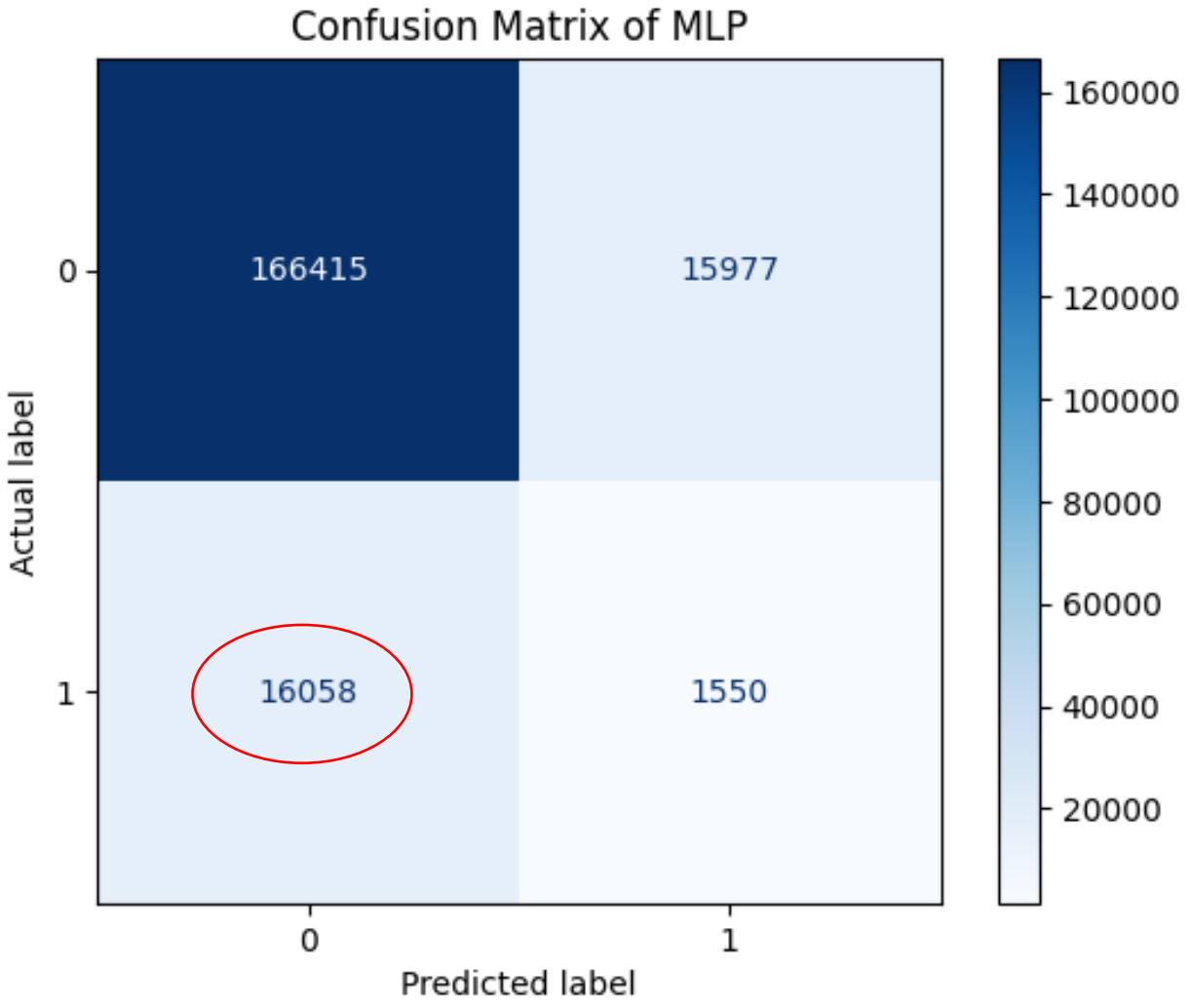
```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 512)               3584

 dense_1 (Dense)             (None, 128)               65664

 dense_2 (Dense)             (None, 64)                8256

 dense_3 (Dense)             (None, 16)                1040

 dense_4 (Dense)             (None, 8)                 136

 dense_5 (Dense)             (None, 1)                 9

=================================================================
Total params: 78689 (307.38 KB)
Trainable params: 78689 (307.38 KB)
Non-trainable params: 0 (0.00 Byte)
```

- **Activation func:**

  중간층에서는 Relu를 사용, 마지막 학습에서는 종속변수가 0,1 값이기 때문에 시그모이드 함수를 사용

- **시그모이드 함수는 y값이 0이나 1로 대응될 확률 값을 의미**
- **오차 함수로는 이진 분류에 적절한 바이너리 크로스 엔트로피 사용**

5.2 인공신경망 평가

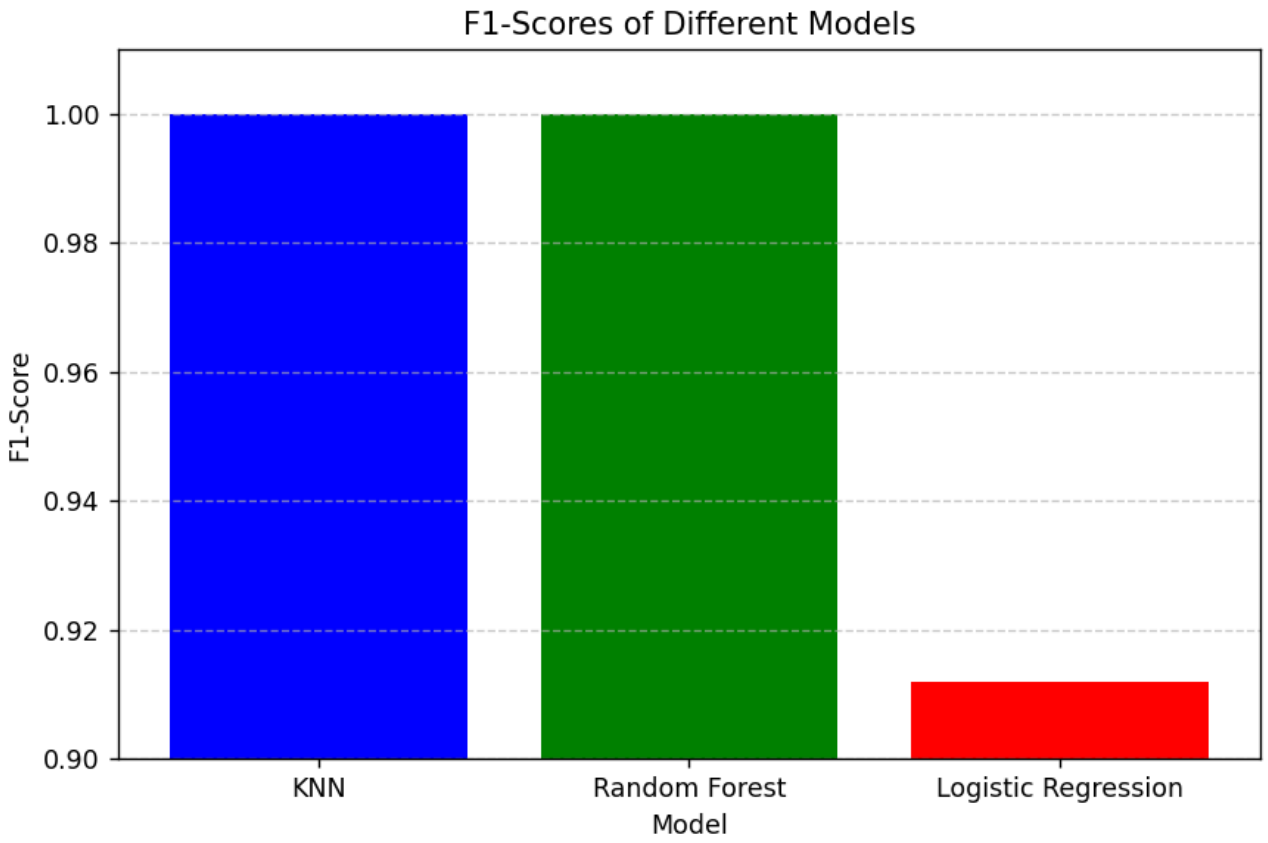


인공신경망의 F1-Score: 0.9122004028887397

- 사기 거래를 정상 거래로 분류한 케이스가 16058번으로 상당히 많은 숫자
- 앞서 사용한 분류 알고리즘과 비교했을 때 낮은 결과
- Why?

6. 연구 결과 정리

- 3가지 방법론 성능 결과



- 왜 분류 알고리즘이 인공신경망보다 성능이 좋게 나왔을까?
- 인공신경망은 이진 데이터의 경우같이 각 특성이 적은 정보를 담고 있을 때 모델이 이 정보를 과대 해석하여 과적합을 일으킨 것이라고 설명할 수 있다.
- 랜덤 포레스트는 인공신경망처럼 복잡한 연산 사용X
- KNN은 특징 공간 N차원에서 기존 데이터 값들 중 가장 가까운 이웃 정보 기반으로 새로운 데이터의 분류하는 단순한 모델

=>랜덤 포레스트와 KNN은 범주형 변수의 정보를 효과적으로 처리

6. 연구 결과 정리

- 3가지 방법론의 시간복잡도

(1) 랜덤 포레스트	(2) 인공신경망(딥러닝)	(3) K-NN
O(m*log(N))	$O(\sum_{i=1}^{L}(N_i \cdot N_{i-1}))$	O(N)

- FDS는 새 데이터가 들어온 후 빠른 분류(탐지)가 가능해야 함
- 시간복잡도까지 고려했을 때 가장 이상적인 방법론은 랜덤 포레스트

  랜덤 포레스트: f1-score 0.99, 시간 복잡도 가장 낮음

*N은 데이터 개수, m: 트리 개수, $N_i \cdot N_{i-1}$:이전 노드와 현재 노드 개수의 곱

7. 결론

- 실제 적용 시 유효성

  (1) 사용한 데이터 - 개인의 특성에 따른 거래 패턴 반영X

  (2) 과거 데이터만 사용하는 방법론

  - 거래 패턴 변화 지속적으로 예측 불가

  (3) 실제 대용량 데이터셋에 라벨 붙이는 비용

7. 결론

- 프로젝트 의의

  (1) 데이터셋 내 주요 변수 선별

  1위 “이전 결재 금액 중앙값과의 비” = 고액 타겟

  2, 3위 ”온라인 거래인지”, “주거지와 거래 장소와의 거리”

  = 비대면 거래 상승에 따라 사용자 정보 해킹 후 원거리의 IP를 통해 결제를 시도하는 인터넷 사기가 지속적으로 증가함을 시사

  (2) 데이터셋의 성격에 따라 분석하는 법을 탐색

  = 주어진 데이터셋: 연속/범주, 라벨 有

  = 비교적 간단한 알고리즘 구조를 가진 지도학습 방법론이 가장 좋은 결과를 도출

Thank you

Q&A