

R을 이용한 신용카드 사기 감지

김승혁

데이터셋 출처: Kaggle의 Credit Card Fraud Detection

The screenshot shows the Kaggle dataset page for 'Credit Card Fraud Detection'. The header indicates it is a 'Reviewed Dataset' with 1280 votes. The title is 'Credit Card Fraud Detection' with a subtitle 'Anonymized credit card transactions labeled as fraudulent or genuine'. It is created by the 'Machine Learning Group - ULB' and was last updated a year ago. The page has tabs for 'Overview', 'Data', 'Kernels', 'Discussion', and 'Activity'. There are buttons for 'Download (68 MB)' and 'New Kernel'. Below these are tags: 'finance', 'crime', 'medium', and 'featured'. The main content area is divided into three sections: 'Top Contributors', 'Kernels', and 'Discussion'. The 'Top Contributors' section lists joparga3 (1st), Currie32 (2nd), and Nell Schnelder... (3rd). The 'Kernels' section lists 'In depth skewed data cl...' (278 votes), 'Predicting Fraud with Te...' (129 votes), and 'GBM vs xgboost vs light...' (78 votes). The 'Discussion' section lists 'does standardizing tim...' (1 reply), 'Fraud detection' (4 replies), and 'Fully Unsupervised Ap...' (0 replies).

Top Contributors	Kernels	Discussion
joparga3 1st	In depth skewed data cl... 278 votes run a year ago	does standardizing tim... 1 reply 5 days ago
Currie32 2nd	Predicting Fraud with Te... 129 votes run a year ago	Fraud detection 4 replies 20 days ago
Nell Schnelder... 3rd	GBM vs xgboost vs light... 78 votes run 6 months ago	Fully Unsupervised Ap... 0 replies 21 days ago

본 데이터셋은 대규모 데이터 마이닝 및 사기 탐지에 대한 ULL (Université Libre de Bruxelles)의 Worldline 및 Machine Learning Group (<http://mlg.ulb.ac.be>)의 연구 협력 과정에서 수집 및 분석되었습니다.

Credit Card Fraud Detection

데이터 개요(1)

신용카드 사기 급증

2017-02-10 오전 10:16 kor 조회 3078

Text Size: + -



카드 사용이 더욱 증가하면서 그와 함께 신용카드 사기 사건도 급증하고 있어 특별한 주의가 요망된다.

신용카드사와 가맹점들이 개인정보보호 등 안전장치를 강화하고 있지만, 미국에서 신용카드 사기는 오히려 급증한 것으로 나타났다.

컨설팅업체인 '자블린 스트러티지 앤 리서치'(Javelin Strategy & Research)와 개인정보 도용 방지업체인 '라이프록'(LifeLock)은 지난해 신용카드 사기 피해자가 전년 대비 18% 증가하며 지난 2003년 이래 최고치를 나타냈으며 신용카드 사기에 따른 피해액도 160억달러에 달했다고 밝혔다.

신용카드 사기:

다른 사람의 카드 정보를 온라인 쇼핑에 이용,
타인의 이름으로 은행계좌를 개설하고 신용카드를 발급.

신용카드 사기가 급증하여 거래가 사기인지 아닌지
높은 신뢰도로 예측할 수 있는 알고리즘이 필요함.

Credit Card Fraud Detection

데이터 개요(2)

이미 PCA된 28개의 컬럼, Time 컬럼, Amount 컬럼, Class 컬럼이 있음.

PCA되어있는 28개 컬럼(V_1, V_2, \dots, V_{28}) : 종모양의 분포를 따르고 있으나 편향성이 큰 경우도 있음.
데이터의 원천이나 특성에 관해서는 개인정보 보호를 위해 공개되지 않음.

Time 컬럼 : 각각의 트랜잭션이 첫번째 트랜잭션을 기준으로 몇초 후의 트랜잭션인지를 나타냄.

Amount 컬럼 : 트랜잭션의 금액을 나타냄. (min: 0, max=25691)

Class 컬럼: 0 과 1의 값을 가짐. (0: 사기 아님, 1: 사기)

데이터 정제 전

Cross Table 비교

KNN/ C5.0/ C&RT	예측: 사기아님	예측: 사기
실제: 사기아님 (100%)	100/ 99.99/ 99.98	0/ 0.0082/ 0.0176
실제: 사기 (100%)	99.81/ 46.43/ 25.00	1.19/ 53.57/ 75.00

정확도: 99.806%, 99.901%, 99.933%

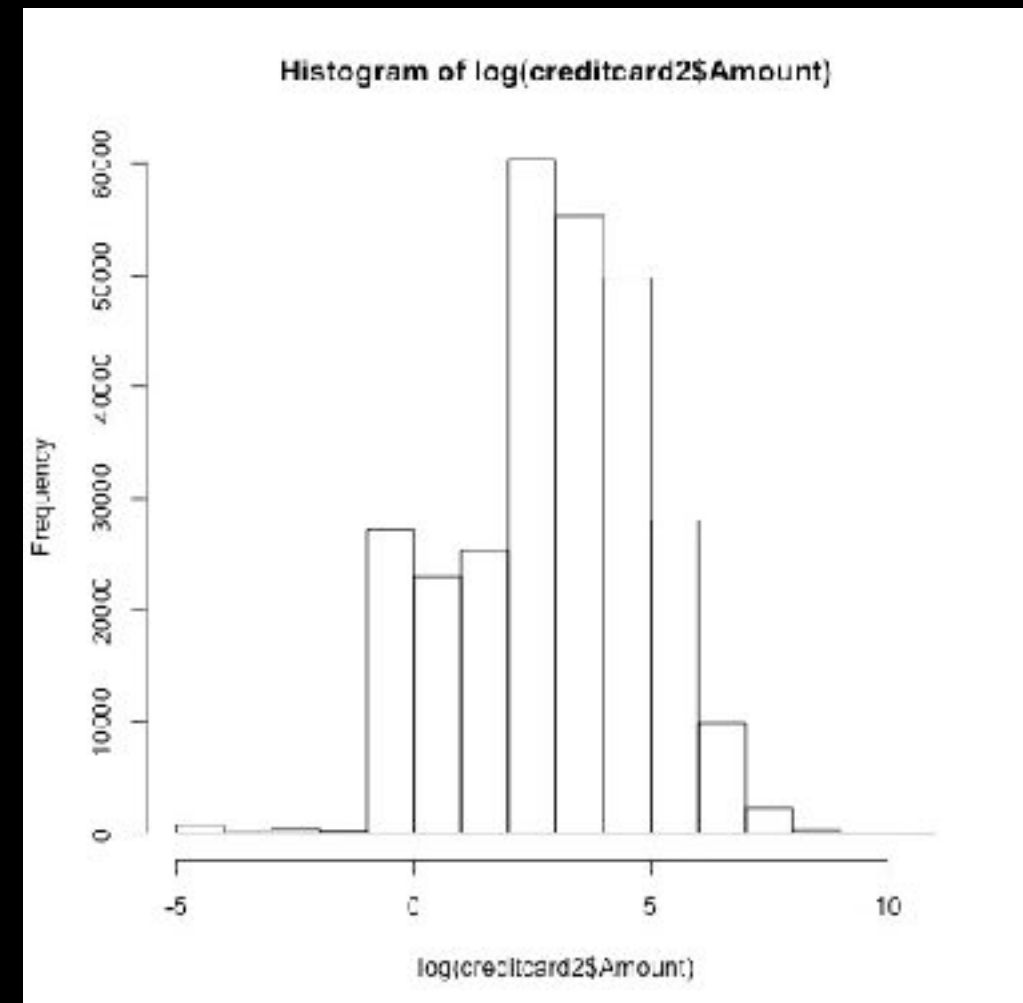
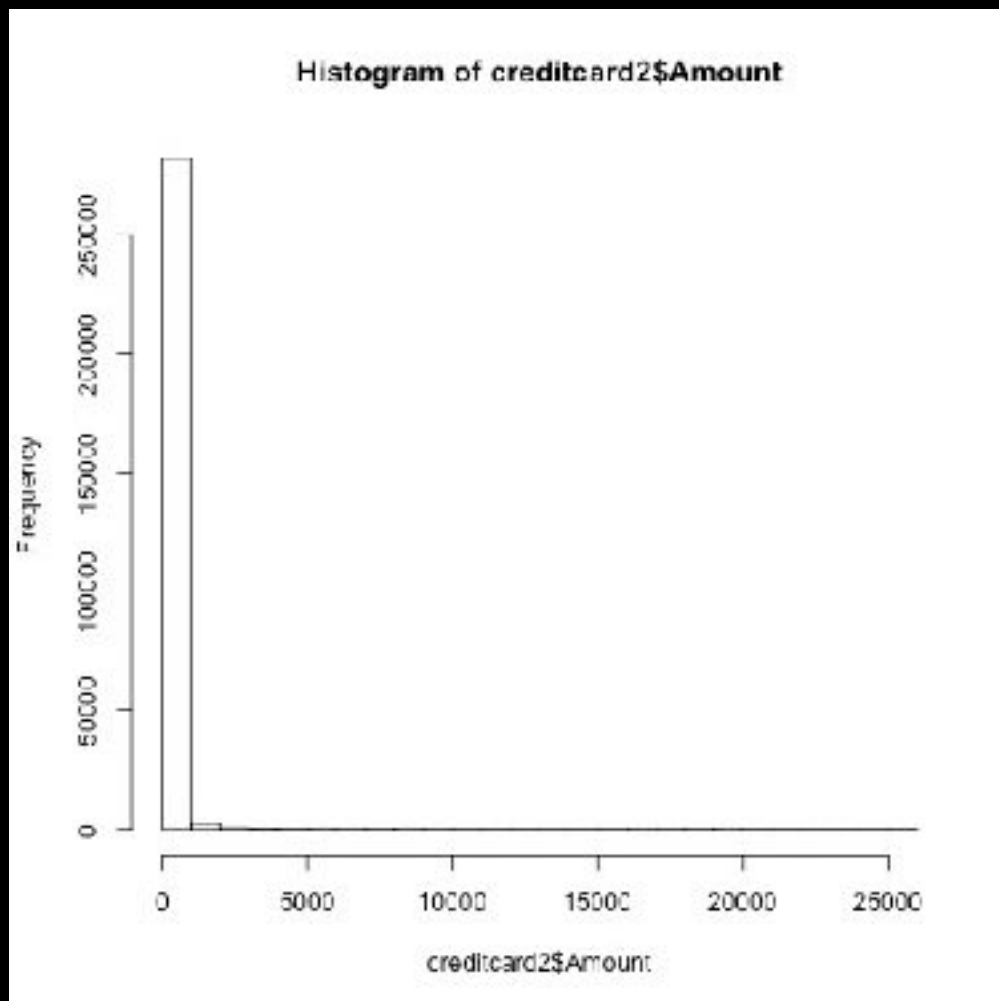
kNN은 쓸모없었고,
C&RT 모델이 가장 예측률이 좋았음.

데이터 가공

1. Time 컬럼 삭제
2. Amount 컬럼 정규화

Time 컬럼은 첫번째 트랜잭션부터 각 트랜잭션 사이의 경과시간임.
의미를 갖기 어려울 것으로 보아 삭제함.

Amount 컬럼은 거래금액을 나타냄. (0~25691) 원 데이터의 분포(왼쪽)와 로그값의 분포(오른쪽). 로그값을 취한 뒤 정규화함.
(금액이 0인 경우는 로그값을 구할 수 없을 뿐 아니라, 분석의 의미가 없다고 판단하여 전부 제외시킴)



분석모델(머신러닝 알고리즘)

1. KNN 알고리즘
2. C5.0 결정 트리 알고리즘
3. C&RT 결정 트리 알고리즘

KNN(k Nearest Neighbors): 유클리디안 거리가 가장 가까운 k개 데이터의 라벨 중에서 가장 많은 라벨로 분류. 분포가 편향되거나 스케일이 크게 차이날 때 분류가 어려움.

C5.0 알고리즘: 정보획득량(엔트로피) 을 기준으로 가지 생성.
Binary tree, multiway 가능.

C&RT(Classification&Regression Trees): Gini Index를 기준으로 가지 생성.
명목형, 숫자형 속성들을 모두 이용할 수 있음.
Cost-Complexity Pruning 이용, 불필요한 가지를 제거.
Binary tree만 가능.

알고리즘 별 결과 비교

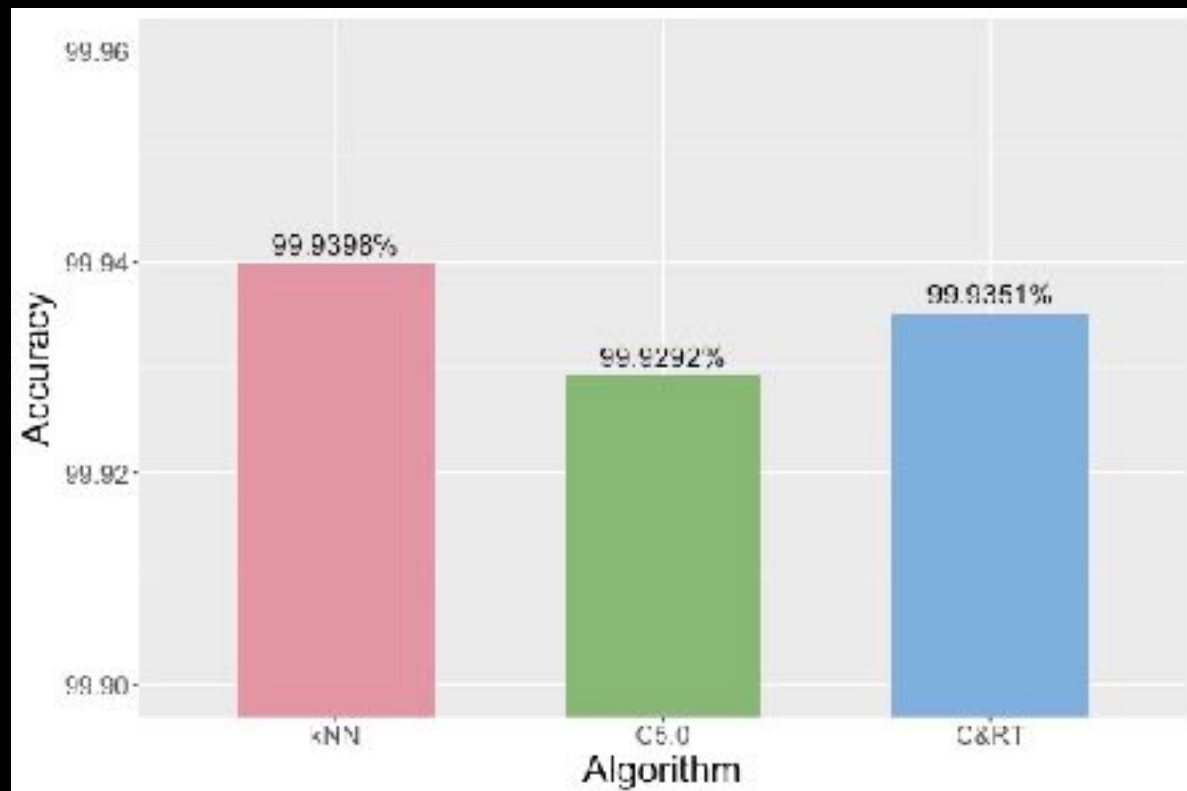
Cross Table 비교

kNN모델의 k=11

KNN/ C5.0/ C&RT	예측: 사기아님	예측: 사기
실제: 사기아님	84,748/ 84,751/ 84,745	18/ 15/ 21
실제: 사기	33/ 45/ 34	95/ 83/ 94

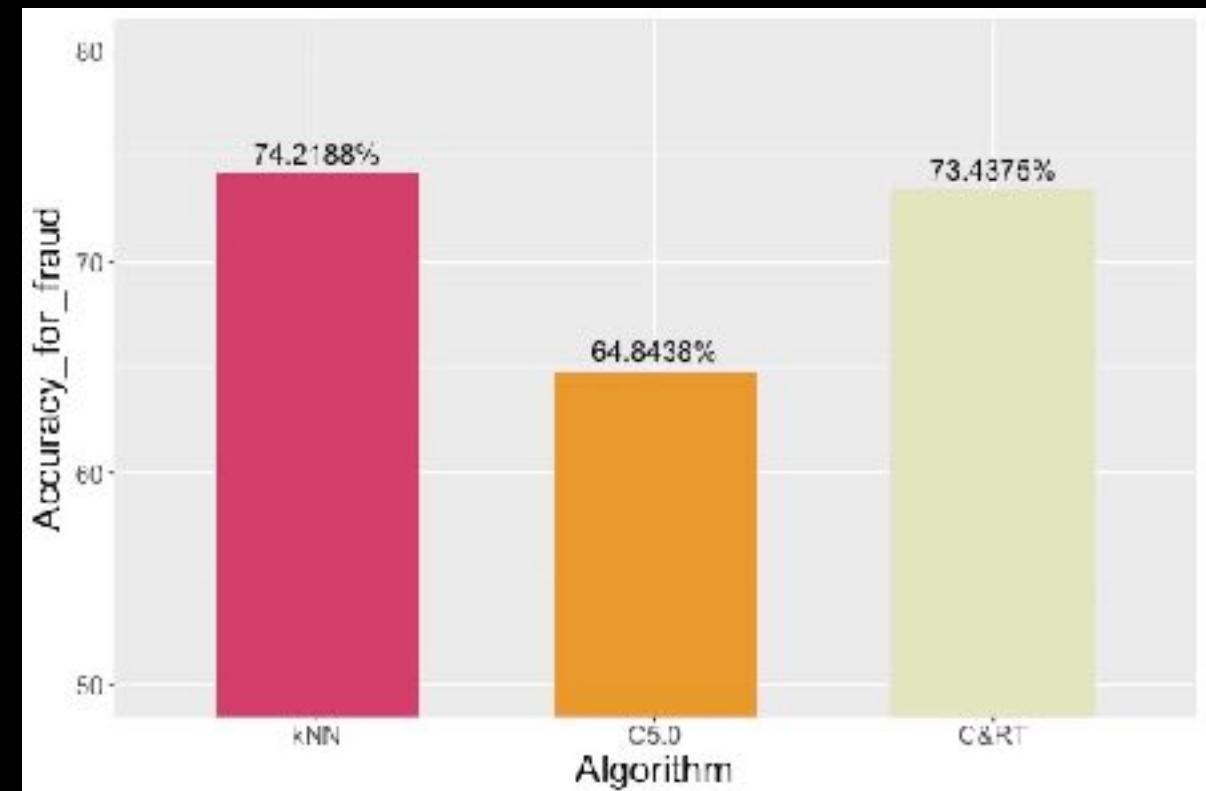
모델별 정확도

99.940%, 99.929%, 99.935%



예측률(실제 사기인 트랜잭션을 사기로 분류한 비율)

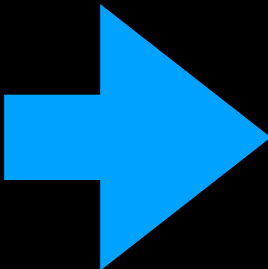
74%, 65%, 73%



데이터 가공 전후 결과 비교

C&RT 모델은 데이터 가공 전후 비슷한 정확도를 보임.
kNN, C5.0 모델은 가공 후 정확도 향상됨.

가공 전
1.19%
(168건 중 2건)



가공 후
74.22%
(128건 중 95건)

Cross Table 비교(%)

KNN/ C5.0/ C&RT	예측: 사기아님	예측: 사기
실제: 사기아님 (100%)	100/ 99.99/ 99.98	0/ 0.0082/ 0.0176
실제: 사기 (100%)	99.81/ 46.43/ 25.00	1.19/ 53.57/ 75.00

99.806%, 99.901%, 99.933%

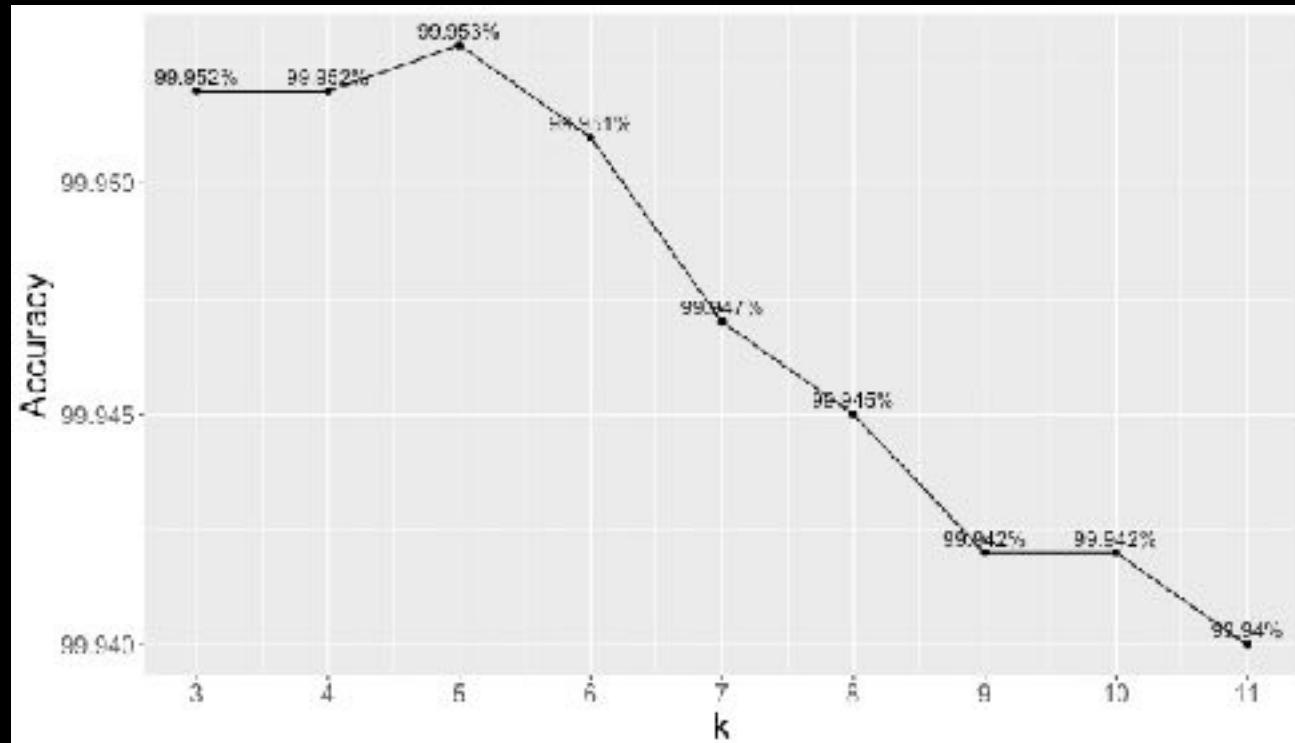
Cross Table 비교(%)

KNN/ C5.0/ C&RT	예측: 사기아님	예측: 사기
실제: 사기아님 (100%)	99.98/ 99.98/ 99.98	0.02/ 0.02/ 0.02
실제: 사기 (100%)	25.78/ 35.16/ 26.56	74.22/ 64.84/ 73.44

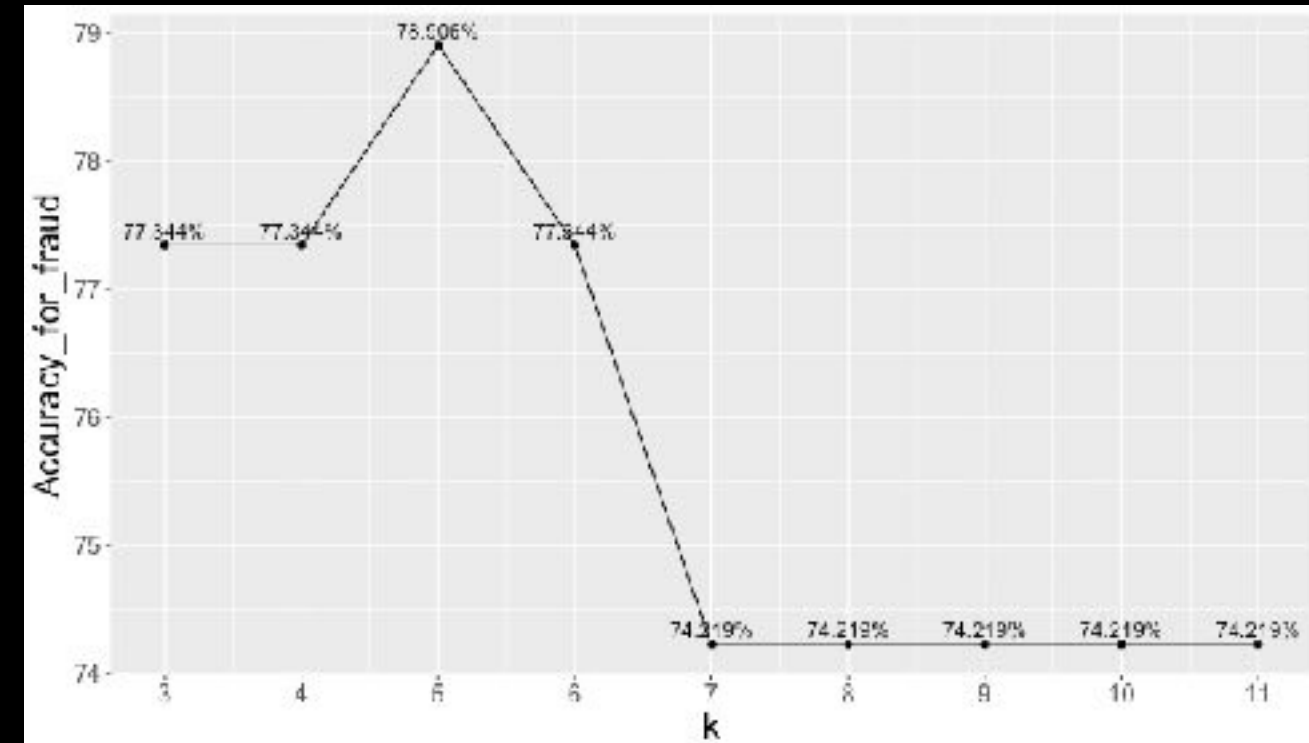
99.940%, 99.929%, 99.935%

kNN 모델 k의 최적값 도출

k값별 정확도(k=3~11)



실제 사기인 트랜잭션을 사기로 분류한 비율 (k=3~11)



Cross Table (k=5)

k=5	예측: 사기아님	예측: 사기
실제: 사기아님	84,753	13
실제: 사기	27	101

최적의 신용카드 사기 감지 모델

데이터 정제 및 k값(kNN알고리즘) 최적화 전
정확도: 99.806%, 99.901%, 99.933%
예측률: 1.19%, 53.57%, 75.00%

데이터 정제 및 k값(kNN알고리즘) 최적화 후
최적 모델(kNN, k=5)
정확도: 99.953%
예측률: 78.91%

Summary

정보가치가 없는 속성(Time) 삭제.

분석에 필요하지 않은 행(Amount=0) 삭제.

정규분포에 가까운 모양으로 정제하고 스케일을 조절(Amount 컬럼을 log화 후 정규화).

kNN 모델에 k=3~11을 적용하여 예측률이 가장 높은 k값을 찾음.