Machine Learning 기반
신용카드 이상 거래 탐지 알고리즘 연구

연구 목적

성인뿐만 아니라 학생들도 현금이 아닌 신용카드를 사용하는 시대에 이르렀다. 특히, 삼성페이, 애플페이 등 모바일 페이의 사용이 확장되면서 자신도 모르는 사이에 결제가 진행되거나 피싱 피해로 인해 고액 결제가 되는 문제가 생기고 있다. 신용카드 회사는 사기가 의심되는 신용카드 이상 거래를 탐지하여 고객이 구매하지 않은 항목에 대해 요금을 청구하지 않는 것이 중요하며, 이러한 문제를 해결하기 위한 사기 거래를 사전에 탐지할 수 있는 도구가 필요하다.
이에 조금이라도 도움이 되고자, 기술과 사회에서 배운 내용을 바탕으로 적절한 데이터 셋을 활용하여 기계학습 및 데이터분석을 이용하여 사기 거래 탐지 시스템을 구축하려 한다. 이러한 시스템은 다양한 데이터 소스를 사용해 거래를 모니터링하고, 머신러닝 알고리즘을 이용해 거래 이상 패턴을 탐지할 것이다.

데이터셋 소개

- 2013년 9월 유럽 카드 소지자가 신용카드로 만든 거래 이틀 간 284,807건의 거래 중 492건의 사기
- 데이터셋이 매우 불균형(사기는 거래의 0.172%)
- 변수에 PCA가 적용되어 있음.
- 기밀 유지 문제로 각 컬럼이 뜻하는 의미는 알 수 없음

데이터 전처리

[Class 피쳐 확인]

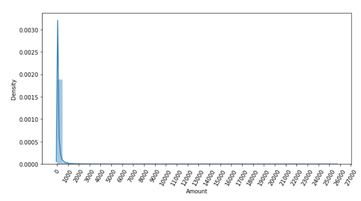
데이터셋의 사기(1) 비사기(0) 데이터가 매우 불균형한 것을 확인.

클래스 값 분포

0 284315
1 492
Name: Class, dtype:int64

[Amount 피쳐 확인]

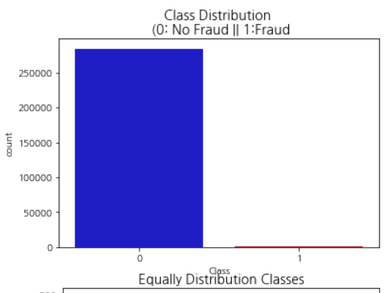
카드 사용 금액에서 1000달러 이하 사용자가 다수임을 확인함



[Train&Test 분리]

표본이 부족했던 데이터 프레임에 모델을 적용시키고 원래 테스트 데이터에서 테스트를 진행하기 위해 데이터 프레임을 분리함.
(Test Set이 전체의 30%가 되도록 분리)

전처리 전



전처리 후

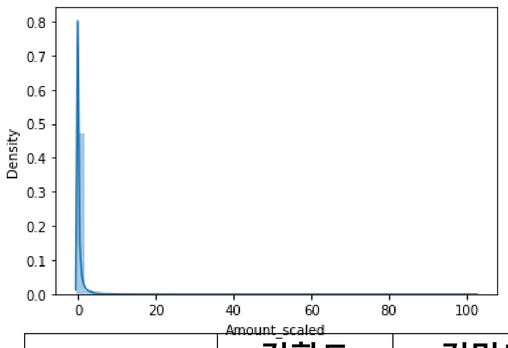
모델 학습/예측/평가

- 데이터 전처리 후의 **성능을 비교**해보기 위해 **원본 데이터를 모델을 학습시키고 예측하고 모델의 예측 성능 평가**
- 로지스틱 회귀 모델과 인기있는 앙상블 방법인 **LightGBM 모델** 사용

	정확도	정밀도	재현율	F1-score	AUC
Logistic Regression	0.9992	0.8762	0.6216	0.7273	0.9582
LighGBM	0.9995	0.9573	0.7568	0.8453	0.9790

데이터 분포도 변환

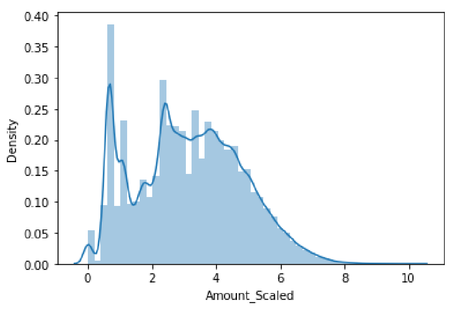
① 정규화



Amonut 피처값을 정규화 한 데이터 세트를 이용한 모델의 성능은 **원본 데이터를 이용한 모델에 비해 성능이 개선되지 않았음을 확인** (StandardScaler 사용)

	정확도	정밀도	재현율	F1-score	AUC
Logistic Regression	0.9992	0.8654	0.6081	0.7143	0.9702
LighGBM	0.9995	0.9569	0.7500	0.8409	0.9779

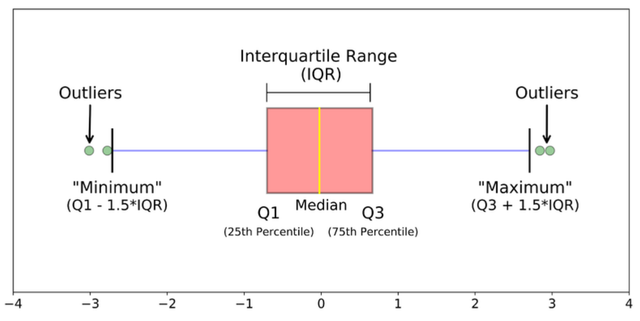
② 로그 변환



Amount 피처값을 로그 변환한 데이터 세트를 이용한 모델의 성능은 **원본 데이터를 이용한 모델에 비해 성능이 약간씩 개선되었음을 확인**

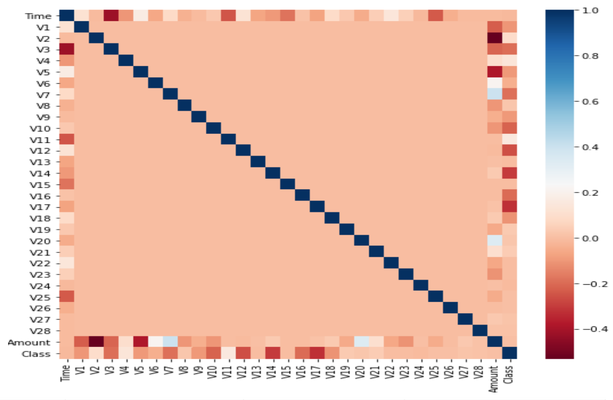
	정확도	정밀도	재현율	F1-score	AUC
Logistic Regression	0.9992	0.8812	0.6014	0.7149	0.9727
LighGBM	0.9995	0.9576	0.7635	0.8496	0.9796

이상치 데이터 제거



Q1-(IQR*1.5) 에서 Q3+(IQR*1.5) 구간을 벗어나면 **이상치 데이터로 간주**

음의 상관관계가 가장 높은 피처인 V14, V17에 대해 이상치 검출 & 제거



재현율이 로지스틱은 67%, LightGBM은 83%로 성능이 상당히 개선됨을 확인

	정확도	정밀도	재현율	F1-score	AUC
Logistic Regression	0.9993	0.8750	0.6712	0.7597	0.9743
LighGBM	0.9996	0.9603	0.8288	0.8897	0.9780

최종 결과

데이터 가공 유형	알고리즘	정밀도	재현율	ROC-AUC
원본 데이터(가공 x)	로지스틱	0.8762	0.6216	0.9582
	LightGBM	0.9573	0.7568	0.9790
정규화	로지스틱	0.8654	0.6081	0.9702
	LightGBM	0.9569	0.7500	0.9779
로그 변환	로지스틱	0.8812	0.6014	0.9727
	LightGBM	0.9576	0.7635	0.9796
이상치 데이터 제거	로지스틱	0.8750	0.6712	0.9743
	LightGBM	0.9603	0.8288	0.9780

불균형한 분포를 가진 피처를 로그 변환을 통해 불균형한 분포를 감소 시켜 주고 target값과 상관관계가 높은 피처의 이상치를 제거한 결과가 가장 best

연구의 미래 지향성

현재 한국뿐만 아니라 전 세계 모든 신용카드 회사들도 자체적으로 이상거래를 탐지하고 있는 시스템을 운영하고 있다. 하지만, 이와 같은 정보들은 영업비밀이기 때문에 오픈되어 있는 정보가 아니라 알 수 없다. 하지만 이 연구가 그런 시스템의 개발 등에 기여할 수 있으리라고 생각한다. 실제 데이터셋으로 비교, 분석해본 결과에 따르면 기존 시스템의 ML기법 부분에서 보완할 수 있으리라 생각한다.

참고 자료

https://www.kaggle.com/janiobachmann/credit-fraud-dealing-with-imbalanced-datasets
신용 사기 탐지기(Credit card fraud detection) 데이터셋 활용