거래 데이터를 이용한 이상거래 탐지

- 데이터 분석과 탐지 방법론을 중심으로

최윤주

2024. 06

목차

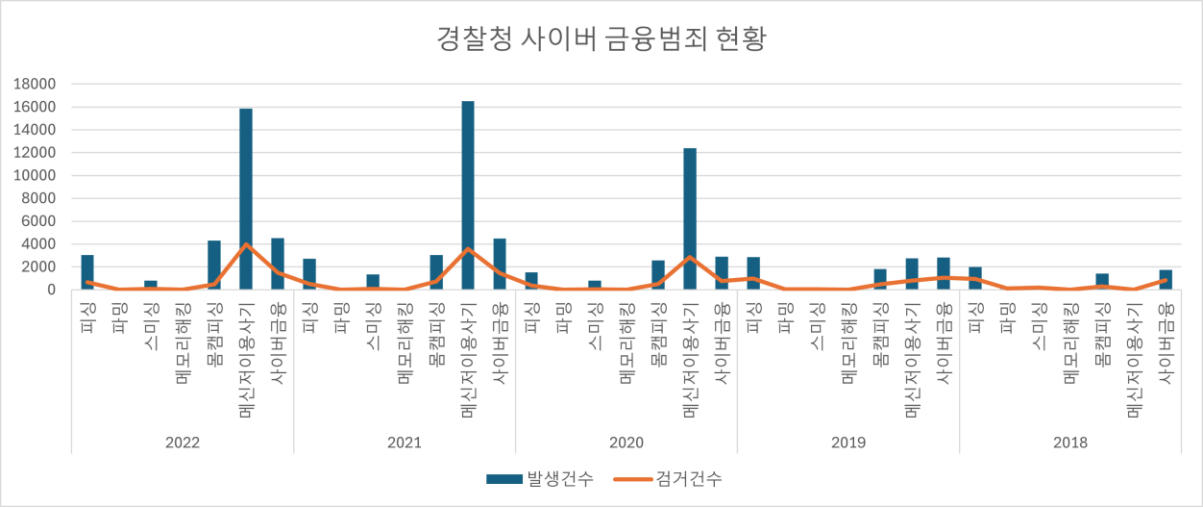
1. 프로젝트 목적 및 필요성

이상거래 탐지 시스템, FDS(Fraud Detection System)은 금융 거래 시에 결제자의 빅데이터를 통해 패턴을 만들고 정상적인 패턴과 다른 패턴을 보이는 이상 결제를 탐지하면 결제 경로를 차단하는 금융 거래 보안 방식을 말한다. 이상 거래 탐지는 20 년전 신용카드가 보편화되기 시작할 때 신용카드 사기를 방지하기위해 각 기업들이 도입을 시작했다. 특히 2010 년대에 들어서면서 빅데이터, 인공지능, 머신 러닝과 같은 it 기술의 발전에 힘입어 FDS 은 더 정교해졌다.

FDS 의 진화와 더불어 사기 수법도 병렬적으로 발전 중이다. 특히 온라인 기반 서비스에 경우 접근이 용이하기 때문에 사기에 더 취약하다. 코로나 이후, 많은 사람들이 집에서 금융 업무를 하게 되면서 디지털 금융 거래의 일상화는 가속화되었다. 이러한 환경 변화는 이상거래 탐지 시스템 (FDS)의 필요성을 더욱 부각시킨다. 카카오뱅크, 토스 등 비대면 계좌 개설 서비스들도 상용화되면서 사기 피해 감지 시스템과 이를 피하려는 사기 수법은 나날이 발전하고 있다.

따라서 금융 기관들은 더욱 정교하고 효과적인 FDS 알고리즘을 개발할 필요가 있다. 금융감독원과 같은 정부 기관은 사기 방지를 위한 지침을 강화하며, 금융 기관의 FDS 구축을 적극적으로 지원하고 권장하게 되었다. FDS 방법론은 기업 비밀로 기밀하게 논의할 정도로 현재 매우 중요한 분야 중 하나이다. 이를 통해 금융 사기 피해를 최소화하고, 안전한 금융 환경을 구축하는 것은 금융 기관의 지속 가능한 성장과 고객 신뢰 확보를 위한 필수 요소라고 할 수 있다.

공공데이터 포털에서 제공하는 경찰청 사이버 금융범죄 현황에 따른 금융 범죄 증가 추이와 검거현황은 다음과 같다.



사이버 금융 피해 건수는 매년 늘어나는데 검거 건수는 그에 한참 미치지 못하는 것을 알 수 있다. 사기 수법도 다양하다. 다양한 수법에 관계없이 빠르게 사기를 탐지하고 대응하는 것이 2 차피해를 막을 수 있는 방법이다. 피해가 발생하면 자산을 다시 회수하는 것은 또 다른 어려운 문제기 때문에 FDS 를 통해 사기를 감지하고 거래 차단이나 실제 고객이 사용한 것이 맞는지 등 직접 연락을 통해 확인해야 한다.

이 프로젝트의 목적은 거래 빅데이터를 통해 사기 거래 탐지 시스템을 만들고 성능을 테스트 데이터 값으로 검증해 보는 것을 목적으로 한다. 실제 사용되고 있는 FDS 는 고객의 거래 시간, 거래 발생 위치, 거래 금액과 유형(온/오프라인 거래)뿐만 아니라, 기존 통계 데이터를 이용해 위험도 높은 계좌를 집중적으로 모니터링하여 이상 거래를 탐지한다. 이와 상응하는 데이터셋을 통해 금융 피해를 최소화할 수 있고, 범죄 예방에도 효과가 클 것으로 기대되는 시스템을 구축하는 것이 이 프로젝트의 목적이다.

2. 선행연구 분석

최신 FDS 연구 논문의 초록을 웹 크롤링을 통해 수집하여 텍스트 마이닝을 진행한다. 어떤 데이터가 활용되고 있고, 연구 방법 동향을 파악하는 것이 주 목적이다. 수집한 대용량 언어 데이터를 토큰화하고 한국어 기준으로 불용어를 제거하는 과정을 거쳐 텍스트 마이닝을 수행했다. 그리고 단어 빈도 수 분석을 사용하여 논문의 집합으로부터 어떤 단어가 빈번하게 제시되고 있는지 확인하고 이를 시각화(워드 클라우드) 했다.

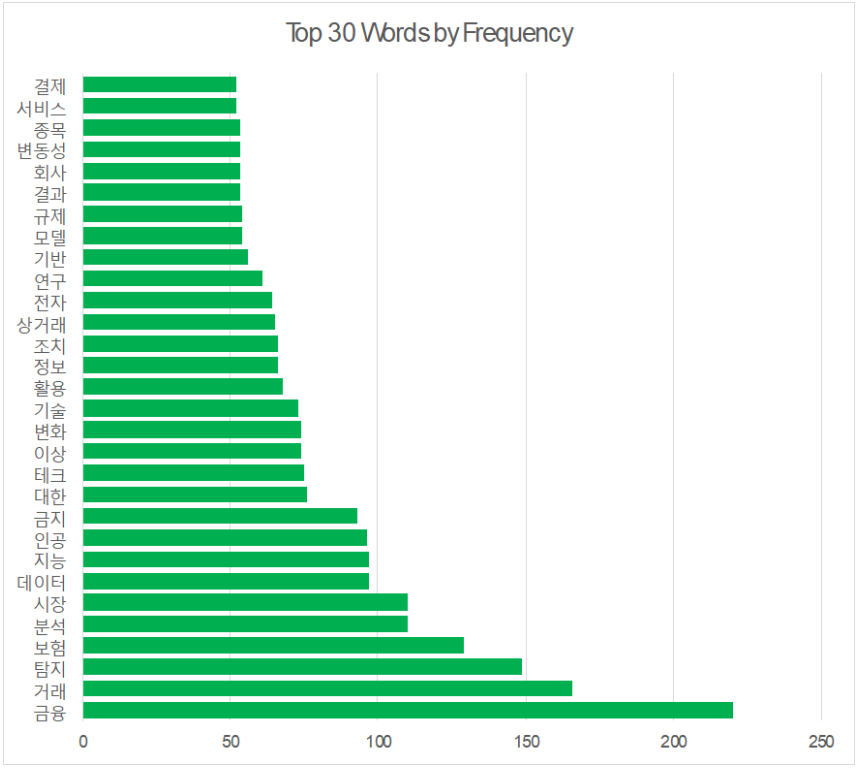
1) 웹 크롤링으로 논문 초록 수집

논문 검색 사이트로는 RISS 를 선택했다. “이상 거래 탐지”라는 키워드가 들어가는 국내 학술 논문 초록 총 82 개를 수집했다. Python 의 BeautifulSoup 및 Selenium 라이브러리를 활용하여 웹 크롤링을 통해 논문 초록을 수집하고, 이를 CSV 파일로 저장했다.

	A	B	C	D
1	Mains	초록	작가명	발간년도
2	순차패턴	정보통신기술의 발달로 전자금융서비스가 활성화됨에	최병호(Byu	2021
3	머신러닝을	전자금융서비스가 활성화됨에 따라 전자금융 거래 건수	최병호(Byu	2022
4	내부자 보	기존의 전자금융 이상거래 분석 및 탐지기술은 전자금융	이재용(Jae	2018
5	전자금융거	본 논문은 금융 사용자의 거래 행태를 반영한 이상거래	최의순(Eui	2015
6	생성적 적대	인공지능이 다루기 어려운 개념에서 아주 익숙한 도구로	김예원	2020
7	전자금융	금융회사가 전자금융 서비스를 제공하기 시작하면서 전	유시완(Si-w	2016
8	생성적 적대	신경망과 딥러닝을 활용한 이상거래탐지 시스템 모형		2019
9	피싱 금융	전자금융 사기범이 전화, SMS, 이메일을 통하여 통신회	김정선(Kim Jung Sun)	
10	Special Report - 핀테크 보안 - 이상거래탐지 동향 및 전망			2013
11	의사결정나	전자금융사기의 고도화와 함께 지능적인 수법들이 동원	박재훈(Jae	2017

2) 텍스트 전처리 및 빈도 시각화

위 CSV 파일의 "초록" 칼럼을 토대로 텍스트 전처리와 word frequency 를 수행하였다.



이를 워드 클라우드로 시각화 한 결과는 다음과 같다.

3) 연구 동향 및 방법론 제시

이상 거래 탐지에 대한 연구 수집에서 “금융”, “거래” 등 연구 주제 상 밀접한 관련이 있는 단어를 제외하면 인공지능을 활용한 연구가 활발히 이뤄지고 있다는 것을 알 수 있었다. 또 “(핀)테크”, ”기술”, ”상거래” 등 it 발전에 따른 연구가 진행되었음을 볼 수 있다. 또 ”보험” 빈도 수를 보면 보험사기에 대한 FDS 연구가 많이 이뤄지고 있다는 것을 알 수 있었다.

이 결과를 토대로 인공지능을 활용한 분석 기법을 사용하기로 했다.

3. 데이터 분석

1) 데이터소개

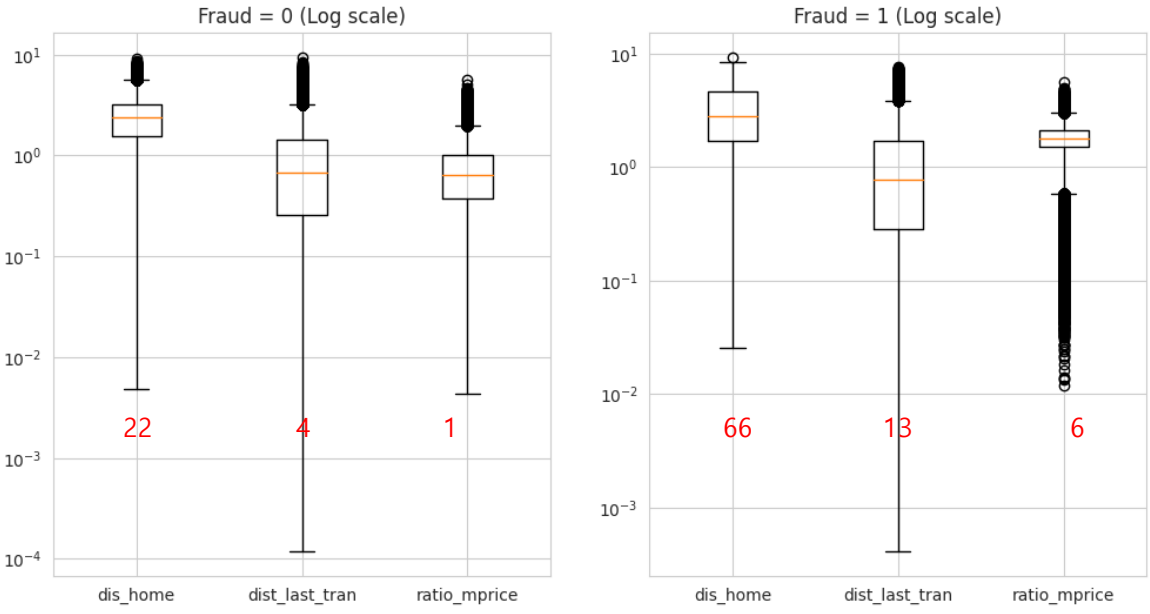
kaggle 의 credit card fraud 데이터를 선정하여 FDS 을 만들어 보도록 한다. 칼럼은 다음과 같다: distance_from_home, distance_from_last_transaction, ratio_to_median_purchase_price, repeat_retailer, used_chip, used_pin_number, online_order, fraud(label). 정리하면 주거지와 거래 장소와의 거리, 마지막 거래장소와의 거리, 이전 결재 금액 중앙값과의 비, 이전에 결제한적 있는 소매업체에서 발생한 거래인지, 신용카드 칩을 사용한 거래인지, PIN 번호를 사용한 거래인지, 온라인 거래였는지, 거래가 사기였는지에 대한 정보를 제공한다.

2) EDA

EDA(Exploratory Data Analysis, 탐색적 데이터 분석)은 데이터를 수집한 후 초기 단계에서 수행하는 분석 방법이다. 데이터의 구조, 이상치, 결측치, 패턴 및 변수 간의 관계를 이해하기 위해 통계적 방법을 사용한다. EDA 의 주 목적은 데이터에 대한 통찰을 얻고, 데이터의 특징을 파악하여 이후의 분석이나 모델링 전략을 설정하는 데 도움을 얻는 것이다.

먼저 데이터 타입과 결측치 등 기본적인 정보를 확인하고, 기초통계량을 확인해 보도록 했다.

모든 변수의 데이터 타입은 float64 로 따로 string 등의 데이터 타입을 처리할 일은 없었다. 결측치는 모든 데이터 column 에서 0 개로 결측치 처리 방법은 필요하지 않았다. 총 데이터의 수는 1,000,000 개였다.



기초 통계량을 데이터 분포를 알아보기 쉽도록 로그 스케일 후 박스그래프로 시각화 해보았다.

주거지와 거래 장소와의 거리는 정상 거래는 평균 약 22 키로미터이고 사기 거래의 경우 평균 약 66 킬로미터로 차이가 3 배였다. 중앙값 또한 정상거래는 약 9 키로미터, 사기거래는 약 15 로 역시 차이가 존재하는 것을 확인 가능하다. 이를 통해 사기 거래의 경우 주거지와 멀리 떨어진 곳에서 주로 발생한다는 것을 알 수 있다.

박스 그래프의 dis_home(주거지와 거래 장소와의 거리) 변수를 보면 사기 거래(fraud=1)인 경우 정상 거래(fraud=0)인 경우보다 집에서의 거리가 일반적으로 더 멀게 나타나는 것을 확인 가능하다.

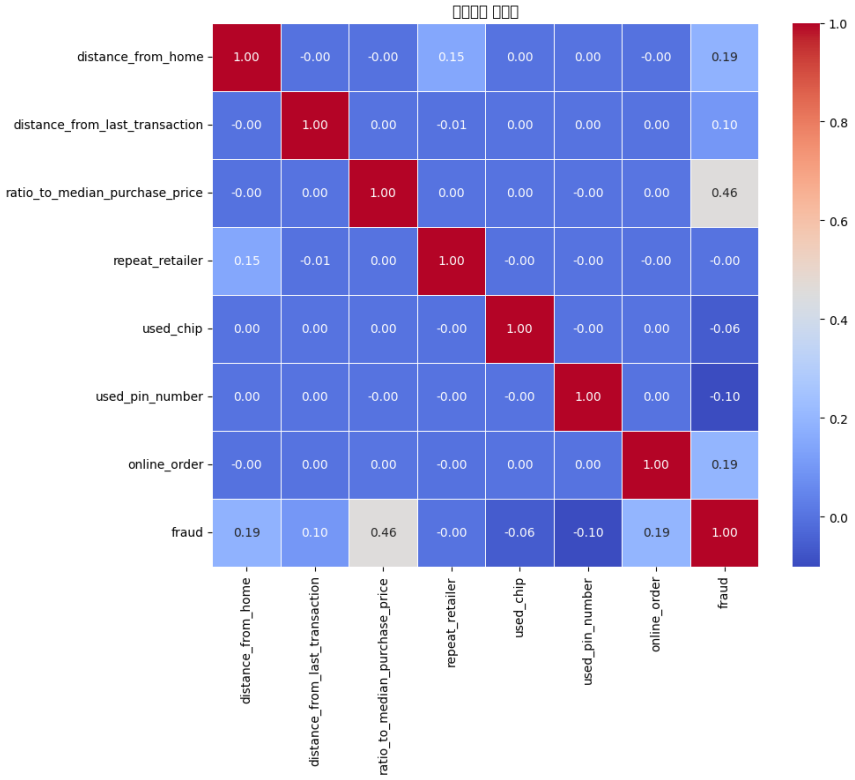
마지막 거래장소와의 거리는 정상 거래의 경우 평균 4.3, 사기 거래는 평균 12.7 킬로미터로 3 배정도의 차이가 있었다. 중앙값의 경우 정상 거래는 약 0.9, 사기 거래는 약 1.1 키로미터로 큰 차이가 존재하지 않았다. 이를 통해 사기 거래는 마지막 거래 장소와 멀리 떨어지지 않은 장소에서도 꽤 일어나지만 평균적으로는 마지막 거래 장소와의 거리가 먼 곳에서 일어남을 알 수 있다.

박스 그래프의 dist_last_tran(마지막 거래장소와의 거리)변수를 보면 사기 거래(fraud=1)인 경우 정상 거래(fraud=0)인 경우보다 마지막 거래 장소와의 거리가 먼 경향을 파악 가능하다. 즉 정상 거래인 경우 마지막 거래 위치와의 거리가 짧다고 볼 수 있다.

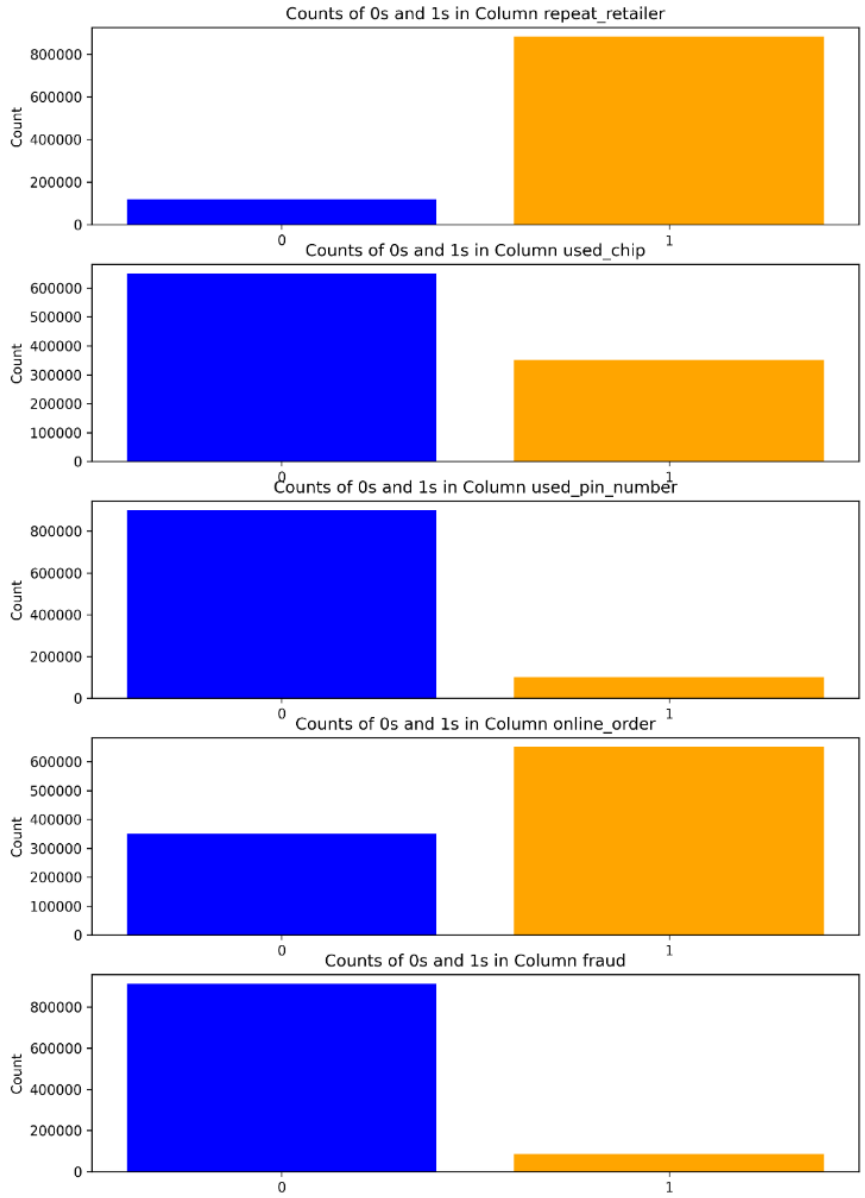
결재 금액 중앙값과의 비는 정상 거래의 경우 평균 1 로 평소에는 비슷한 금액대를 결제한다는 것을 알 수 있고 사기거래의 경우 6 으로 카드 소유자의 결제 데이터의 중앙값 금액 보다 약 6 배가량을 탈취했다는 것을 알 수 있다.

ratio_mprice(결재 금액 중앙값과의 비) 변수의 박스 위치를 주목하면 사기거래(fraud=1)와 정상거래(fraud=0)의 차이를 한눈에 알아볼 수 있다. 중앙값 구매 가격에 비해 큰 금액의 거래 시, 사기 거래임을 특히 의심할 필요가 있다는 것을 시사한다. 또 상자의 크기도 작은 것을 보아 데이터가 특정 구간에 몰려 있음을 알 수 있다. 따라서 방법론을 통해 모델 구축 시 꼭 사용해야 할 중요 변수이다.

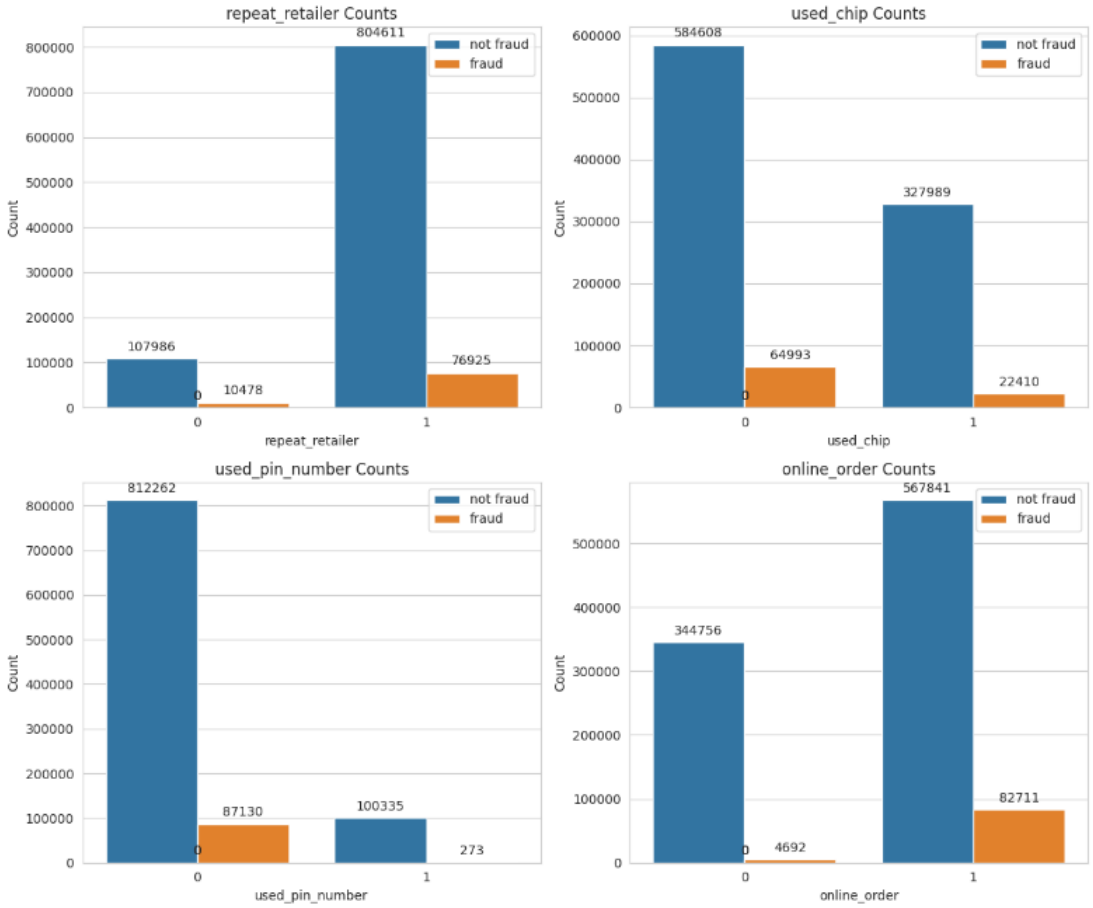
보통 데이터 전처리 과정에서 이상치를 탐지 후 제거하여 모델 과적합을 예방하지만, 이상거래 탐지에는 이상치가 중요 정보를 포함하고 있을 가능성이 높으므로 이상거래의 이상치 제거는 생략했다. 대신 정상 거래의 이상치는 제거하여, 모델 학습 시 일반적인 정상 거래의 패턴을 더 잘 학습하도록 하였다.



위는 변수 간 상관행렬 히트맵이다. 빨간색일수록 변수간 선형성이 크고, 파란색일수록 변수 간 선형성이 적다고 판단한다. 상관행렬 값이 0 인 경우, 두 변수 사이에 선형적인 관계가 없다고 할 수 있다. 데이터 셋이 연속형과 범주형 변수를 모두 포함하기 때문에 그래프에서 볼 수 있듯 fraud 열의 범주형 변수의 상관행렬 값은 모두 0 에 가깝다. 양적 변수를 봐도, 상관계수가 0.19, 0.10, 0.46 이며 선형성이 비교적 낮기 때문에 선형모델은 적용할 수 없다는 결론을 도출해낼 수 있다.



각 범주형 데이터의 수를 시각화한 것은 위와 같다. 사람들은 대부분 구매 이력이 있는 상점에서 쇼핑하는 것을 알 수 있다. 신용카드 칩을 통한 거래는 카드 복제가 용이한 마그네틱 스트라이프로 결제하는 것보다 안전하지만 칩으로 거래하는 사람보다 거래하지 않는 사람의 수가 두배정도 높았다. Pin number, 개인식별번호는 거래 시 신원을 인증할 수 있으므로 일반적으로 더 안전하지만 사용하는 비중이 매우 낮았다. 또 온라인 거래의 수는 오프라인 거래 수보다 2 배가량 높았다. 마지막으로 사기거래(87403 개) 데이터보다 정상거래 데이터(912597 개) 수가 월등히 많다는 것을 알 수 있다. 전체 거래의 수에 비하면 사기거래 수가 적은 것은 당연하지만 방법론 적용 시, 불균형 데이터에 대한 처리가 필요함을 알 수 있다.



각 데이터를 다시 정상거래(blue), 사기거래(orange)로 나눠본 것은 위와 같다. 이전에 방문한 적 있는 소매업체 인지에 대한 변수에서 이전에 방문한적 없는 소매업체의 정상: 사기거래 비율은 약 10:1 로 방문한적 있는 경우에도 약 10:1 이었기에 큰 차이가 없었다.

신용카드 칩을 사용했는지에 대한 변수를 보면 칩을 사용하지 않은 거래의 경우 정상: 사기거래 비율은 약 8:1 로 신용카드 칩을 사용한 경우의 정상: 사기거래 비가 약 14:1 이었기에 신용카드 칩을 사용한 경우 사기 거래의 수는 적음을 확인할 수 있다.

핀 번호를 사용했는지에 대한 변수를 보면 핀 번호를 사용하지 않은 거래의 경우 정상: 사기거래 비율은 약 9:1 로 핀 번호를 사용한 경우의 정상: 사기거래 비가 약 367:1 이었기에 핀 번호를 사용한 경우 사기 거래의 수는 매우 적음을 확인할 수 있다.

온라인 거래였는지에 대한 변수를 보면 오프라인 거래의 경우 정상: 사기거래 비율은 약 73:1 로 온라인 거래의 정상: 사기거래 비가 약 6:1 이었기에 오프라인 거래보다 온라인 거래가 사기거래가 매우 많음을 확인 가능하다.

방금 전 그래프의 검증을 위해 추가적인 데이터 분석 방법으로 카이 제곱 검정을 사용한다. 카이 제곱 검정은 두 범주형 변수에 대한 분석법으로 두 변수 간 관련성 여부를 파악 가능하다. 범주형 데이터들이 fraud 변수와 관련성이 있다고 볼 수 있는지 분석해 보도록 한다. 위에서 확인했듯 종속변수(fraud)의 불균형 문제가 있어 카이 제곱 통계량이 과대평가될 가능성이 있다는 것은 염두 해야 한다.

```
Chi-square Statistic: 0.4746628733585097
P-value: 0.4908498137021666
Degrees of Freedom: 1
Expected Frequencies:
 [[11824.23894916  1124.76105084]
 [88592.76105084  8427.23894916]]
There is no significant association between the two variables.
```

repeat_retailer 의 카이 제곱 검정 결과는 위와 같고 카이 제곱 통계량이 매우 작음으로 사기 거래와의 연관성이 없다고 볼 수 있다. 따라서 추후 방법론을 사용하여 모델 구축 시 이 변수는 제외하도록 한다.

```
Chi-square Statistic: 3717.4490433572664
P-value: 0.0
Degrees of Freedom: 1
Expected Frequencies:
 [[592823.923797  56777.076203]
 [319773.076203  30625.923797]]
There is a significant association between the two variables.
```

used_pin_number 의 카이 제곱 검정 결과는 위와 같고 카이 제곱 통계량이 매우 높음으로 사기 거래와의 강한 연관성이 있다고 볼 수 있다. 모델 구축 시 이 변수를 포함한다.

```
Chi-square Statistic: 10057.412546099067
P-value: 0.0
Degrees of Freedom: 1
Expected Frequencies:
 [[820782.441024  78609.558976]
 [ 91814.558976   8793.441024]]
There is a significant association between the two variables.
```

Used_chip 의 카이 제곱 검정 결과는 위와 같고 카이 제곱 통계량이 매우 높음으로 사기 거래와의 강한 연관성이 있다고 볼 수 있다. 모델 구축 시 이 변수를 포함한다.

```
Chi-square Statistic: 36852.02374794533
P-value: 0.0
Degrees of Freedom: 1
Expected Frequencies:
 [[318905.196456  30542.803544]
 [593691.803544  56860.196456]]
There is a significant association between the two variables.
```

online_order 의 카이 제곱 검정 결과는 위와 같고 카이 제곱 통계량이 매우 높음으로 사기 거래와의 강한 연관성이 있다고 볼 수 있다. 모델 구축 시 이 변수를 포함한다.

카이 제곱 검정에서 확인한 각 범주형 변수와 사기거래의 연관성 정도는 online_order>Used_chip>used_pin_number>repeat_retailer 로 그래프의 비율만 확인했을 때와는 다르게 각 변수 간 사기거래와의 정확한 연관 순서를 얻을 수 있었다.

4. 방법론

앞서 EDA 를 통해 이 데이터셋은 3 개의 연속형 변수 그리고 4 개의 범주형 변수(0,1 값의 이진 값)와 라벨(0: not fraud, 1: fraud)데이터로 이루어져 있다는 것을 알았다. 이를 바탕으로 사용할 알고리즘은 범주형 변수 처리에 적합해야 한다는 조건이 필요하다. 예를 들어 연속형 변수 처리에 알맞은 PCA 를 이 데이터셋에 사용한 결과 모든 주성분의 분산이 거의 동일했기 때문에 차원 축소에 어려움이 있었다. 또 클러스터링(k-means)의 경우 특징 벡터의 평균값을 기준으로 군집을 생성한다. 그런데 범주형 변수의 경우 인코딩을 통해 특징을 0,1 로 나타내는데 평균을 사용하면 범주형 범주의 의미가 유실되어버린다.

이를 통해 데이터 성격에 따라 적용가능한 모델은 정해져 있다는 것을 알 수 있다. 그리고 사용할 데이터셋에는 라벨 값(fraud)이 있으므로 지도학습을 사용하기로 했다. 따라서 이번 FDS 구축에 사용할 방법론은 K- Nearest Neighbors, 랜덤 포레스트, 인공신경망이다. EDA 를 통해 알아본 데이터셋의 성격에 알맞은 방법으로 이들은 독립변수가 연속형, 범주형이 혼합 되어있을 때 이를 포괄할 수 있는 분석 능력을 지닌다고 알려져 있다. 또 종속변수(Fraud)가 더미변수인 것도 고려했다.

데이터 전처리로는 양적 변수에 한하여 정규화를 했다. 또 데이터 분포가 편향되어 있으면, 모델이 적은 데이터를 갖는 클래스보다는 많은 데이터를 갖는 클래스를 더 잘 예측하는 방향으로 학습이 이루어지는 문제가 있다. 지도학습의 경우 데이터셋이 많은 라벨의 경향성대로 학습하기 때문에 데이터 샘플링을 거쳤다. Fraud 값이 불균형하기에(정상거래의 수가 훨씬 많은 문제) 오버 샘플링을 통해 정상거래와 사기거래 클래스 간의 비율을 맞춰 줌으로서 이를 조정했다. 이때 트레인 세트만 오버 샘플링을 사용하고 테스트 데이터는 그대로 놔둬야 올바르게 검증할 수 있다.

성능 평가 지표로는 F1-Score 을 사용한다. F1-Score 는 분류 모델의 성능을 평가하는 지표 중 하나로, 모델의 정밀도(Precision)와 재현율(Recall) 간 조화 평균 산식을 사용한다. 정밀도와 재현율 모두 중요한 경우에 F1-Score 를 사용한다. 정밀도는 모델이 True 로 예측한 것 중 실제로 True 인 비율이다. 재현율은 실제 True 인 것 중 모델이 True 로 올바르게 예측한 비율이다. F1-Score 는 정밀도와 재현율이 모두 높을 때 높은 값을 가지고, 정밀도와 재현율 중 한가지라도 값이 낮으면 결과값이 낮게 나타난다. 이러한 특징 때문에 균형 잡힌 성능지표로 널리 활용된다.

정밀도와 재현율을 더 설명하기 앞서 정확도에 대해 설명해야 한다. 정확도란 단순히 예측 데이터 건수에서 모델이 실제로 예측에 성공한 데이터 건수를 뜻한다. 이진분류 데이터의 경우 모델 성능을 평가할 때 정확도만을 사용하기에는 불충분하다. 그 이유는 데이터셋의

사기거래와 정상거래의 비가 다를 수 있기 때문이다. 이 데이터셋 같은 경우 정상거래는 전체의 약 90%, 사기거래는 전체의 약 10%이기 때문에 예측 데이터 값만으로 평가를 했을 때 무조건 정상 거래라고만 답해도 정확도가 90%이 나올 수 있다.

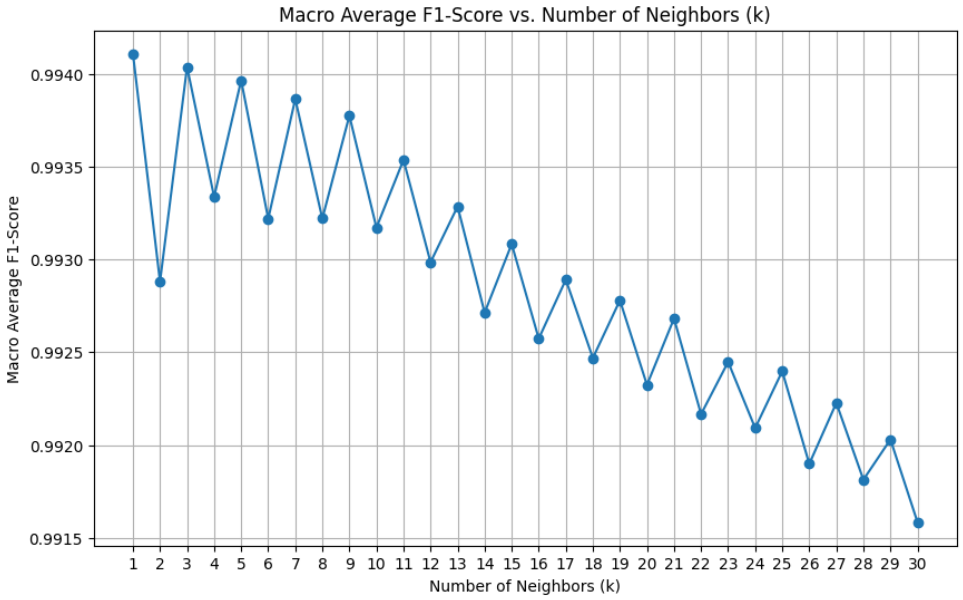
정밀도와 재현율은 이러한 맹점을 극복하고자 오차 행렬을 도입한다. 오차 행렬은 예측 클래스와 실제 클래스 간 경우의 수로 TP(true positive), FP(false positive), FN(false negative), TN(true negative)인 경우로 분류 가능하다. 여기서 모델이 예측을 성공한 TP, TN 인 경우뿐만 아니라 FP, FN 와 같이 실패한 경우를 주목하여 모델 성능을 세밀하게 파악할 수 있다.

이 프로젝트에서 사용한 방법론에서는 원 데이터셋을 트레인셋과 테스트셋으로 8:2 로 나눈 후, 트레인셋으로 모델을 구축하고 테스트셋은 f1-score 를 통해 모델의 성능을 평가할 것이다.

1) K- Nearest Neighbors

K-means 는 특징벡터의 평균값을 기반으로 한 군집 생성 비지도 학습이고 K- Nearest Neighbors 는 특징벡터의 위치를 기반으로 가까운 포인트의 클래스를 참조하는 분류 지도학습이다. 두 방법론은 관련이 없다. K-means 는 평균값을 기반으로 군집을 형성하며 연속형 데이터로만 이루어진 데이터셋에 적절하기에 프로젝트 초반에 기각한 방식이다. 비슷한 군집 생성 알고리즘이고 실제 데이터셋에 적용 가능한 K-prototype 방법을 적용했을 때의 성능은 f1-score 0.3 정도로 좋지 못했다. 군집 생성 알고리즘에서 분류 알고리즘으로 대체 이유는 프로젝트 주제가 이상거래 탐지이고 라벨 값이 있는 데이터를 사용하기에 KNN 을 사용하는 것이 모델의 정확도도 높이고 주제에 부합한다고 생각했기 때문이다.

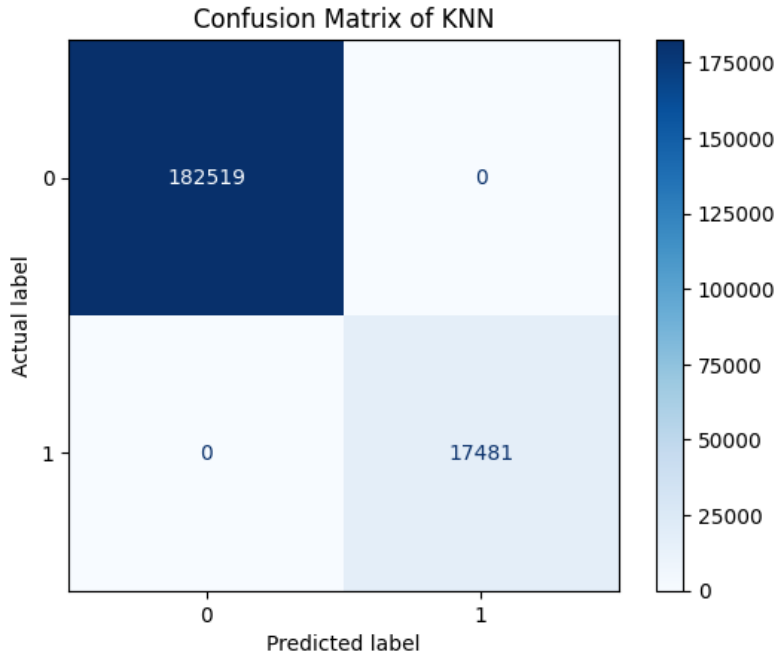
KNN 은 분류하려는 새로운 데이터 포인트와 기존 데이터 포인트들 간의 거리를 측정하고 측정된 거리를 바탕으로 가장 가까운 K 개의 이웃을 선정한다. 그 후 분류 작업에서는 K 개의 가장 가까운 이웃 중 다수가 속한 클래스를 새로운 데이터의 클래스로 예측한다. 이 알고리즘에서는 근접 이웃을 몇 명 참고할 것인지 정해야 하는데 이는 각 k 개 이웃별로 성능을 검증해보면 가능하다. 성능 측정 지표로는 f1-score 를 사용했다.



K=1 즉 근접 이웃을 1 개 참고했을 때 성능이 가장 좋았던 곳으로 보이며 짝수에서 성능이 낮게 나오는 것은 이웃의 클래스가 정상과 사기의 비율의 1:1 일 때 제대로 분류를 수행하지 못함에 있다. 또 이 프로젝트에서는 5-Fold Cross-Validation 을 사용하여 데이터 세트를 5 개의 동등한 부분으로 나눈 후 한 부분은 검증 세트로 사용하고 나머지 4 개의 부분을 훈련 세트로 사용하여 각 부분에 대해 순차적으로 모델 학습을 진행했다. 이를 통해 모델의 일반화 능력을 더 정확하게 평가 가능하다.

KNN의 F1-Score: 1.0000000000000000

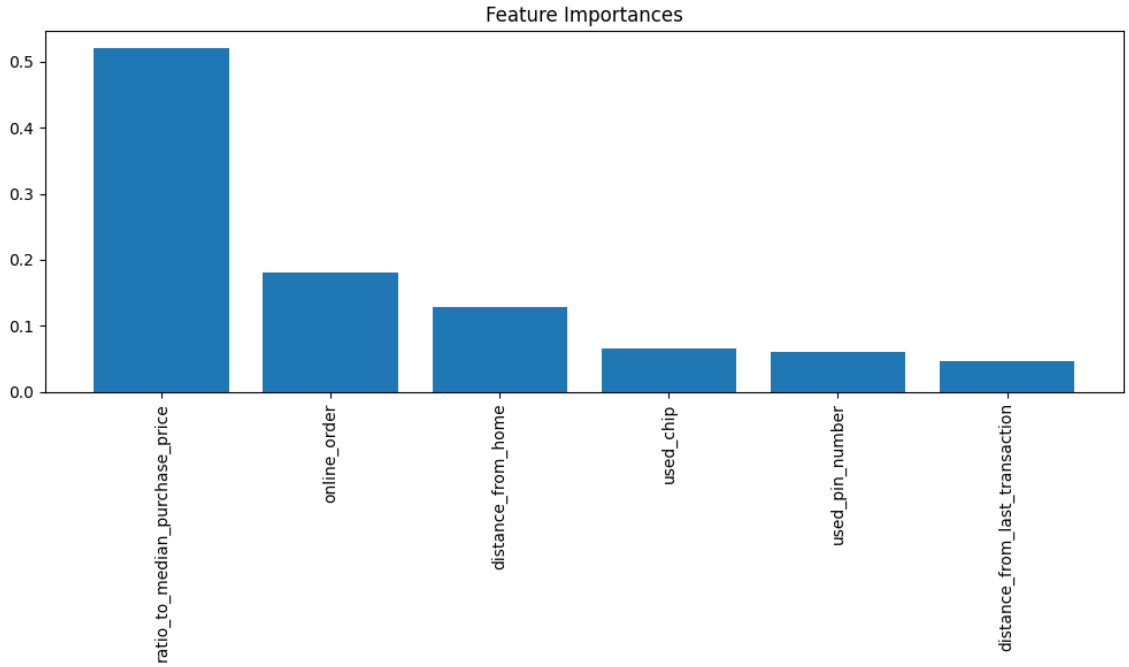
위는 최종 모델 학습의 결과이며 f1-score 가 1 이었다. 이는 모델이 테스트셋에서 정상과 사기 거래를 완벽하게 분류했음을 알 수 있다. 결과를 오차 행렬로 그려본 결과는 다음과 같다.



사기를 정상거래로, 정상을 사기거래로 분류한 케이스가 0 이었음을 나타낸다.

2) 랜덤 포레스트

랜덤 포레스트(Random Forest)는 분류와 회귀 작업에 사용되는 기계학습 알고리즘 중 하나이다. 알고리즘 작동 방식은 먼저 부트스트랩 샘플링을 통해 전체 데이터에서 샘플을 복원 추출한다. 그렇게 추출된 데이터 샘플들은 분개시마다 변수들을 랜덤하게 선택하여 만든 의사 결정 트리에 학습된다. 앙상블 학습은 여러 개의 학습 알고리즘을 조합하여 예측 성능을 극대화하는 방식인데, 랜덤 포레스트는 여러 개의 결정 트리를 훈련하고 예측을 결합하는 방법을 사용하여 성능을 극대화시킨다. 각 트리는 전체 데이터셋 중 일부 특성만을 이용해 훈련되어 과적합을 막고, 여러 개의 결정 트리를 사용하기 때문에 랜덤 포레스트는 결정 트리보다 향상된 성능을 가진다고 평가된다.

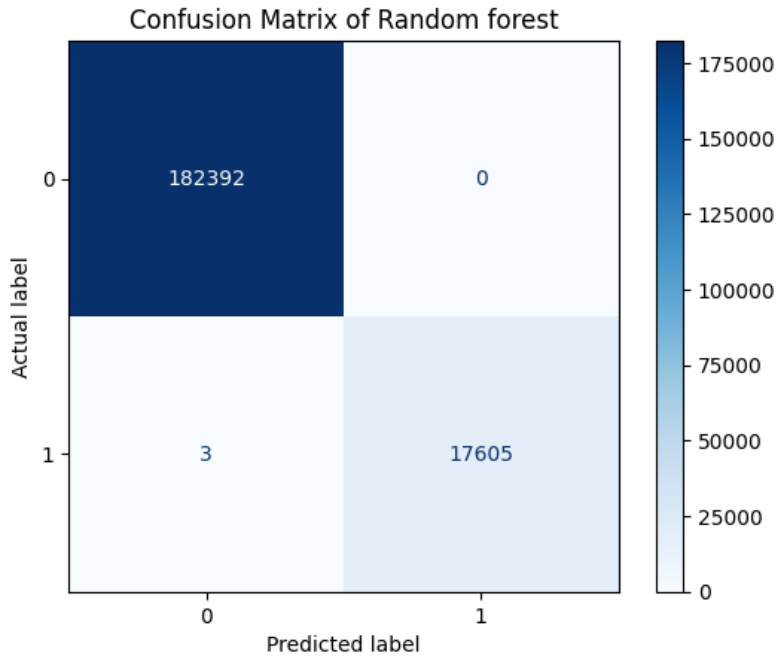


랜덤 포레스트는 기본적으로 블랙박스 모형이지만 변수의 중요도를 설명 가능한 특징이 있다. 블랙박스 모형은 모델 학습 결과에 대한 설명이 어렵다는 것이다. 랜덤 포레스트는 독립변수가 종속변수에 어떻게 영향을 줬는지는 설명할 수 없더라도 어떤 변수가 예측에 중요한 역할을 했는지 변수 중요도 계산을 통해 알 수 있다.

변수 중요도 계산 결과 EDA 에서 중요한 변수라고 예측했던 이전 결재 금액 중앙값과의 비가 가장 중요한 예측을 담당한 것을 알 수 있다. 그 다음으로 사기 거래 예측에 중요한 변수는 “온라인 주문이었는지”, “주거지와의 거리”, “신용카드 칩을 사용했는지”, “핀 넘버를 사용했는지”, “마지막 거래 장소와의 차이”였다.

```
RandomForest의 F1-score: 0.999943204407338
```

위는 랜덤 포레스트를 사용한 모델의 성능 평가 결과이다. F1-score 가 거의 1 에 가까운 결과로 보아, 모델의 예측력은 매우 강력했다.



F1-score 의 오차 행렬을 시각화 한 것은 위와 같다. 정상 거래를 사기 거래로 잘못 예측한 경우는 0 번이고 사기 거래를 정상 거래로 예측한 횟수는 3 번이었다.

3) 인공신경망

*참고 - 로지스틱의 f1-score

로지스틱 회귀분석의 F1-Score: 0.7146068946722988

이 데이터셋의 로지스틱 회귀모형의 f1-score 가 상대적으로 낮은 이유는 범주형 데이터의 수가 너무 많이 때문이다. 일반적으로 로지스틱 회귀분석의 경우 연속형 변수로 이루어진 데이터에 사용된다. 범주형 데이터가 다수 포함될 경우 0 이나 1 의 단순 값을 가지는 데이터들이 모델의 설명력을 낮출 수 있다.

더 적절한 기울기를 구하기 위해 로지스틱보다 범주형 데이터를 더 효율적으로 처리할 것으로 예상되는 인공 신경망 방법을 적용해 보았다. 인공 신경망은 오차 역전파를 통해 모델을 반복적으로 조정하면서 최적 기울기를 찾아내는 능력을 가지고 있다. 이는 복잡한 패턴과 상호작용을 모델링 할 때 특히 유용하다고 알려져 있다.

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

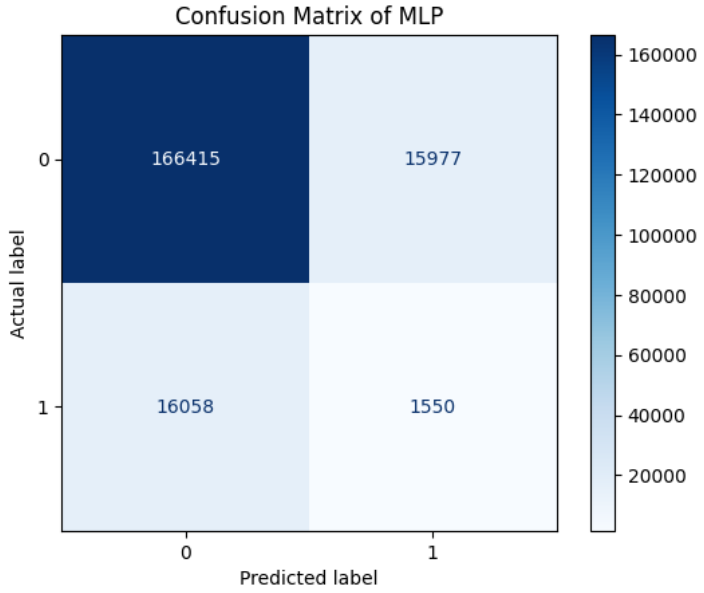
Total params: 78689 (307.38 KB)
Trainable params: 78689 (307.38 KB)
Non-trainable params: 0 (0.00 Byte)

모델 설계에 사용한 노드 개수는 다음과 같다. 노드 개수를 512 개에서 점차 줄이다가, 마지막 노드는 결과를 도출하기 때문에 하나로 설정했다. 액티베이션 함수로는 중간층에서는 Relu 를 사용하고 마지막 학습에서는 종속변수가 0,1 값이기 때문에 sigmoid 함수를 사용한다. sigmoid 함수는 y 값이 0 이나 1 로 대응될 확률 값을 의미한다. 오차 함수로는 이진 분류에 적절한 바이너리 크로스 엔트로피를 사용했다

학습 데이터셋이 800,000 개인 점을 고려하여 각 batch 의 크기를 1,000 으로 설정했다. 총 학습 횟수인 epoch 는 50 번으로 설정하고 MLP 를 설계했다.

성능 측정결과는 다음과 같다.

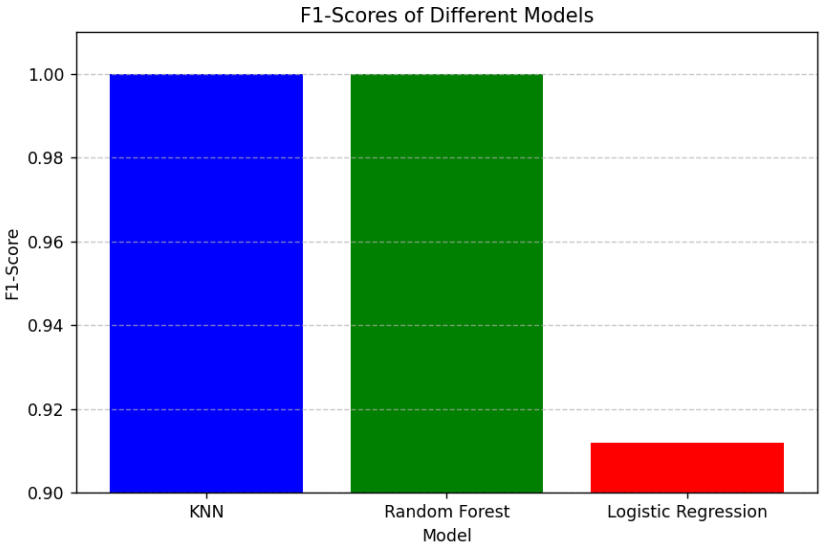
인공신경망의 F1-Score: 0.9122004028887397



인공신경망의 잘 알려진 강력한 성능에도 불구하고 앞서 사용한 분류 알고리즘과 비교했을 때 낮은 결과를 보였다. 오차 행렬을 확인해 봐도 사기 거래를 정상 거래로 분류한 케이스가 16058 번으로 상당히 많은 숫자였다.

5. 결론

1) 연구결과정리



KNN, 랜덤 포레스트, 인공신경망의 성능은 위와 같다.

본 프로젝트 데이터 셋을 사용한 이상거래 탐지에서 왜 앞선 2 개의 분류 알고리즘이 인공신경망보다 성능이 좋게 나왔을 지 분석해 보면, 인공신경망은 이진 데이터의 경우같이 각 특성이 적은 정보를 담고 있을 때 모델이 이러한 정보를 과대 해석하여 과적합을 일으킨 것이라고 설명할 수 있다. 복잡한 패턴을 모델링 할 때 유용한 방법론일수 있지만 단순한 데이터가 주어졌을 때의 분류 문제에는 적합하지 않을 수 있다.

반대로 결정 트리 기반 모델(랜덤 포레스트 등)은 인공신경망처럼 복잡한 연산을 사용하지 않아 범주형 변수의 정보를 효과적으로 처리할 수 있다. 트리의 각 노드는 조건(특징) 하에서 독립적이라는 것도 장점으로 작용한다. 또 이 데이터셋의 경우 범주형 데이터가 모두 이진 데이터였기 때문에 트리의 복잡성을 증가시키지 않아 성능이 좋았다고 평가할 수 있다. 또 KNN 의 경우에도 N 차원 특징 벡터 스페이스에서의 기존 데이터 값들을 위치시키고 그 기준으로 새로운 데이터를 분류하기 때문에 단순한 알고리즘으로 평가받는다.

성능과 별개로 실제 적용했을 때 테스트값 입력 시 모델의 시간 복잡도도 고려해야 한다. 사기 거래 탐지는 빠르게 이루어져야 하므로 모델 학습시의 시간 복잡도가 높은 것은 감안하더라도 테스트 시의 시간 복잡도는 낮을수록 좋다. N 을 기존 데이터셋의 크기라고 했을 때 KNN 은 테스트 시의 시간 복잡도가 O(N)으로 셋 중에 가장 높다. 학습 시간(O(1))이 거의 존재하지 않는 대신 테스트가 입력되었을 때야 비로소 거리 비교가 가능해지기 때문이다.

랜덤 포레스트의 경우 테스트 시의 시간 복잡도는 O(m*log(N))으로 m 은 트리 개수를 의미한다. 기본 파이썬 코드에서 설정하는 트리 개수는 100 개로 m 은 거래 빅데이터 개수(N)에 비해선 시간 복잡도에 거의 영향을 못 미치는 수준이다. 인공신경망(딥러닝) 또한

테스트시의 시간 복잡도는 $O(\sum_{i=1}^{L}(N_i \cdot N_{i-1}))$, L 은 레이어 개수이고 $N_i \cdot N_{i-1}$ 는 각 레이어 노드 수, 로 아무리 딥러닝이 깊은 레이어를 가지고 있더라도(이는 과적합으로 이어질 가능성이 크다.) 거래 빅데이터 수보다는 적을 것이다.

각 모델의 시간 복잡도를 낮은 순서 순으로 정리하면 다음과 같다.

(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)

2) 결론

세가지 방법론의 성능 측면에서는 KNN>랜덤 포레스트>인공신경망 순이지만 테스트 시의 시간 복잡도는 랜덤 포레스트가 가장 낮고 KNN 이 압도적으로 가장 높았다. 사기 거래를 빠르게 탐지하는 것이 FDS 에서 중요한 요소라는 것을 감안하면 랜덤 포레스트가 가장 좋은 방법론이었다고 결론 내릴 수 있다.

본 프로젝트를 통해 구축한 모델들을 현실에서 적용 시 유효할지에 대해서는 다음 세가지 측면에서의 우려가 있다.

첫번째로 사용한 데이터 종류가 개인의 특성에 따른 거래 패턴을 반영한 정보는 아니라는 것이다. 실제 사용할 정도의 예측이 가능 하려면 각 개인의 방대한 거래 시계열 데이터를 사용한 맞춤형 서비스가 제공되어야 할 것인데 프리 소스 데이터셋은 거래 단위의 데이터셋 밖에 찾을 수 없었다. 이 데이터셋에서는 직전 거래 정보와의 차이만 제공하고 있으므로 현실 적용에는 어렵다는 의견이다.

두번째는 이 프로젝트에서는 과거 데이터만 사용했으므로 거래 패턴의 변화를 지속적으로 예측 불가능하다. 즉 해당 데이터셋 내에서만 유효할 수 있다. 과거 데이터만 사용하는 방법론을 사용했다는 것도 한계이다. 실제 적용시의 유효성을 생각하면 새로운 데이터를 기반으로 지속적으로 학습하는 연속 학습 알고리즘의 도입이 필요하다.

세번째로 실제 모델을 만드는 데 가장 큰 문제는 방대한 데이터셋에서 이상 거래와 정상 거래를 구별해서 라벨을 만들 수 있을지에 대한 것이다. 일부 데이터셋에는 적용할 수

있겠지만 거래마다 라벨을 붙여 지도학습을 진행하는 것은 비용이 크기 때문에 일반적으로 사용하기엔 문제가 있을 수 있다. 실제 탐지 모델은 여러 고객의 데이터셋에서 얻은 pre-trained model 을 개인 맞춤형 모델에 적용하는 것이나 유사한 거래 고객과의 비교법을 사용해야 할 것이다.

그럼에도 데이터셋 내 주요 변수를 선별할 수 있었다. EDA 와 랜덤 포레스트 방법론에서 확인한 이상거래탐지에 중요한 변수는 “이전 결재 금액 중앙값과의 비”, ”온라인 거래인지”, “주거지와 거래 장소와의 거리” 였다. “이전 결재 금액 중앙값과의 비”가 큰 거래에서 사기 거래의 설명력이 높아지는 것은 사기꾼들이 고액을 노리기 때문에 어찌 보면 당연하지만 가장 중요한 변수였다는 것을 확인 가능하다.

그리고 특히 ”온라인 거래인지”, “주거지와 거래 장소와의 거리”가 주 변수에 나란히 있는 것은 주목할 만하다. 비대면 거래 상용화 이후 사용자 정보 해킹하여 원거리의 IP 를 통해 결제를 시도하는 인터넷 사기가 지속적으로 증가했기 때문이다. 온라인 거래의 경우 취약점이 많기에 이상거래 탐지뿐만이 아닌, 사기 거래가 사전에 발생하지 못하도록 해킹을 통한 도용 등을 예방하는 보안 기술이 필수인 것을 알 수 있다. 또한 의심 거래가 나타났을 시 추가 인증 등을 요구해서 실제 고객이 맞는 지 확인해야 한다.

랜덤 포레스트에서 중요도가 큰 변수가 아니었던 핀 번호나, 주민등록번호 같은 경우는 타인 도용 위험이 크다. 따라서 생체인증(지문, 얼굴)을 요구하는 것이 바람직하다. 다만 생체 인증 관련 최신 이슈 중 하나가 딥 페이크인 것을 고려하면 금융 기업들은 딥 페이크를 가려내는 기술력을 갖추고 있어야 할 것이다.

“신용카드 칩을 사용한 거래”, “마지막 거래장소와의 차이”가 중요도 낮은 변수였던 이유는 분실된 카드를 사용하는 사기 거래가 자주 일어나는 것으로 보인다. 실제 카드를 활동지 근처에서 분실 시 정상거래와 사기거래의 경계는 모호해진다. 신용카드 칩은 마그네틱 스트라이프에 비해 복제가 훨씬 어렵기 때문에(카드 스키밍 방지) 안전한 결제 방식이라고 알려져 있지만 이 경우에는 효력이 없어진다. 이 경우, 중요도 높은 변수인 “이전 결재 금액 중앙값과의 비” 등을 통한 이상거래 탐지로 추가 인증을 요구할 수 있을 것이다.

또 이 프로젝트의 의의는 데이터셋의 성격(연속/범주, 라벨 有)에 맞춰 분석하는 법을 탐색했다는데 있다. PCA, K-mean 및 비지도 군집생성 알고리즘, 로지스틱 회귀분석 등 여러 방법론을 사용해 보았을 때 그 방법들이 유효하지 않았던 점과 낮은 설명력과 예측을 했다는 것을 알 수 있었다. 이를 토대로 제한된 데이터셋 환경 속에서 최대의 예측을 가질 것이라 생각하는 방법론을 3 개 선정하고 이를 평가해 보았다. 그 결과, 비교적 간단한 알고리즘 구조를 가진 지도학습 방법론이 가장 좋은 결과를 도출함을 알 수 있었다.

6. 참고문헌

RISS 논문 키워드="이상 거래 탐지", 순차패턴 분석을 통한 이상금융거래탐지 연구: 선불전자지급수단 거래를 중심으로 외 81 건

https://www.data.go.kr/data/15064566/fileData.do

Anomaly Detection using PCA in Time Series Data, https://ieeexplore-ieee-org-ssl.glibproxy.gachon.ac.kr/document/10502929?arnumber=10502929&SID=EBSCO:edseee

https://bitnine.tistory.com/397

https://www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud/code

https://velog.io/@jadon/F1-score%EB%9E%80

https://www.kaggle.com/code/linakeepgoing/6-ai-auto-encoder

https://velog.io/@zlddp723/%EB%A1%9C%EC%A7%80%EC%8A%A4%ED%8B%B1-%ED%9A%8C%EA%B7%80Logistic-Regression

https://velog.io/@pyose95/Data-Analysis-14.-%EC%83%81%EA%B4%80%EB%B6%84%EC%84%9D-Correlation-Analysis
https://rfriend.tistory.com/773

https://yaeyang0629.tistory.com/entry/%EB%8D%B0%EC%9D%B4%ED%84%B0-%EB%B6%88%EA%B7%A0%ED%98%95%ED%95%B4%EA%B2%B0%EB%B0%A9%EC%95%88-Random-UnderSampling

https://medium.com/@reddyyashu20/k-means-kmodes-and-k-prototype-76537d84a669

https://mozenworld.tistory.com/entry/%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D-%EB%AA%A8%EB%8D%B8-%EC%86%8C%EA%B0%9C-4-%EB%9E%9C%EB%8D%A4-%ED%8F%AC%EB%A0%88%EC%8A%A4%ED%8A%B8-Random-Forest#google_vignette

https://www.tensorflow.org/tutorials/generative/autoencoder?hl=ko

https://velog.io/@hyesoup/KNN-K-Nearest-Neighbor-%EA%B0%9C%EB%85%90

https://blog.naver.com/sjy5448/222427780700

https://www.tta.or.kr/data/androReport/ttaJnal/172-2-3-7.pdf

https://eiec.kdi.re.kr/policy/materialView.do?num=243367

https://news.einfomax.co.kr/news/articleView.html?idxno=4247562

https://maloveforme.tistory.com/221