report

Data EDA

```
[Data Head]
   Time        V1        V2        V3        V4        V5        V6        V7  \
0   0.0 -1.359807 -0.072781  2.536347  1.378155 -0.338321  0.462388  0.2395
99
1   0.0  1.191857  0.266151  0.166480  0.448154  0.060018 -0.082361 -0.07880
3
2   1.0 -1.358354 -1.340163  1.773209  0.379780 -0.503198  1.800499  0.7914
61
3   1.0 -0.966272 -0.185226  1.792993 -0.863291 -0.010309  1.247203  0.2376
09
4   2.0 -1.158233  0.877737  1.548718  0.403034 -0.407193  0.095921  0.5929
41

        V8        V9  ...       V21       V22       V23       V24       V25  \
0  0.098698  0.363787  ... -0.018307  0.277838 -0.110474  0.066928  0.12853
9
1  0.085102 -0.255425  ... -0.225775 -0.638672  0.101288 -0.339846  0.16717
0
2  0.247676 -1.514654  ...  0.247998  0.771679  0.909412 -0.689281 -0.32764
2
3  0.377436 -1.387024  ... -0.108300  0.005274 -0.190321 -1.175575  0.647376
4 -0.270533  0.817739  ... -0.009431  0.798278 -0.137458  0.141267 -0.20601
0

       V26       V27       V28  Amount  Class
0 -0.189115  0.133558 -0.021053  149.62      0
1  0.125895 -0.008983  0.014724    2.69      0
2 -0.139097 -0.055353 -0.059752  378.66      0
3 -0.221929  0.062723  0.061458  123.50      0
```

```
4  0.502292  0.219422  0.215153   69.99      0

[5 rows x 31 columns]
...
1       492
Name: count, dtype: int64
Normal(0): 284315, Fraud(1): 492
Fraud Ratio: 0.0017
```

```
[Class Counts after Sampling]
Class
0    10000
1      492
Name: count, dtype: int64
Normal(0): 10000, Fraud(1): 492
```

데이터 요약:

class불균형 - 0과1 레이블의 수량 차이가 많다

Amount과 Time변수의 scale를 다름으로, normalization를 필요한다다

데이터 전처리과 분할

```
Features shape: (10492, 30)
Target shape: (10492,)

------------------------------
Train set shape: (8393, 30)
Test set set shape: (2099, 30)

[Train Class Ratio]
Class
0    0.953056
1    0.046944
```

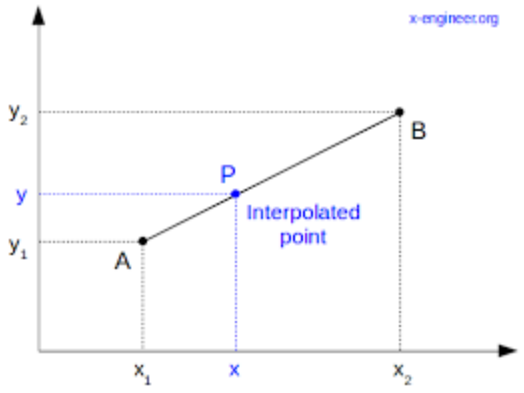
```
Name: proportion, dtype: float64

[Test Class Ratio]
Class
0    0.953311
1    0.046689
Name: proportion, dtype: float64
```

데이터셋 중에 있는 Amount변수를 변환하고 제거 했으며, test/ train dataset을 나누었다.

SMOTE

```
Before SMOTE - Class 1 counts: 394
After SMOTE  - Class 1 counts: 7999
```

데이터 불균형(Imbalanced Data) 문제는 머신러닝 모델이 다수 클래스(Normal)에 편향되게 학습하도록 만듭니다.
단순히 소수 클래스(Fraud)를 복제하는 Random Oversampling은 과적합(Overfitting) 위험이 있습니다.
SMOTE(Synthetic Minority Over-sampling Technique)는 소수 클래스 데이터 포인트 사이의 보간(interpolation)을 통해
새로운 가상의 데이터를 생성함으로써 과적합 위험을 줄이면서 데이터 균형을 맞추는 효과적인 방법입니다.



linear interpolation

Train and Test

```
Target: Recall >= 0.8, F1 >= 0.88, PR-AUC >= 0.9
------------------------------
Class 1 (Fraud) - Recall: 0.8265, F1: 0.8950
Class 0 (Normal) - Recall: 0.9990, F1: 0.9953
PR-AUC: 0.9157
------------------------------
결과: 목표 달성 성공! (Goals Achieved)
```

과제의 요구를 만났다

- 본 프로젝트에서는 `Time` 변수를 제거하지 않고 학습에 포함시켰습니다. 실험 결과, `Time` 변수 제거 시 모델의 전반적인 성능(특히 Recall)이 오히려 하락하는 현상을 확인했습니다.
- **원인 분석**: 사기 거래(Fraud)가 특정 시간대에 집중적으로 발생하는 패턴이 존재하며, 모델이 이 시간적 특징을 유의미한 변별자로 학습한 것으로 판단됩니다.