

5(1)-Basic_ML

5(1)-Basic ML Report

Machine Learning Modeling Process

1. Define Problem

2. Collect & Explore Raw Data

3. Sample from Raw Data

if Class Imbalance exists

4. Data Preprocessing

Standardize some variables if needed

Split data to dataframe (X, y)

5. Split Train Set and Test Set (and Validation Set)

6. Oversample

if Class Imbalance exists

SMOTE (Synthetic Minority Over-sampling Technique)

적용 이유:

Random sampling이 진행된 이후에도, 현재 데이터는 정상 거래(데이터 전체의 95.31%)가 사기 거래(데이터 전체의 4.69%)보다 훨씬 많은 불균형 상태입니다. 클래스 불균형 상태에서 학습한다면 모델이 '정상' 클래스만 예측하려는 편향이 생겨 분류 문제를 해결하기 어렵습니다. Random sampling과 다르게 SMOTE는 단순히 데이터를 복제하지 않고 소수 클래스 데이터를 분석해 유사한 데이터를 생성하므로 Overfitting 위험을 줄이면서 클래스 균형을 맞출 수 있습니다.

7. Train ML Model using Train Set

Performance Metrics

- Recall: 실제 positive 데이터 중 모델이 positive로 옳게 분류한 비율
- Precision: 모델이 positive로 분류한 데이터 중 실제 positive의 비율
- F1-Score: $2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$

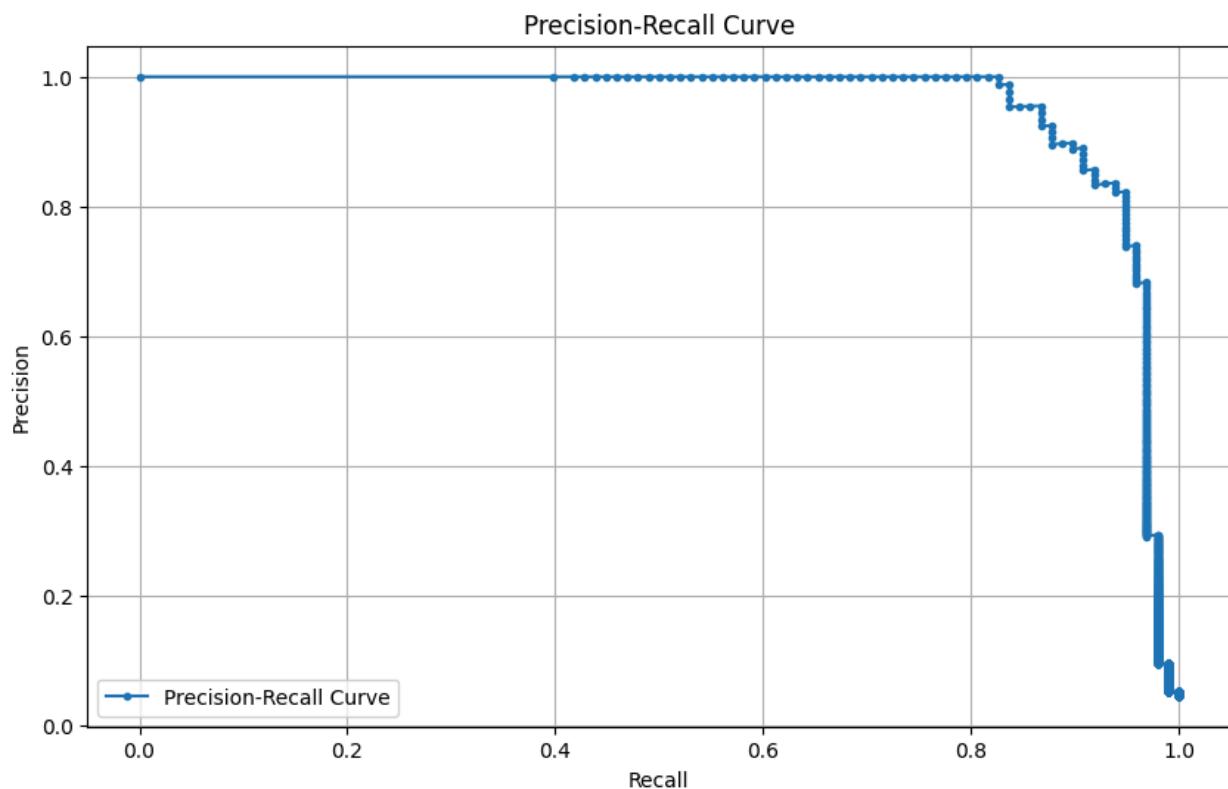
8. Improve Performance

- Hyperparameter tuning (본 실험에서는 이루어지지 않음)
- Threshold adjustment

9. Test ML Model using Test Set

모델 학습 성능 분석

Precision-Recall Curve



Precision과 Recall의 tradeoff를 확인할 수 있는 곡선입니다.

Initial training

Class	Precision	Recall	F1 Score
0	1.00	0.97	0.98
1	0.62	0.97	0.76

PR-AUC Score : 0.9559

분석: Class 0 (정상 거래)의 F1 score는 매우 우수합니다. 그러나 Class 1 (사기 거래)는 recall이 높으나, precision이 높은 편이 아닙니다. 다시 말해서 모델이 실제 사기 거래를 사기 거래라고 예측할 가능성은 높으나, 사기 거래가 아닌 거래를 사기 거래라고 잘못 예측하는 경우가 꽤 많았다는 것(38%)을 의미합니다. 이에 Class 1 분류 능력을 향상시켜야 할 필요성이 있습니다.

After threshold adjustment

Class 1의 precision을 높이려면 임계값(threshold)을 0.5보다 많이 높여야 합니다. 이에 threshold를 0.9로 확 높게 잡아서 recall을 조금 낮추고 precision을 높이는 선택을 했습니다.

Threshold = 0.9

Class	Precision	Recall	F1 Score
0	0.99	0.99	0.99
1	0.89	0.90	0.89

PR-AUC Score : 0.9559

분석: 임계값(threshold)을 0.9로 상향 조정한 결과, Class 1(사기 거래)의 precision이 0.62에서 0.89로 크게 향상되었습니다. 이는 모델이 사기 거래라고 확신이 높다고 판단할 때 사기 거래라고 분류하도록 기준을 강화했기 때문입니다. Precision-Recall tradeoff에 따라 recall은 0.97에서 0.90으로 소폭 감소했습니다. 그러나 잘못 판단하는 경우가 훨씬 줄어들었기 때문에 성능이 유지되며 전체 시스템의 안정성도 확보되었습니다.

After applying optimal threshold for highest F1 score

Class 1의 F1 score가 가장 높은 threshold를 찾는다면 성능을 가장 높일 수 있을 것이라고 생각했습니다. 이에 precision-recall curve에서 가장 적절한 precision-recall 짹을 찾을 수 있는 threshold를 확인했습니다.

Threshold = 0.9581

Class	Precision	Recall	F1 Score
0	0.99	1.00	1.00
1	0.96	0.87	0.91

PR-AUC Score : 0.9559

분석: F1-score를 최대화하는 최적 threshold를 적용한 결과, Class 1의 F1-score가 0.91로 가장 높은 값을 기록하였고, precision(0.96)과 recall(0.87) 간의 균형이 가장 잘 이루어졌습니다. F1 score가 최고점이기 위해서는 높은 recall보다도 높은 precision이 더 중요한데, 이는 모델이 사기 거래라고 판단한 경우 대부분 실제 사기 거래임을 보장한다는 점에서(신뢰도가 높아서)입니다. 이 단계는 전반적인 탐지 성능이 가장 우수한 지점이라고 볼수 있으며, 실제 사기 탐지 시스템에서 적용하기에 가장 합리적인 threshold 선택으로 평가됩니다.

Final Result

- **Recall:** 0.80 이상 달성 (성공)
- **PR-AUC:** 0.90 이상 달성 (성공)
- **F1-score:** 0.88 이상 달성 (성공)

만약 달성하지 못했다면, Logistic Regression 모델의 한계일 수도 있어서 Random Forest 모델 등 타 모델을 통해 성능을 올릴 수 있었을 것이라고 생각합니다. 하지만 실제 실험에서는 logistic regression만으로 최적의 모델 목표치를 달성하였습니다.