01 Introduction

- Background
  - Dataset with a class imbalance requires oversampling/undersampling
    - SMOTE synthetic minority oversampling technique
  - In datasets with an extreme class imbalance, oversampling might cause overfitting
    - Overfitting causes models to capture unnecessary noise
    - Resulting in misleading predictions/classifications

01 Introduction

Background

- Dataset with a class imbalance requires oversampling/undersampling
  - SMOTE synthetic minority oversampling technique
- In datasets with an extreme class imbalance, oversampling might cause overfitting
  - Overfitting causes models to capture unnecessary noise
  - Resulting in misleading predictions/classifications

Research Question

RQ1) How can we use SMOTE in a dataset with an extreme class imbalance?

RQ2) Is XGBoost a good choice for a model to overcome the risk of overfitting in extremely unbalanced datasets?

02 EDA

■ Credit Card Fraud Detection dataset

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.128539	-0.189115	0.133558	-0.021053	149.62	0
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.167170	0.125895	-0.008983	0.014724	2.69	0
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.689281	-0.327642	-0.139097	-0.055353	-0.059752	378.66	0
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1.175575	0.647376	-0.221929	0.062723	0.061458	123.50	0
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	0.141267	-0.206010	0.502292	0.219422	0.215153	69.99	0

- ○ V1, V2, ⋯ V28 – the principal components obtained by PCA
- ○ Time – the seconds elapsed between each transaction and the first transaction in the dataset
- ○ Amount – the transaction amount
- ○ Class – the response variable (1: fraud / 0: otherwise)

02 EDA

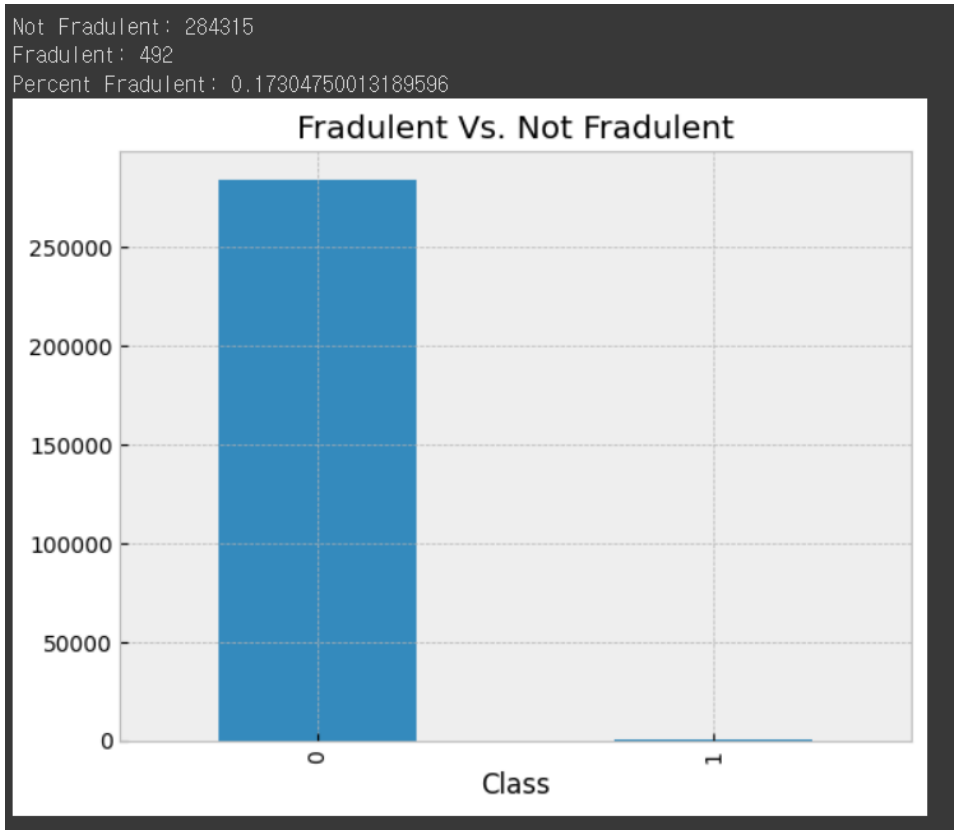
- Key insights from summary statistics
  - Mean transaction amount is 88.35
    - Also, the std. is quite high at 250.12
    - ➔ High variability in transaction amounts
  - Mean of Class column is 0.001727
    - ➔ Suggests that it is a highly imbalanced dataset
    - ➔ Very few fraudulent transactions
      - ➔ Frauds: only 0.1727% of the total

	Amount	Class
count	284807.000000	284807.000000
mean	88.349619	0.001727
std	250.120109	0.041527
min	0.000000	0.000000
25%	5.600000	0.000000
50%	22.000000	0.000000
75%	77.165000	0.000000
max	25691.160000	1.000000

02 EDA

■ Fraudulent vs. Not Fraudulent

○ Extremely imbalanced

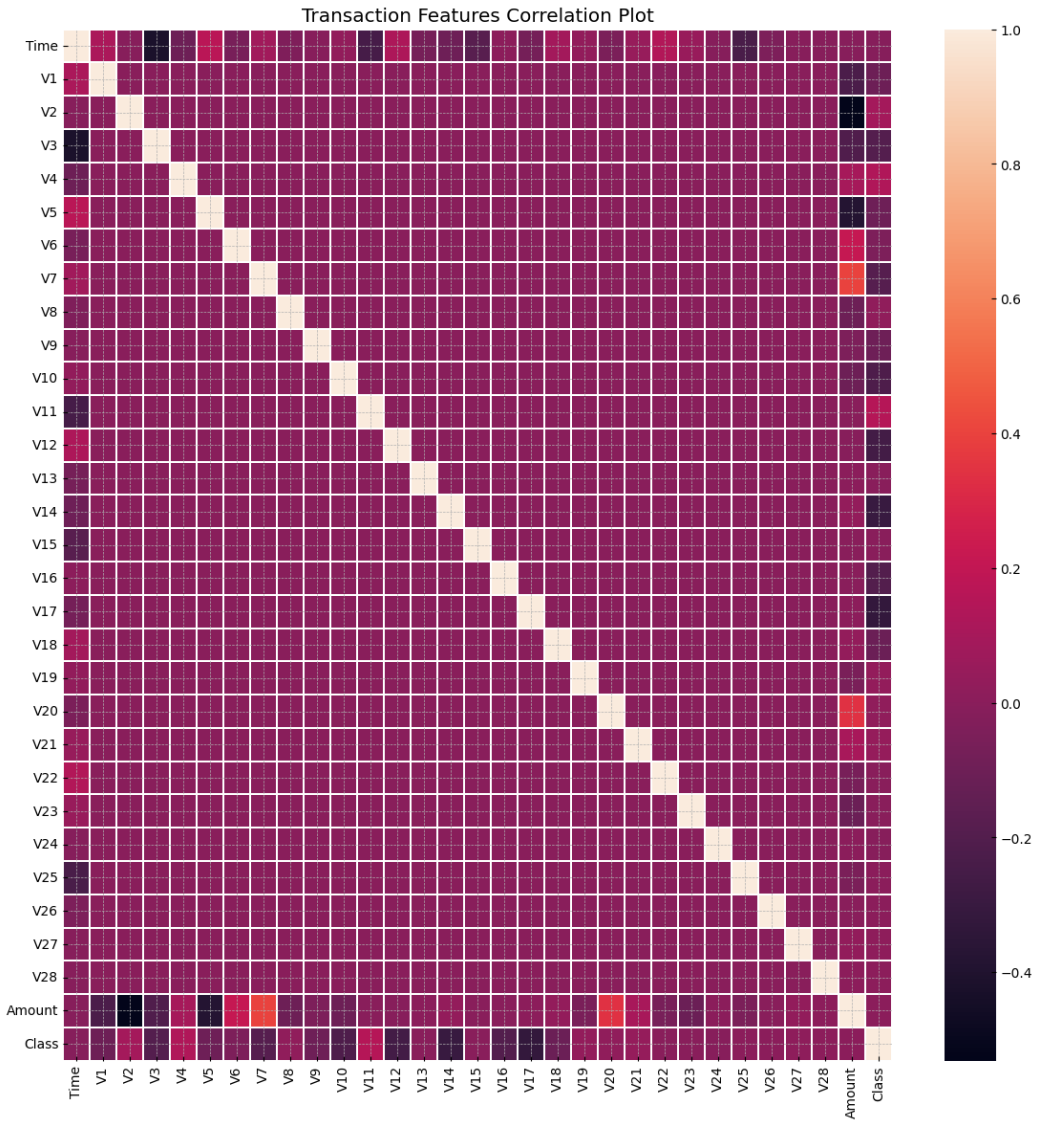


Using accuracy alone will reward models that:
- correctly predicts the majority class
- even if it performs poorly on the minority class

➔ use AUPRC (area under Precision–recall curve) for future model performance measures

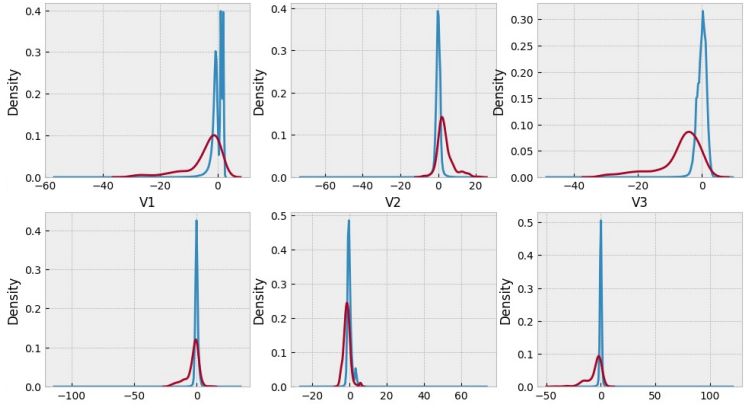
02 EDA

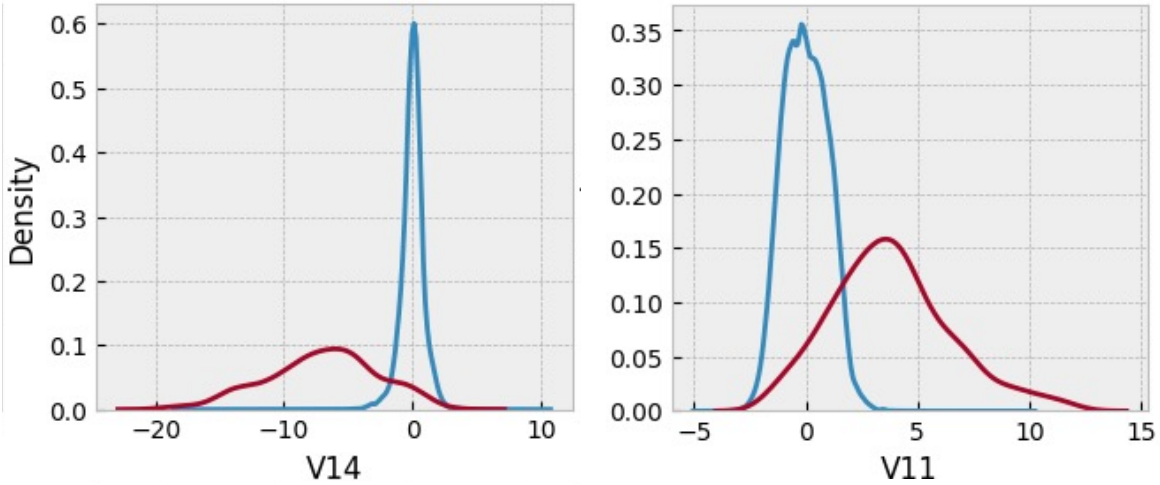
- Fraudulent vs. Not Fraudulent
  - As a result of the PCA transformations, there are not many notable correlations
    - Although there are some exceptions:
      - Negative correlations of V2 and V5 with Amount
      - Positive correlations of V20 and V7 with Amount



02 EDA

- Normal transactions
  - Most features are concentrated around zero
    - Distributed around the mean value (around zero).
- Fraudulent transactions
  - Features such as V11 and V14 exhibit highly skewed distributions





03 SMOTE

- A technique used to address the class imbalance issue in a dataset

☑ Mechanism of SMOTE

1. Identify minority class samples
2. Select a sample and find its nearest neighbor data points
3. Generate synthetic samples for the randomly selected sample
    1. Select one of the k nearest neighbors of x_j at random
    2. Calculate the difference between the feature vector of x_j and x_closest
    3. At random number between 0 and 1, called the coefficient, is chosen and multiplied to (x_j – x_closest) to create a variation
    4. This variation is then added to x_j to create x_synthetic
4. Add synthetic samples to the dataset

03 SMOTE

- A technique used to address the class imbalance issue in a dataset

- Concerns about overfitting:

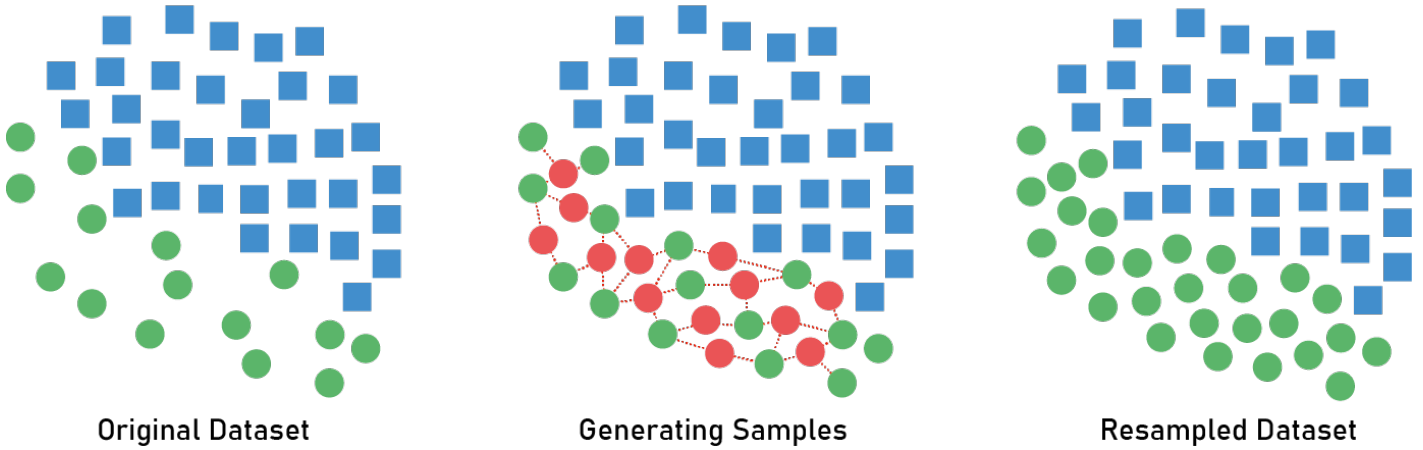
1. Can handle overfitting by adjusting its parameters (e.g., # of k)
2. Data preprocessing (e.g., PCA) can prevent overfitting when SMOTE is applied to an extremely unbalanced dataset

03 SMOTE

- Applied SMOTE with k=5
  - Resulted in a balanced the number of fraud and non-fraud transactions

- Before SMOTE
  - Class 0: 213233
  - Class 1: 372
- After SMOTE
  - Class 0: 213233
  - Class 1: 213233

```
Before oversampling: Counter({0: 213233, 1: 372})
After oversampling: Counter({0: 213233, 1: 213233})
```



03 SMOTE

- Before SMOTE

```
Accuracy score for train: 1.0
Accuracy score for test: 0.9996207971686188
```

```
AUPRC: 0.86
AUROC: 0.98
```

- After SMOTE

```
Accuracy score for train: 1.0
Accuracy score for test: 0.9995084407741356
```

```
AUPRC: 0.86
AUROC: 0.97
```

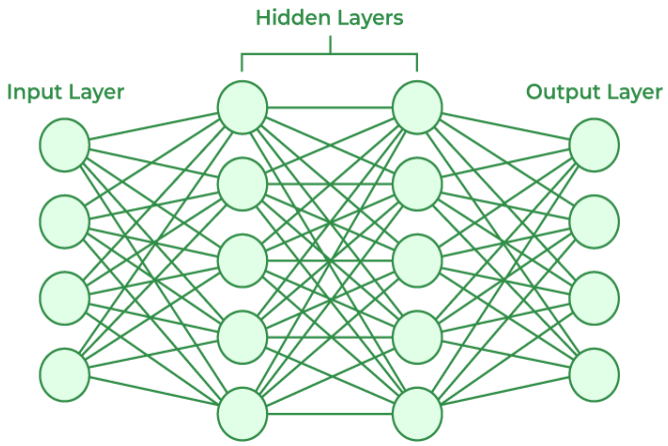
➔ Appropriate preprocessing prevented SMOTE from causing overfitting in an extremely imbalanced dataset

  ➔ As we can know from the fact that the test scores did not decrease dramatically after SMOTE

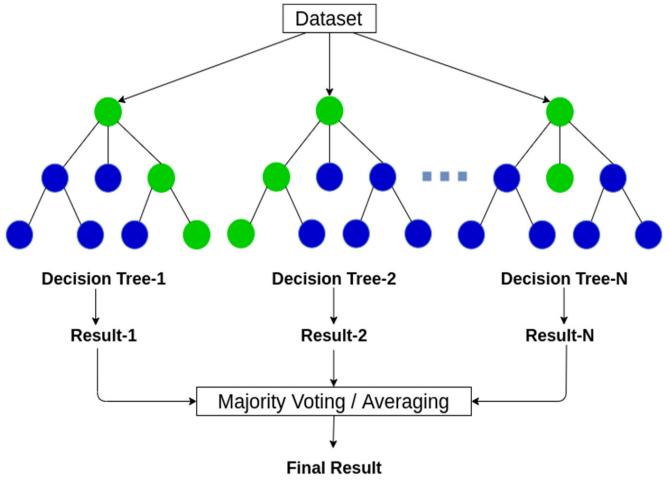
04 Model Comparison

- Compared three models' performances:

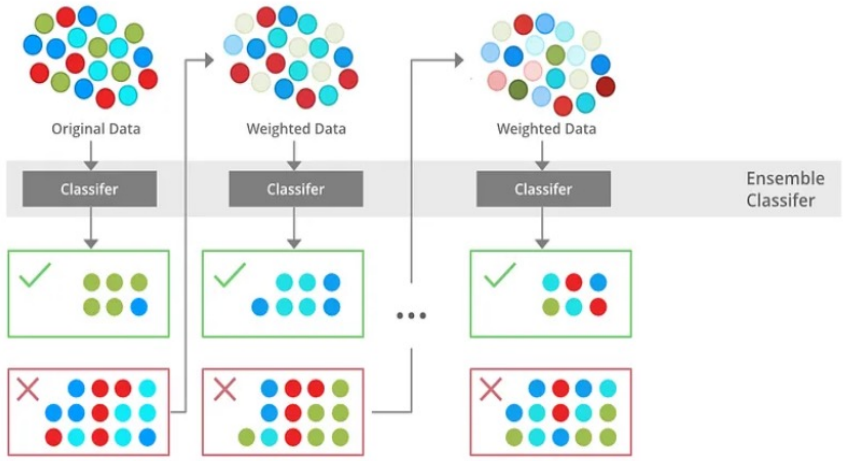
Artificial Neural Network



RandomForest



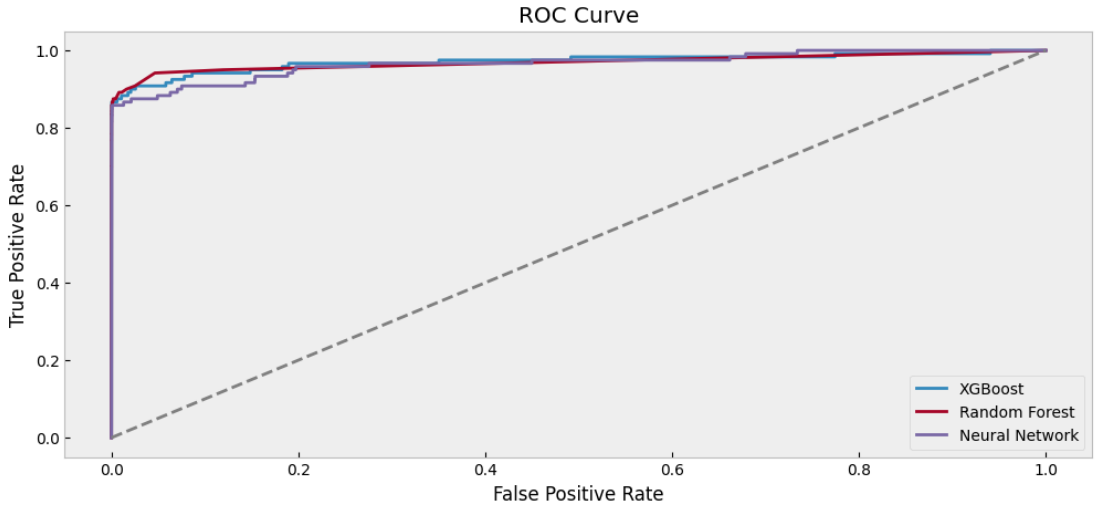
XGBoost

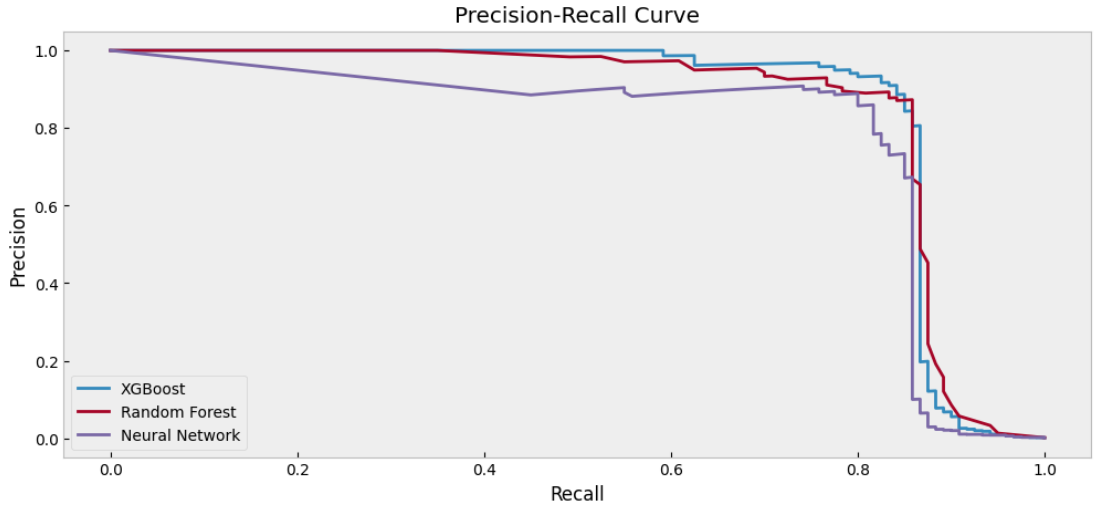


04 Model Comparison

- AUPRC is a good measure with higher values (closer to 1 indicating better performance)
  - XGBoost model had the highest value for both AUPRC and AUROC

Model	AUROC	AUPRC
Artificial Neural Network	0.965551	0.759720
Random Forest	0.969668	0.848327
XGBoost	0.971260	0.858027





04 Model Comparison

XGBoost (eXtreme Gradient Boosting)

☑ Mechanism of XGBoost

1. Sequentially adds new Decision Trees to correct the errors of the previous models
2. Optimizes the model by using Gradient Descent
3. Applies regularization techniques to control the complexity of the model and prevent overfitting (e.g., L1 or L2 for penalizing large parameter values)
4. Provides the evaluation of each feature's importance in predicting

05 Conclusion

■ Conclusion

- ○ Appropriate data preprocessing (e.g., PCA) can prevent overfitting when SMOTE is applied to an extremely unbalanced dataset

- ○ XGBoost model shows better scores on both AUPRC and AUROC than the other two models
  - – This could be due to XGBoost's mechanism of regularization
    - • Regularization controls model's complexity and prevents overfitting by penalizing large parameters

Model	AUROC	AUPRC
Artificial Neural Network	0.965551	0.759720
Random Forest	0.969668	0.848327
XGBoost	0.971260	0.858027

Thank you