Imbalance data and how to handle it

Data science, Machine learning & AI knowledge sharing #1

Speaker: Peerapat.t, Data analyst at Kasikorn asset management (KAsset)

Level: Advance

Agenda

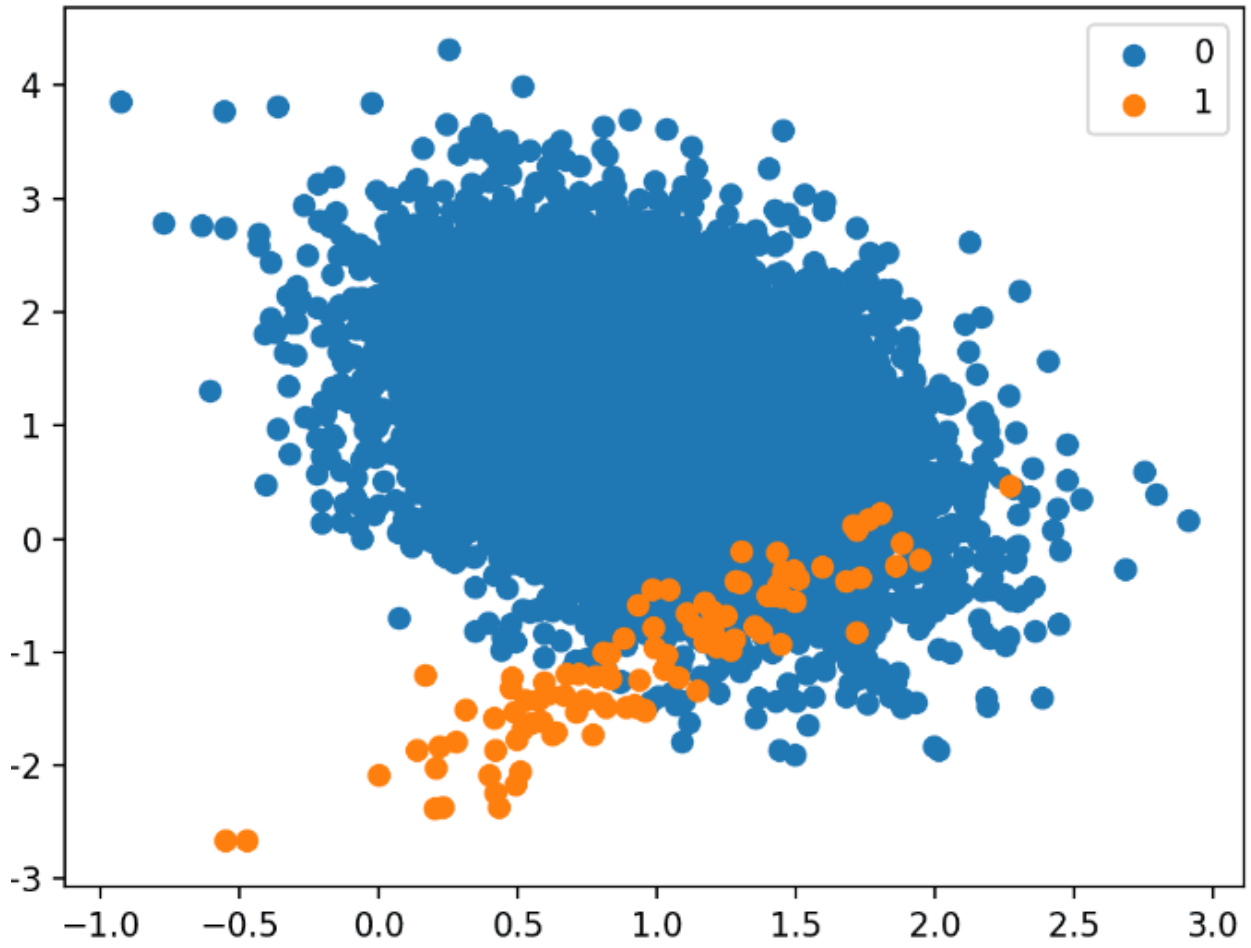
1. What is imbalance data
2. Example scenario
3. Problems of imbalance data
4. How to handle imbalance data***
5. Show case

6. Further reading
7. Appendix

What is Imbalance data

- Imbalanced data refers to a situation in a dataset where the classes are not represented equally.

- Characteristics
  - Majority Class: The class with the most instances.
  - Minority Class: The class with fewer instances.

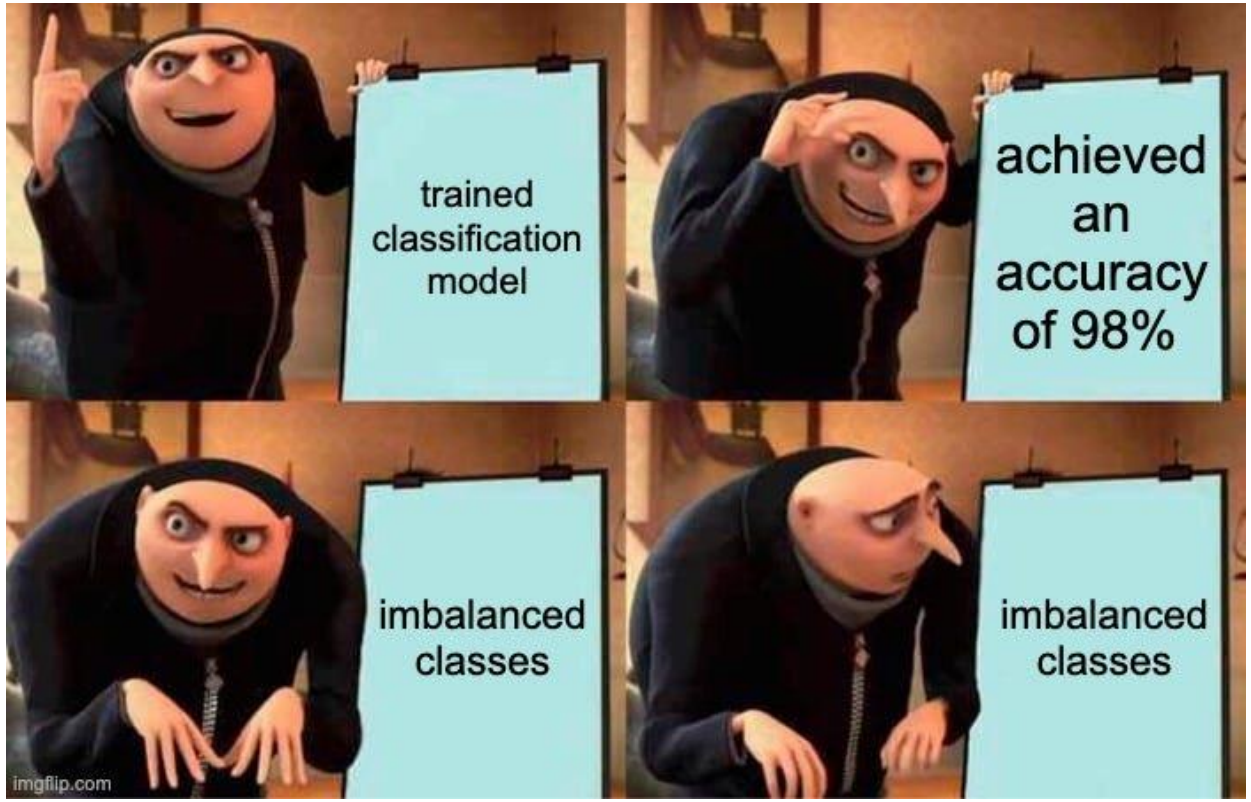


Example Scenario

- Fraud Detection in Credit Card Transactions
- Spam Detection in Emails
- Detection of Rare Diseases
- Churn prediction model
- Propensity-to-buy model

Problems of imbalance data

- Poor Model Performance
- Bias in Performance Metrics



Example Scenario:

Fraud Detection Consider a dataset for credit card transactions with the following distribution: Legitimate transactions (Majority class): 98,000 instances Fraudulent transactions (Minority class): 2,000 instances Total transactions: 100,000 instances

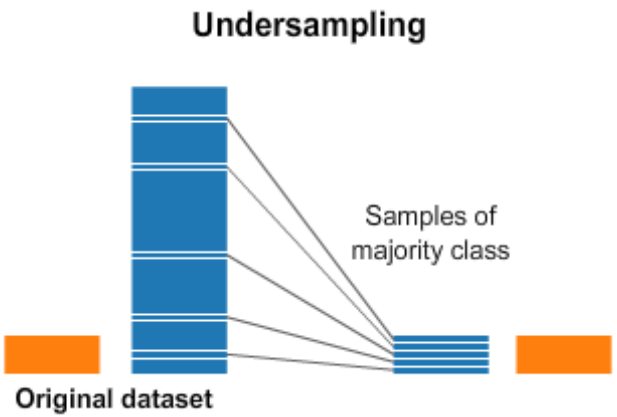
If model predict only majority class: model's accuracy = 98%

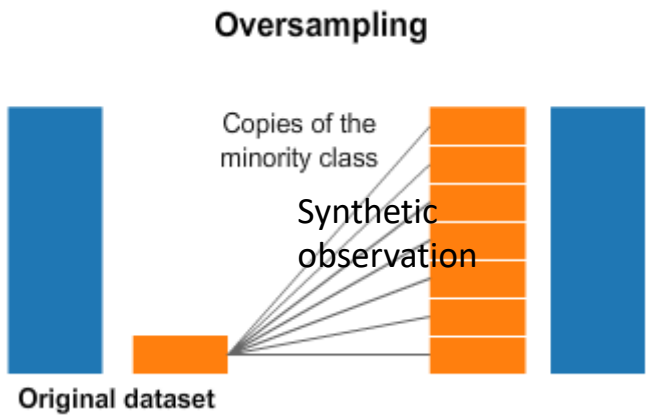
How to handle imbalance data

- Resampling technique* (prepare)
- Evaluation metrics adjustment* (evaluate)
- Algorithm-level method (train)
- Hybrid methods**

How to handle imbalance data

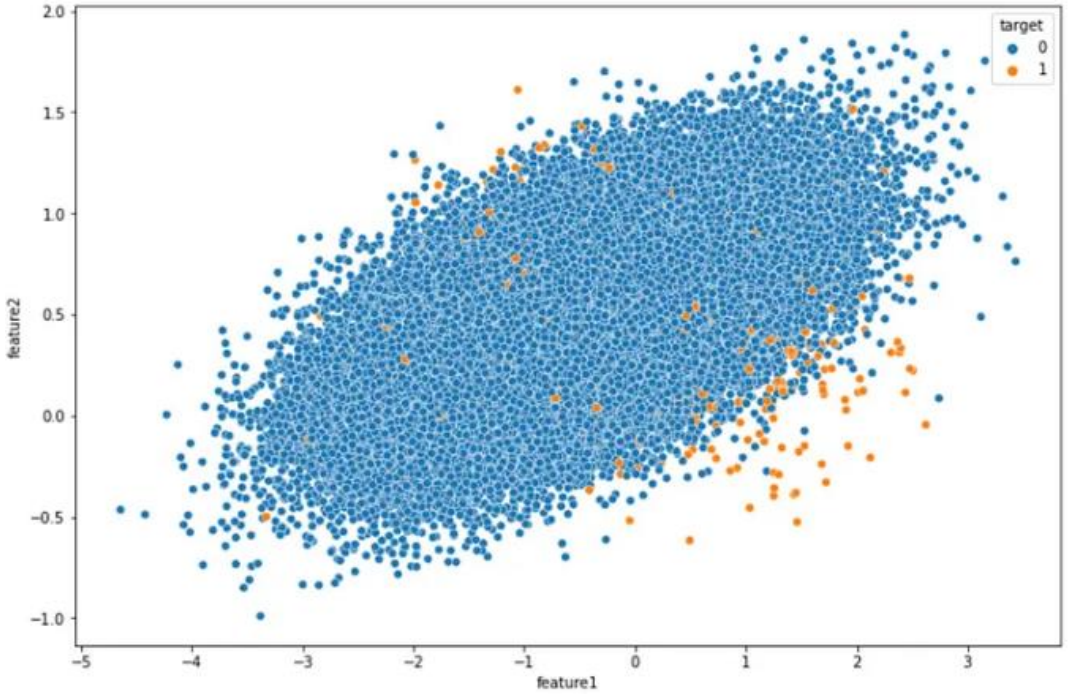
1. Resampling technique
   - Under sampling
     - Random under sampling
   - Over sampling
     - SMOTE
     - Borderline-SMOTE
     - ANASYN

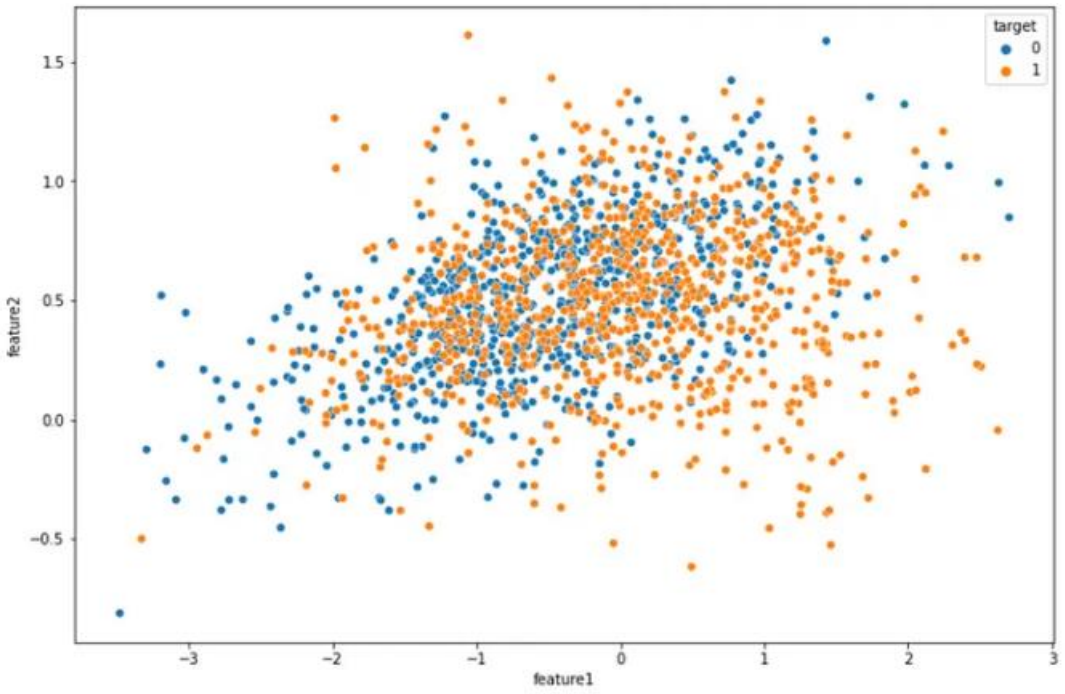




How to handle imbalance data

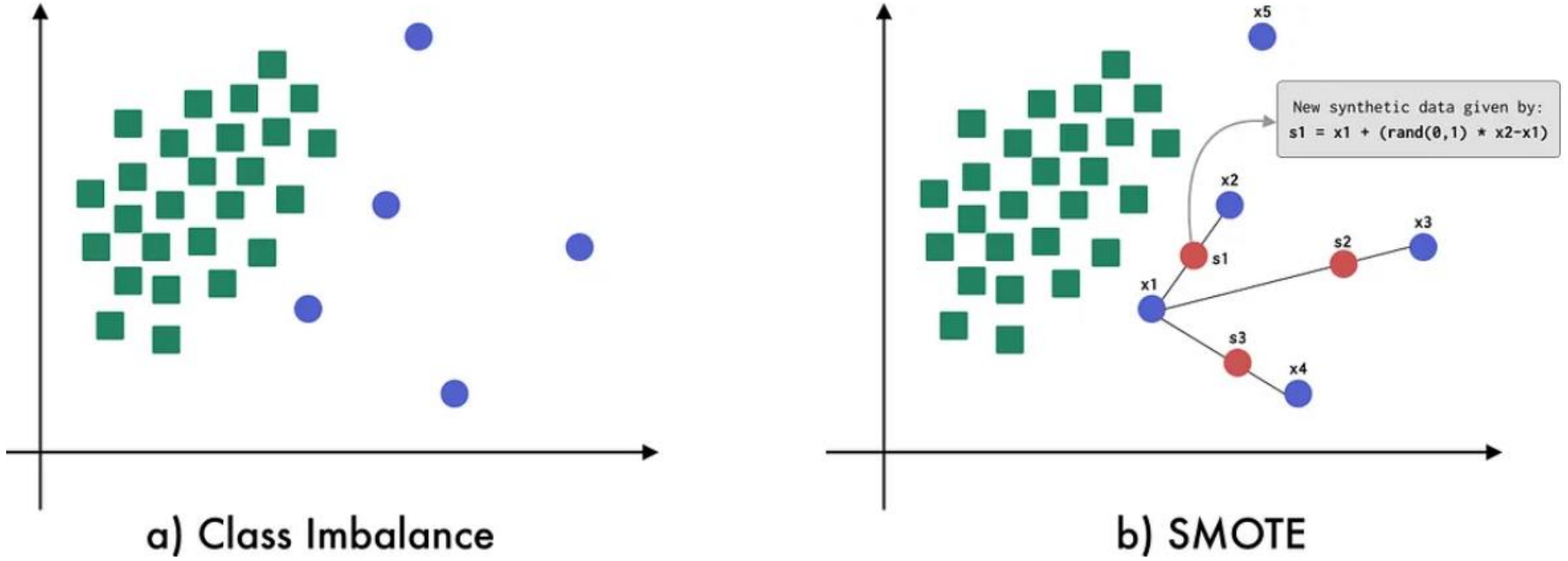
Random under sampling





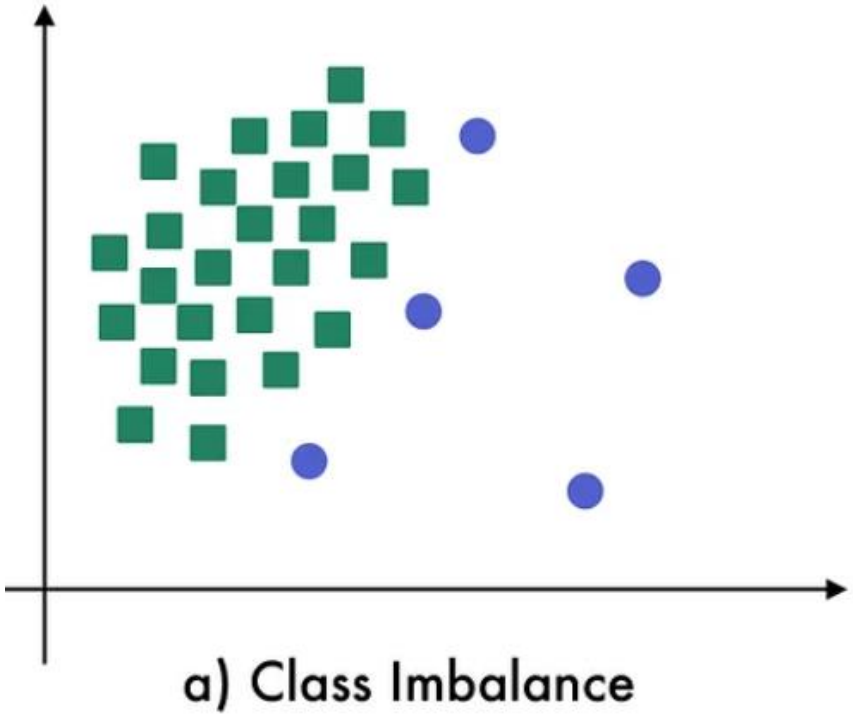
How to handle imbalance data

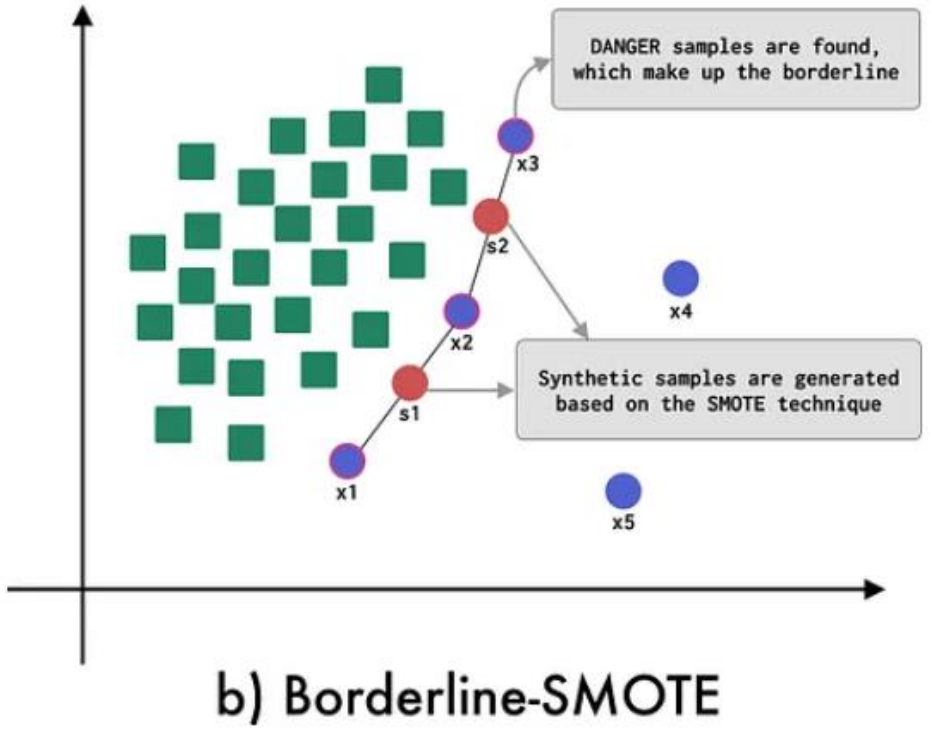
SMOTE (Synthetic Minority Over-sampling Technique)



How to handle imbalance data

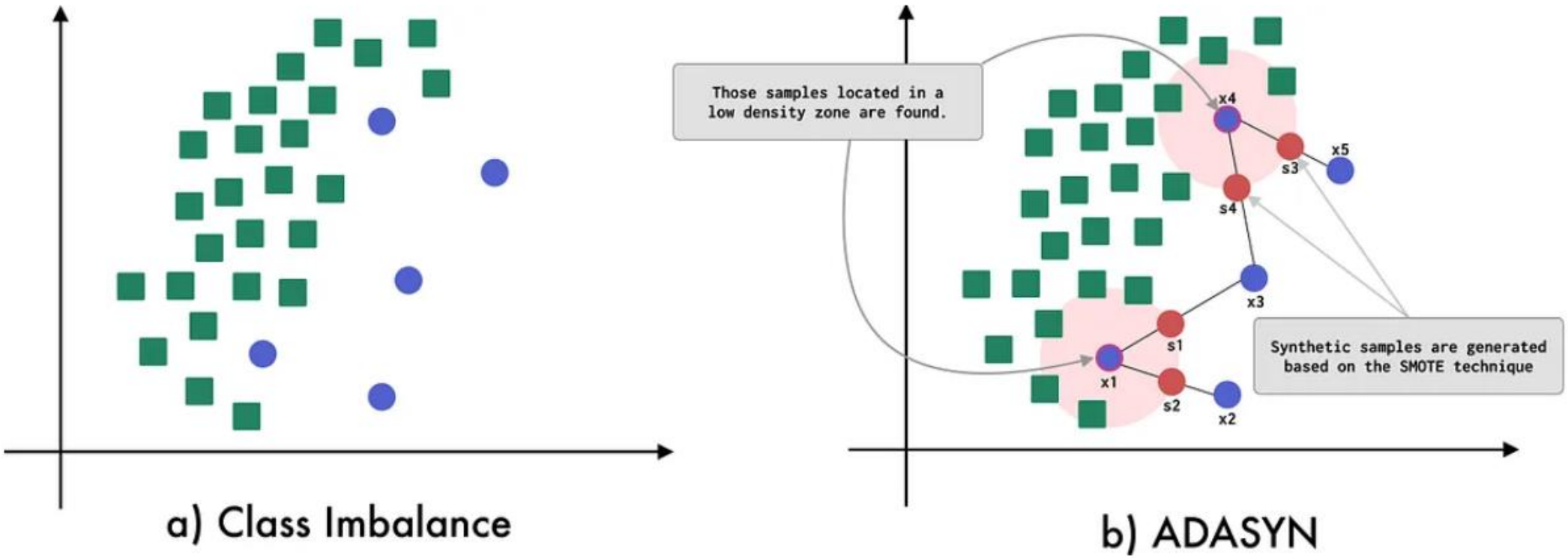
Borderline-SMOTE





How to handle imbalance data

ADASYN



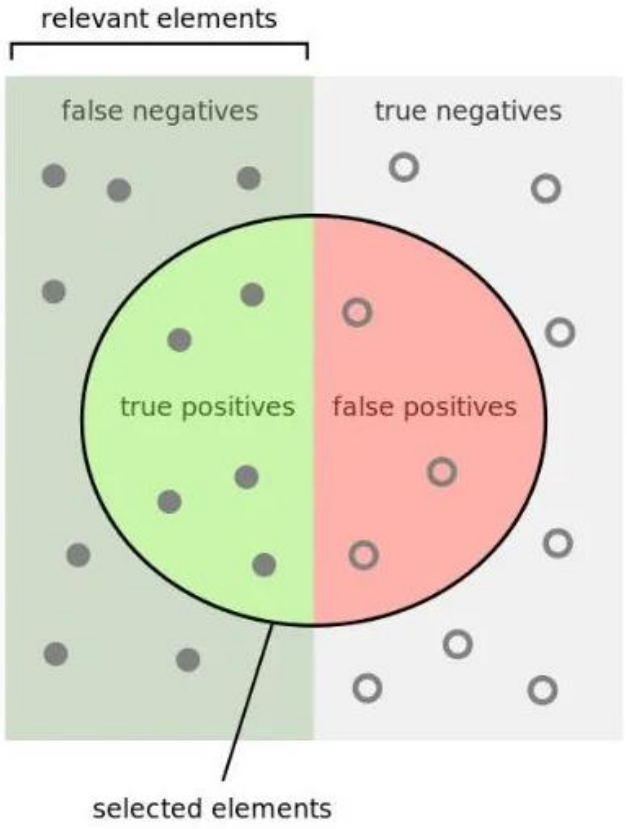
How to handle imbalance data

2. Algorithm-level method
   - Cost-sensitive learning (adjust class weight)

3. Evaluation metrics adjustment
   - ~~Accuracy~~
   - Precision and Recall
   - F-beta
   - ROC-AUC
   - PR-AUC

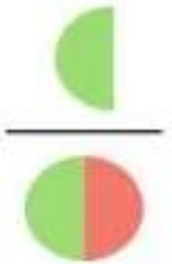
How to handle imbalance data

Precision and recall



How many selected items are relevant?

Precision =



How many relevant items are selected?

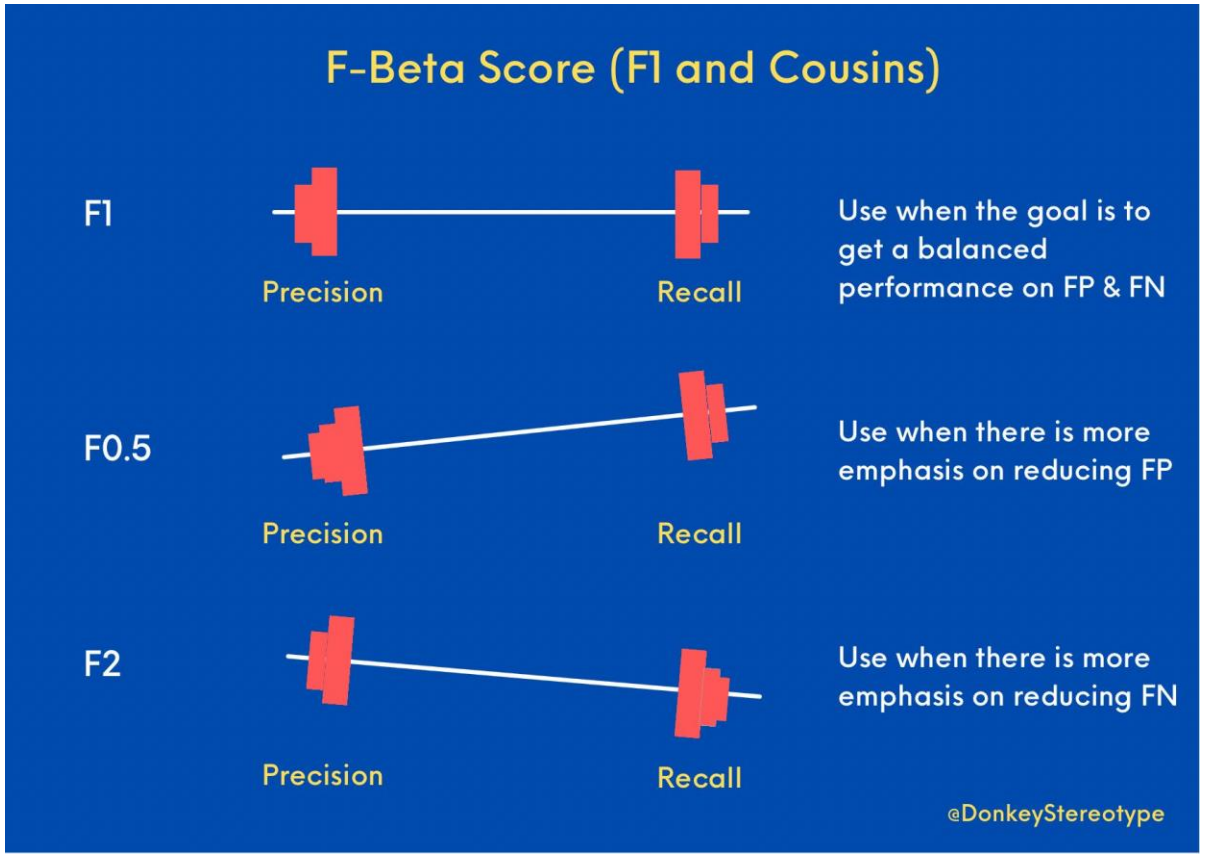
Recall =

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

How to handle imbalance data

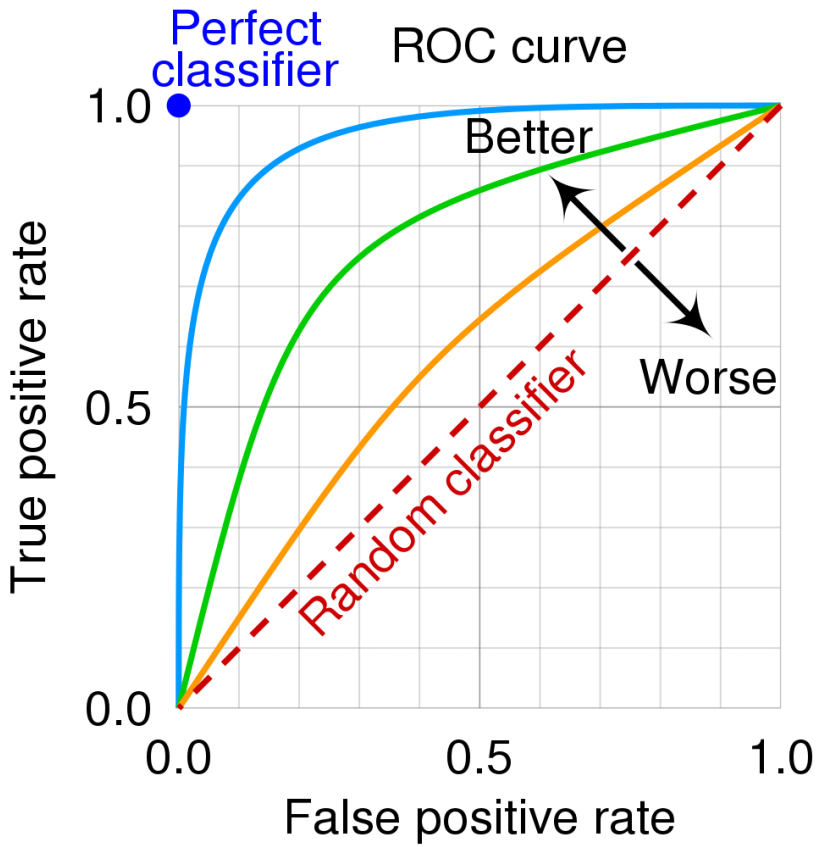
F-beta

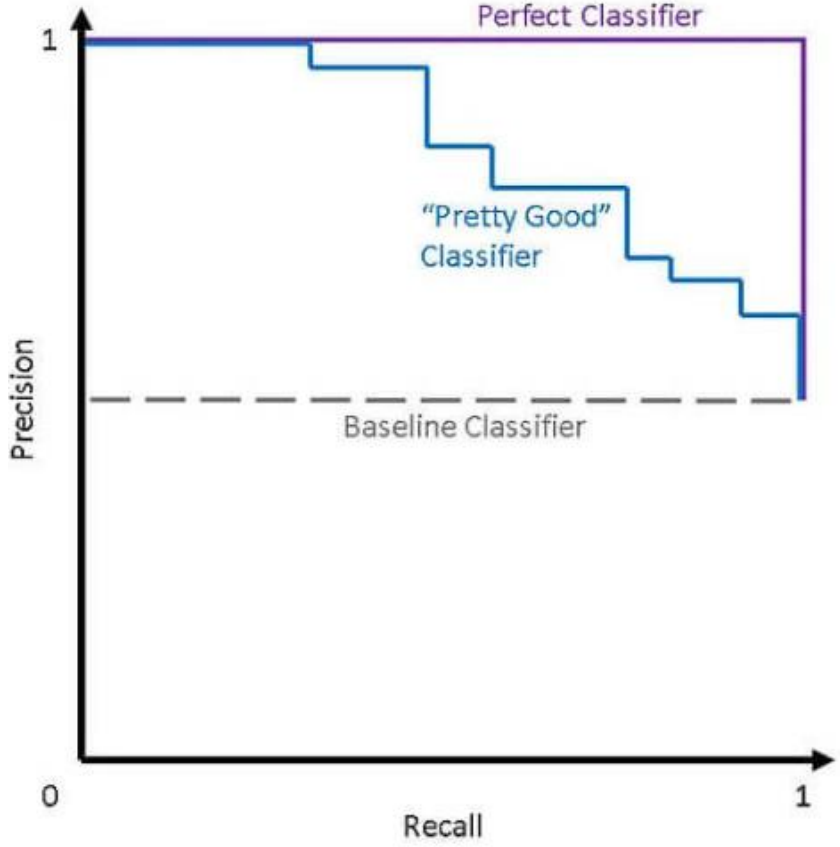


$$F_\beta = \frac{1 + \beta^2}{\frac{\beta^2}{Recall} + \frac{1}{Precision}}$$

How to handle imbalance data

ROC-AUC and PR-AUC





How to handle imbalance data

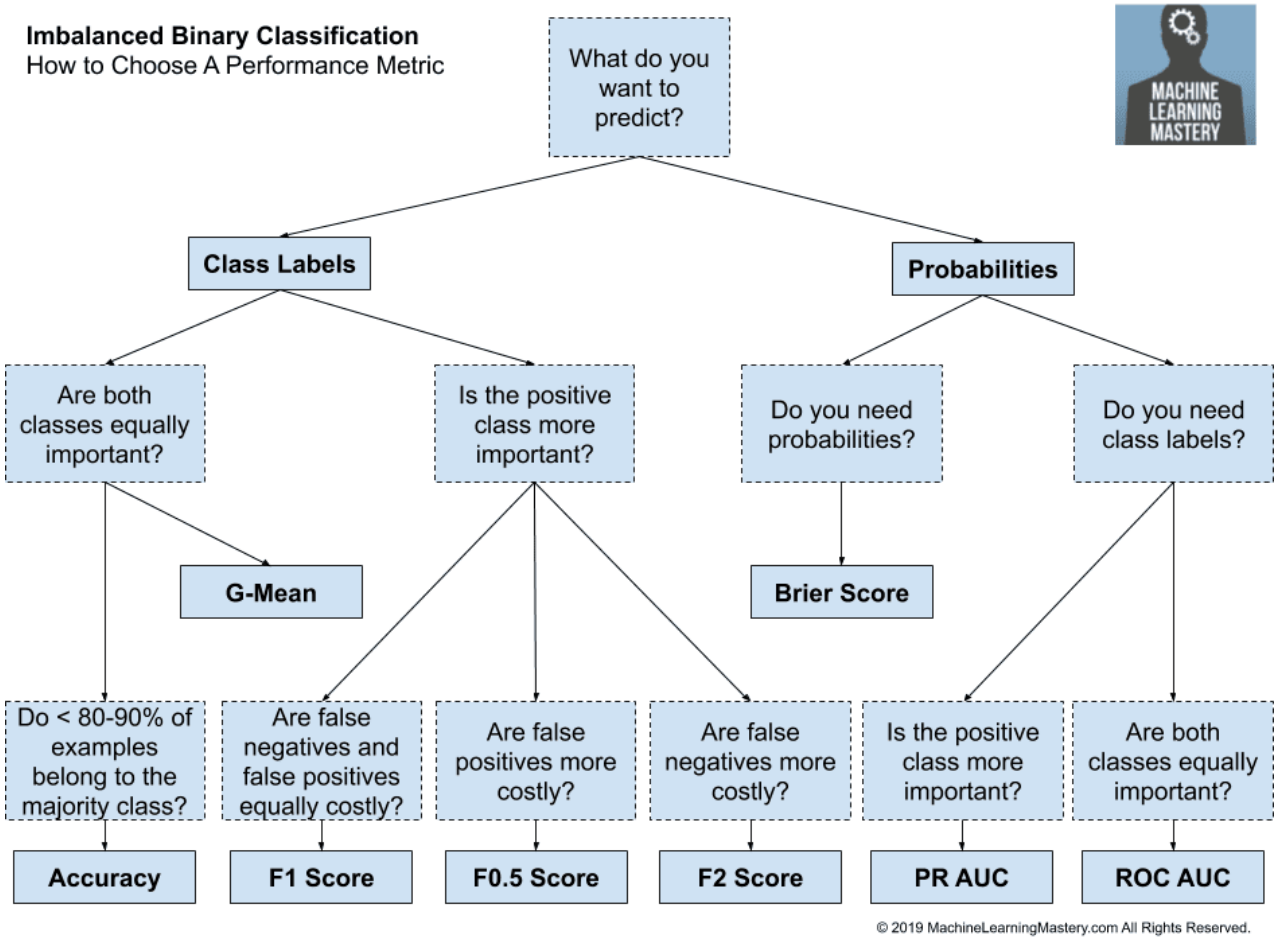
3. Hybrid method
   - Combining several techniques
   - No free lunch theorem**

Show case

1. Telco churn prediction
2. Bitcoin false signal trading system detection

Further reading

1. Multi-class imbalance data
2. Performance metrics
3. Imbalance in regression
4. Over under sampling



Appendix

The 5 Most Useful Techniques to Handle Imbalanced Datasets – Kdnuggets

Resampling strategies for imbalanced datasets (kaggle.com)

Four Oversampling and Under-Sampling Methods for Imbalanced Classification Using Python | by Amy @GrabNGoInfo | GrabNGoInfo | Medium

SMOTE: Synthetic Data Augmentation for Tabular Data | by Fernando López | Towards Data Science

Tour of Evaluation Metrics for Imbalanced Classification - MachineLearningMastery.com

Is F1 the appropriate criterion to use? What about F2, F3,…, F beta? | by Dr Barak Or | Towards Data Science

How to Deal With Imbalanced Classification and Regression Data (neptune.ai)