INL/MIS-23-72864-Revision-0

Imbalanced Classification

June 2023

Changing the World's Energy Future

Cody McBroom Walker

Idaho National Laboratory

INL is a U.S. Department of Energy National Laboratory operated by Battelle Energy Alliance, LLC

DISCLAIMER

This information was prepared as an account of work sponsored by an agency of the U.S. Government. Neither the U.S. Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness, of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the U.S. Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the U.S. Government or any agency thereof.

INL/MIS-23-72864-Revision-0

Imbalanced Classification

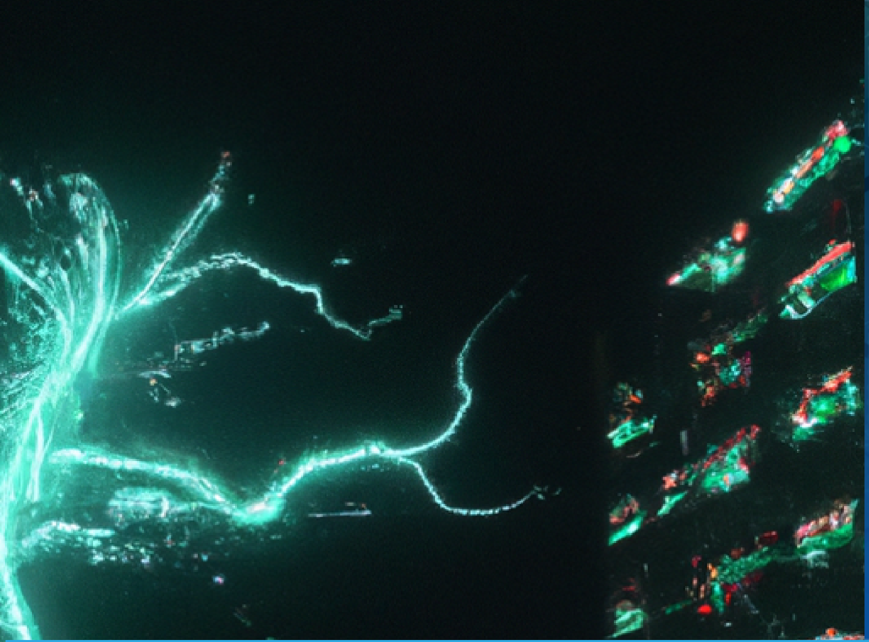
Cody McBroom Walker

June 2023

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

June 22, 2023

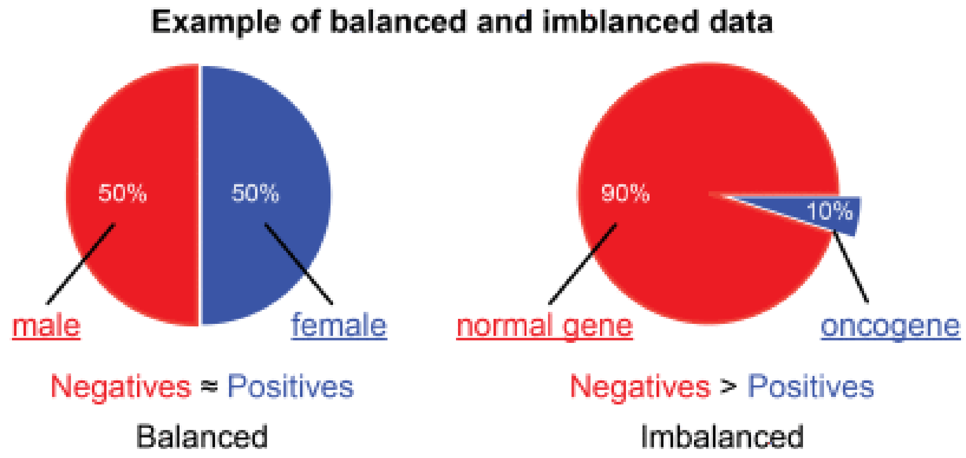
Dr. Cody Walker
Research Scientist

Imbalanced Classification

AI ML
Summer Symposium
2023

Overview of imbalanced classification problem in machine learning

- **Imbalanced classification** is where the classes of interest have significantly different sample sizes. 1:100, 1:5000.
- Examples of imbalanced classification problems include as fraud detection, disease diagnosis, and faults inside a nuclear power plant.
- Without properly handling imbalanced datasets, you'll end up with biased models and inaccurate predictions.
- Correctly identifying the minority class is often the most important thing as it often represents critical or rare instances.

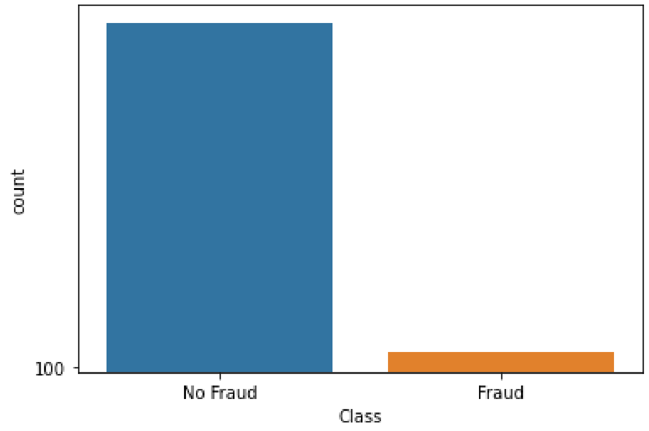


What Is Balanced And Imbalanced Dataset?
By Himanshu Tripathi Sept 24, 2019

Understanding Imbalanced Classification

- **The imbalance ratio quantifies the severity of class imbalance.**
- **Common causes of data imbalance:**
  - Natural class distribution (rare disease diagnosis)
  - Sampling bias (bias towards certain classes)
  - Data skewing (overrepresented or oversampled)
  - Rare event (anomaly detection)
  - Data loss or noise (data lost during preprocessing)
- **Data imbalance can negatively impact model performance and evaluation:**
  - Biased Decision Boundaries
  - Low Sensitivity to the Minority Class
  - Misleading Evaluation Metrics

Degree of imbalance	Proportion of Minority Class
Mild	20-40% of the data
Moderate	1-20% of the data
Extreme	<1% of the data



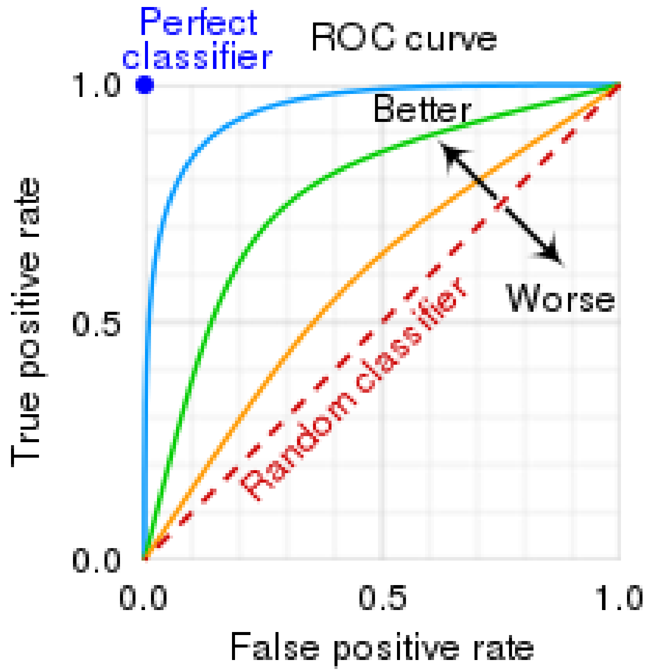
The majority class (no fraud) has many more samples than the minority class (fraud).

A Gentle Introduction to Imbalanced Classification
by Jason Brownlee on December 23, 2019 in Imbalanced Classification

Challenges in Imbalanced Classification

- Model Skewness - Imbalanced data can lead to models that are skewed towards the majority class in their predictions. This skewness can be problematic, especially when the misclassification of the minority class has severe consequences.
- Accuracy =
- Precision =
- Recall =
- F1-score =
- ROC curve is insensitive to changes in class distribution. It allows for a comprehensive evaluation of the model's ability to discriminate between the classes, regardless of their prevalence in the data.

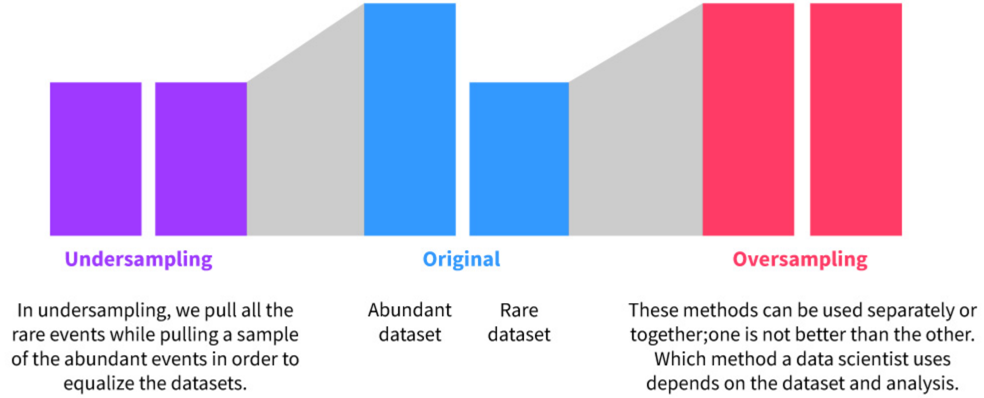
Example: If a dataset with 95% N and 5% P. A naive classifier that predicts all instances as N would achieve an accuracy of 95%. However, this accuracy does not reflect the model's performance on the minority class, which is of particular interest.



Receiver operating characteristic, Wikipedia

Techniques for Handling Imbalanced Data (Resampling methods)

1. Undersampling techniques (majority class instances are reduced to balance the dataset.)
   - **Random undersampling** - simple and computational efficient, but information loss and underutilization of majority class samples.
   - **Cluster-based undersampling** (preserve more information.)
     - aim to retain representative instances from the majority class by clustering them.
     - can be applied to identify dense regions in the majority class and selectively remove instances.
     - preserves important patterns and reduces the risk of removing informative samples.



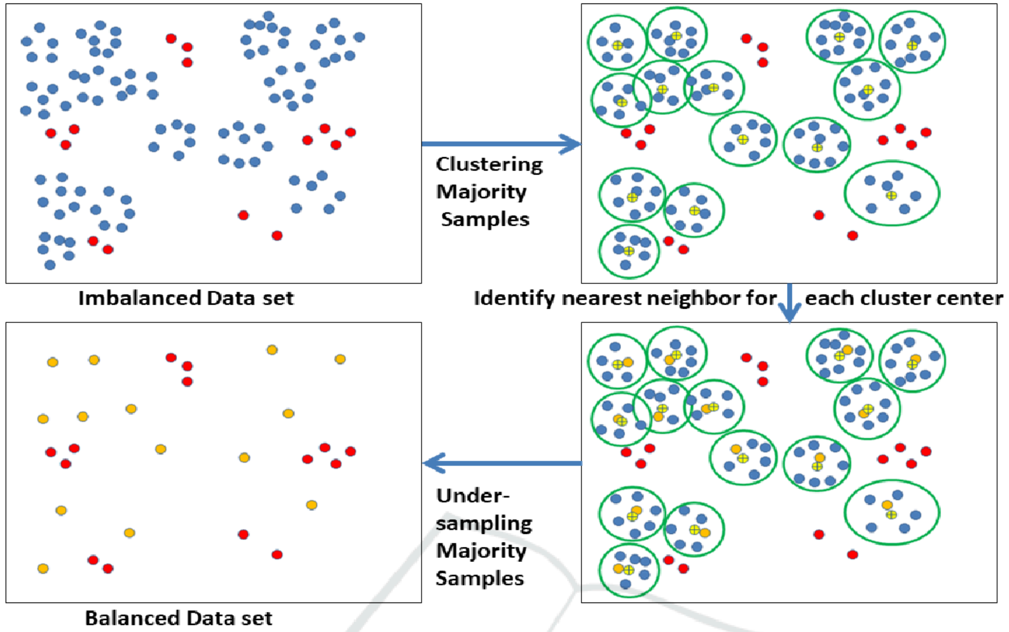
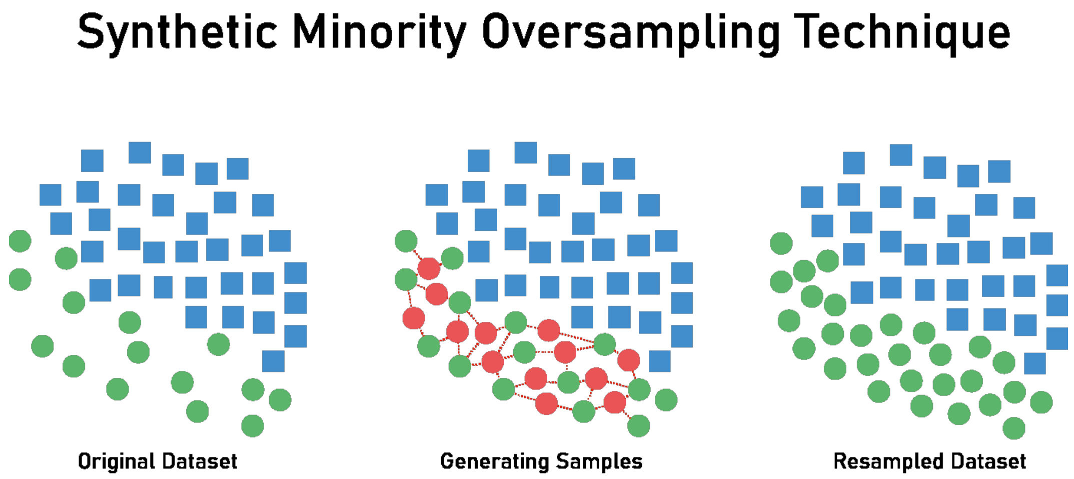


Figure 1: Clustering-based Under-sampling majority samples.

Henein, Moheb M. R. et al. "Clustering-based Under-sampling for Software Defect Prediction." *International Conference on Software and Data Technologies* (2018).

Techniques for Handling Imbalanced Data (Resampling methods)

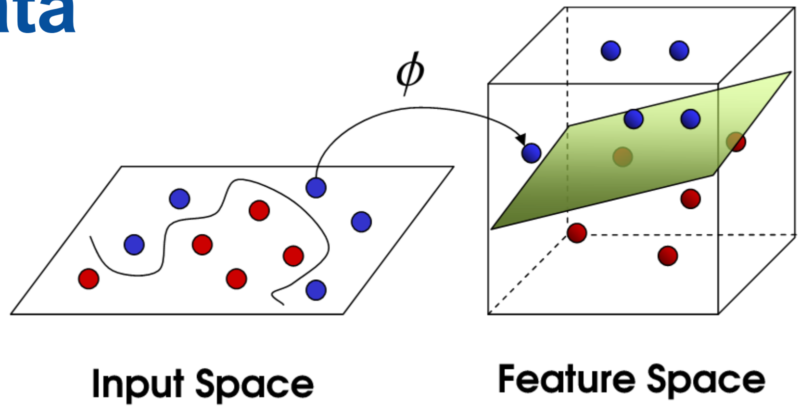
2. Oversampling techniques (generate synthetic samples to increase the minority class representation.)
   - **Random oversampling -** the minority class instances are replicated randomly to increase their representation.
     - Potential issues include overfitting and the introduction of duplicate patterns.
   - **SMOTE** (Synthetic Minority Over-sampling Technique)
     - create synthetic examples along the line segments connecting minority class instances.
     1. Select a minority instance,
     2. Identify its k nearest neighbors
     3. Create synthetic samples along the line segments connecting them.



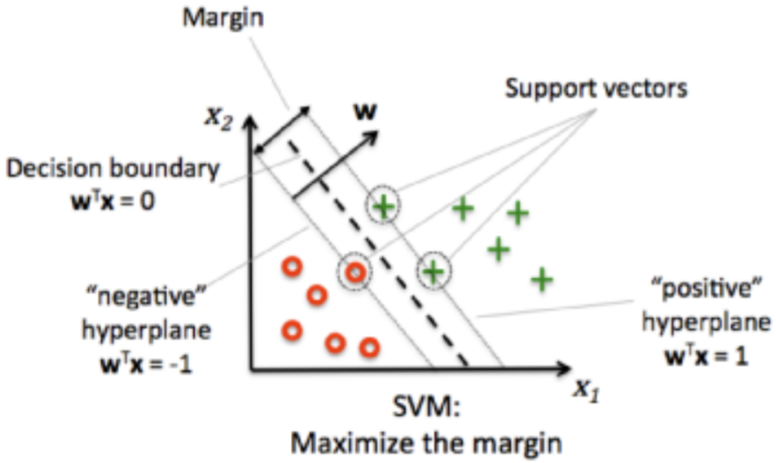
SMOTE by Emilia Orellana. Dec 9, 2020.
https://emilia-orellana44.medium.com/smote-2acd5dd09948

Techniques for Handling Imbalanced Data (Algorithmic approaches)

1. Cost-sensitive learning
   - Assigning different **misclassification costs** to balance their importance.
     - E.g., weighting a tumor classification as highly important.
     - to determine appropriate costs requires domain knowledge or likely experimental tuning.
   - **Incorporating class weights** (modifies the training process to reflect the costs)
     - Class weights can be assigned to each class to adjust the impact during model training.
     - Popular algorithms that support class weights, such as decision trees and support vector machines.



The kernel Trick in Support Vector Classifcation by Drew Wlimitis
https://towardsdatascience.com/the-kernel-trick-c98cdbcaeb3f

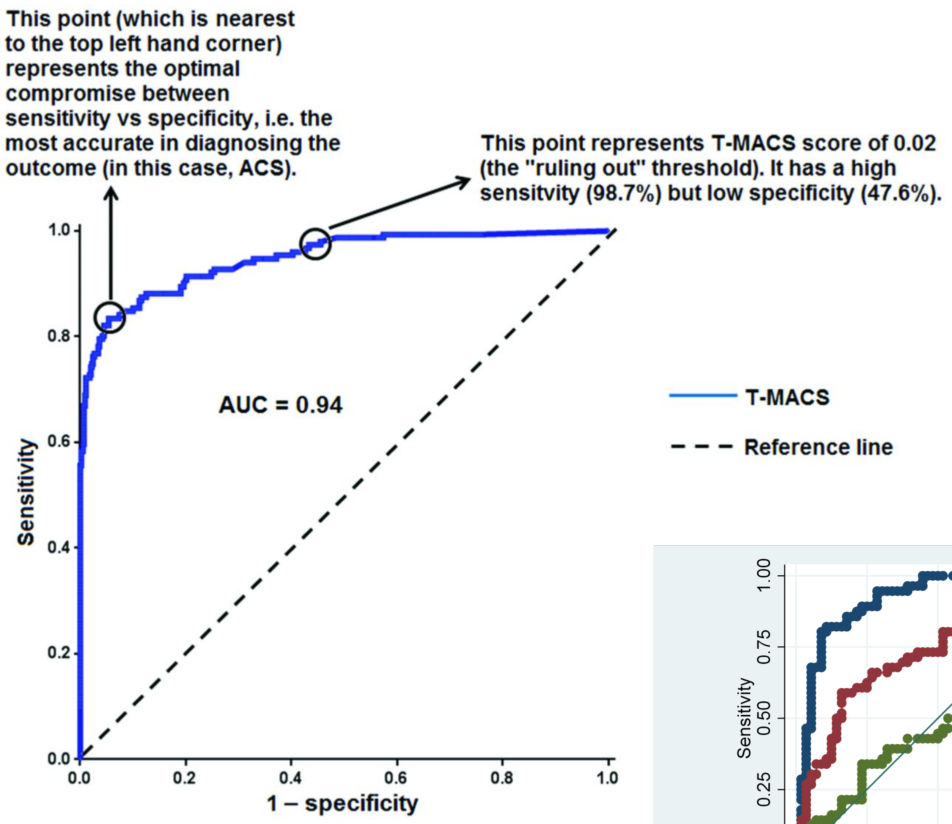


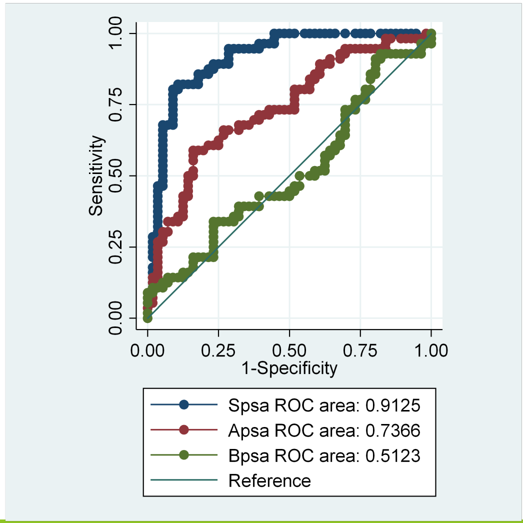
The weight matrix helps to determine the optimal hyperplane which can be exploited by modifying the class weights.

"Python Machine Learning" by Sebastian Raschka

Techniques for Handling Imbalanced Data (Algorithmic approaches)

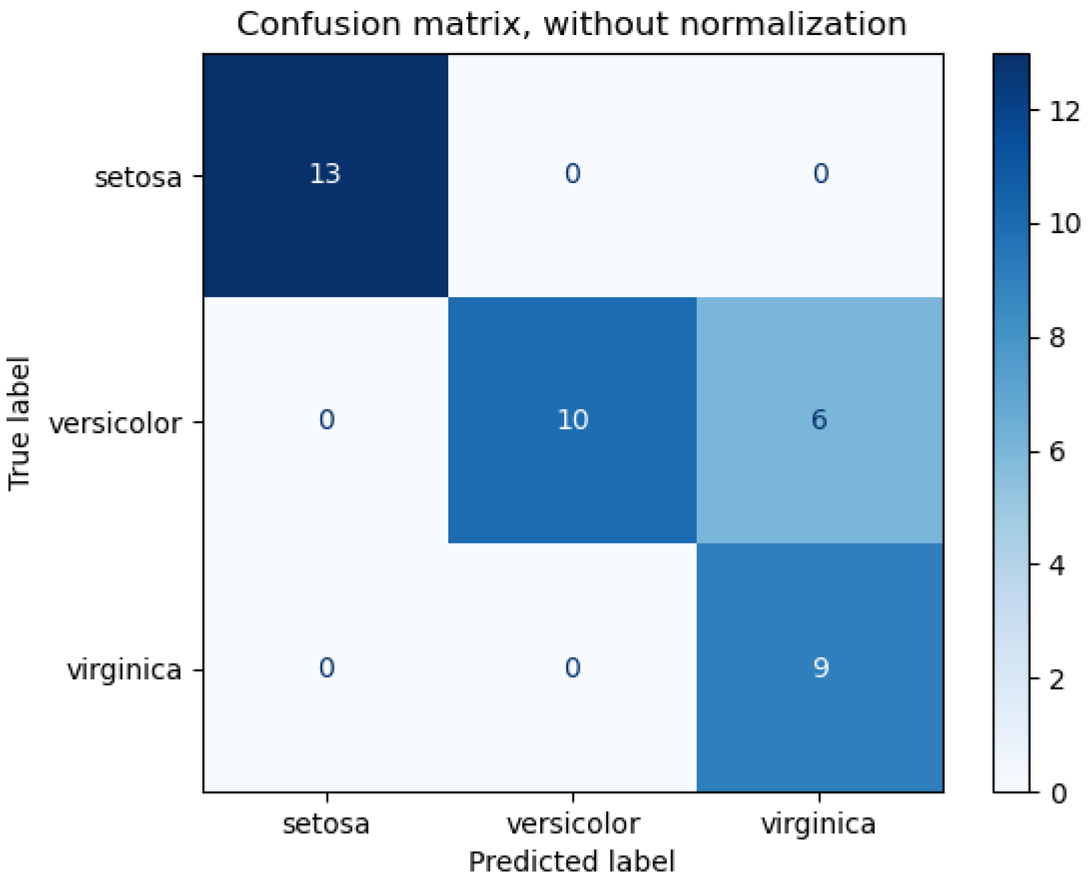
2. Threshold-moving strategies
   - Adjust decision thresholds (to address class imbalance)
     - decision thresholds on the trade-off between precision and recall.
     - can help in balancing the prediction biases towards the majority or minority class.
   - Receiver Operating Characteristic (ROC) curve analysis
     - the trade-off between true positive rate (sensitivity) and false positive rate.
     - can help determine an optimal decision threshold for imbalanced datasets.
3. Hybrid methods combining resampling and algorithmic techniques





Evaluation Metrics for Imbalanced Classification

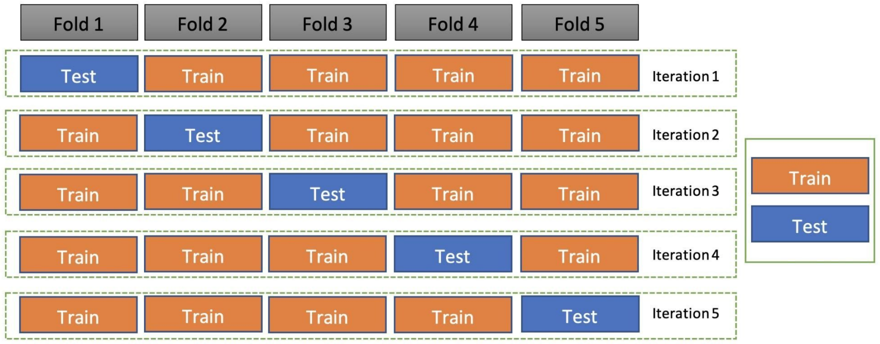
- **Confusion matrix** -tabular representation of the model's predictions.
  - the confusion matrix provides insights into the true positives, true negatives, false positives, and false negatives.
  - Can be used to determine which classes are being misclassified and with what.



https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html

Practical Considerations and Best Practices

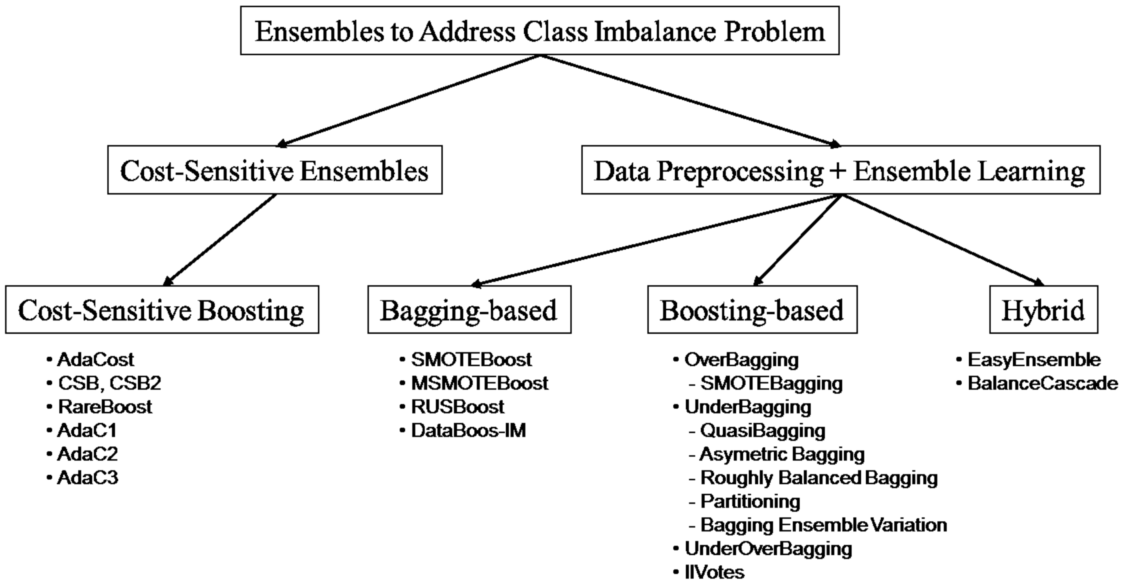
- **Feature engineering and selection**
  - Imbalanced datasets often suffer from overlapping feature distributions making it hard to accurately distinguish between them.
  - Feature engineering can help create informative features that capture the nuances between classes. E.g., creating interaction terms, deriving ratios or differences between variables.
  - Techniques such as feature scaling, dimensionality reduction, and creating informative features to improve model performance.
  - May need domain knowledge or at least exploratory data analysis to identify relevant features.
- **Cross-validation and hyperparameter tuning**
  - Cross-validation with imbalanced datasets provides a more reliable estimate of how well the model will perform on unseen data. Performance can be highly sensitive to the particular data split.
  - Other strategies include stratified sampling and k-fold cross-validation to ensure representative evaluation.
  - Hyperparameter tuning to optimize model performance includes model parameters as well as adjusting class weights, misclassification penalties, and fine-tuning thresholds.



https://www.turing.com/kb/different-types-of-cross-validations-in-machine-learning-and-their-explanations

Practical Considerations and Best Practices

- **Model selection and ensemble techniques**
  - selecting appropriate models for imbalanced classification. (e.g., SVM and decision tree can weight class importance)
  - Ensemble techniques, such as bagging and boosting, can improve the performance and robustness of models.
  - Combining multiple models can help in handling class imbalance.
- **Monitoring model performance over time**
  - Concept drift can lead to changes in the class distribution over time. For a minority class, this could have large implications on the mode.
  - Monitor model performance over time and adapt to changes in the data distribution.
  - Techniques such as online learning and updating models periodically to maintain their effectiveness.

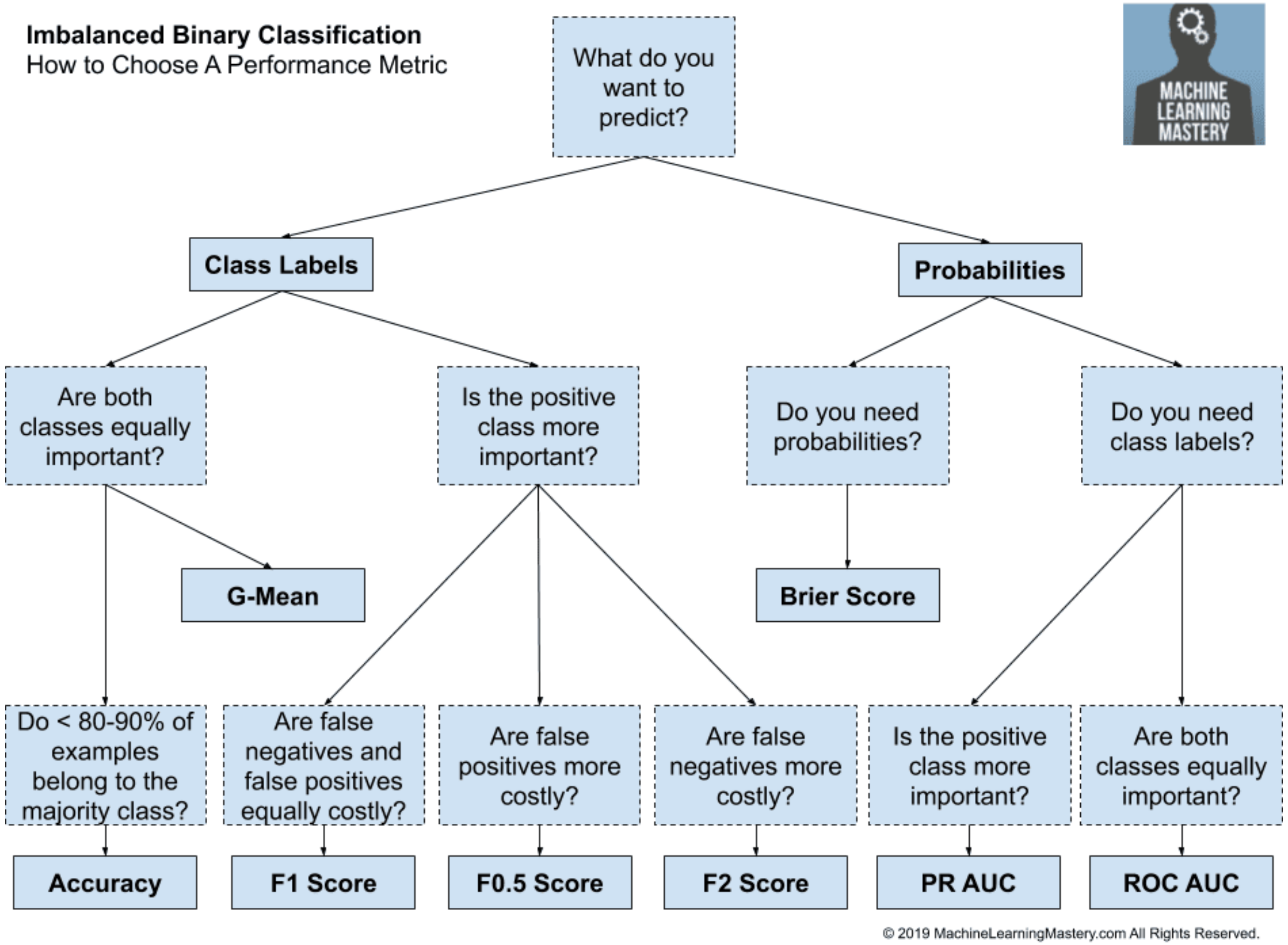


V. López, A. Fernandez, S. Garcia, V. Palade and F. Herrera, *An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics.* Information Sciences 250 (2013) 113-141 doi: 10.1016/j.ins.2013.07.007

From start to finish for fraud detection.

1. **Problem Identification**: The imbalance exists in credit card fraud detection, where fraudulent transactions are relatively rare compared to legitimate ones. The goal is to develop a model that can accurately identify fraudulent transactions while minimizing false positives.
2. **Data Collection**: Historical credit card transaction data is collected, including features such as transaction amount, location, time, cardholder information, and more. The dataset contains a small proportion of fraudulent transactions compared to legitimate ones, resulting in class imbalance.
3. **Data Preprocessing**: Initially, data preprocessing techniques are applied to handle class imbalance. This may involve undersampling the majority class, oversampling the minority class, SMOTE, or using hybrid approaches.
4. **Feature Engineering**: Feature engineering techniques are employed to extract meaningful information from the data. This may involve creating new features such as transaction frequency, aggregating transaction amounts over time, or incorporating external data sources to enhance the model's ability to capture fraud patterns.
5. **Model Selection and Training**: Various classification algorithms, such as logistic regression, decision trees, random forests, or gradient boosting, are trained on the imbalanced dataset. During model training, techniques like stratified sampling, cross-validation, or using appropriate evaluation metrics like precision, recall, or F1 score are employed to assess and compare model performance.
6. **Hyperparameter Tuning**: Hyperparameter tuning is performed to optimize the model's performance on the imbalanced data. This may involve adjusting parameters related to class weights, regularization strength, learning rates, or tree depths to find the optimal balance between sensitivity and specificity for detecting fraud.
7. **Model Evaluation**: The model is evaluated using appropriate evaluation metrics that emphasize the performance on the minority class. Metrics such as recall, precision, F1 score, or area under the precision-recall curve (AUC-PR) are commonly used to assess the model's ability to identify fraudulent transactions effectively.
8. **Threshold Adjustment**: As imbalanced classification problems often involve a trade-off between false positives and false negatives, the decision threshold of the model can be adjusted to balance the desired level of fraud detection with acceptable false positive rates. This adjustment can be based on the costs associated with misclassifications or domain-specific considerations.
9. **Monitoring and Adaptation**: Once the model is deployed in a real-time environment, it needs to be continuously monitored for concept drift. Drift detection techniques can be employed to identify changes in the data distribution and trigger model retraining or adaptation when necessary. This ensures the model maintains its performance over time.
10. **Iterative Improvement**: The entire process is iterative, involving continuous monitoring, evaluation, and retraining of the model as new data becomes available and the fraud landscape evolves. The model can be refined further by incorporating additional features, experimenting with different algorithms, or exploring ensemble methods to enhance performance.

Questions?



Tour of Evaluation Metrics for Imbalanced Classification
by [Jason Brownlee] on January 8, 2020 in [Imbalanced Classification]

Battelle Energy Alliance manages INL for the U.S. Department of Energy's Office of Nuclear Energy. INL is the nation's center for nuclear energy research and development, and also performs research in each of DOE's strategic goal areas: energy, national security, science and the environment.