AMMnet Annual Meeting
Kigali 2024

Training Session on:
Computational Techniques for Handling Health (Malaria) Data Imbalance in Machine Learning-Based Modeling (18th April 2024)

Instructor: Prof. (Dr.) O. Olawale Awe (olawaleawe@gmail.com)

Summary: This intensive 3-hour training program aims to empower participants with a thorough understanding and practical skills in mitigating imbalanced health (malaria) datasets through computational techniques in machine learning, utilizing the R software. Focused on interactive learning, it includes presentations, hands-on coding exercises, group discussions, and an engaging case study. This training fosters collaboration, enabling participants to share experiences and collectively tackle real-world challenges, ultimately enhancing their ability to make informed decisions when modeling malaria data.

Duration: 3 hours

Training Aims:

1. **Develop Basic Understanding:**
   - Provide participants with a comprehensive understanding of the challenges of imbalanced health (malaria) datasets in machine learning-based modeling.

2. **Hands-on Skill Development:**
   - Equip participants with practical skills in employing computational techniques, specifically using R, to address data imbalance issues in health-related, malaria datasets.

3. **Enhance Policy Decision-Making:**
   - Enable participants to make informed decisions when selecting and implementing appropriate machine learning algorithms, preprocessing techniques, and evaluation metrics for imbalanced health data.

4. **Promote Collaboration and New Research Groups:**
   - Foster a collaborative learning environment where participants can share insights, experiences, and challenges related to malaria modeling and diagnosis using machine learning.

A.Introductory Notes: Advantages of Using Machine Learning for Diagnosing Imbalanced Malaria Data

Machine learning (ML) offers several advantages for diagnosing imbalanced malaria data, a common challenge in medical research where the number of cases in different classes (e.g., infected vs. non-infected) is unevenly distributed. These advantages improve the accuracy, efficiency, and overall effectiveness of malaria diagnosis and research. Here are some key benefits:

1. **Improved Prediction Accuracy:** ML algorithms can be trained to recognize complex patterns in malaria data that may not be immediately apparent to humans. Even with imbalanced datasets, certain techniques like weighted classes, oversampling the minority class, or undersampling the majority class can help in adjusting the learning process to improve prediction accuracy for both common and rare events.

2. **Automated Feature Selection**: Machine learning can automatically identify the most relevant features for distinguishing between different outcomes, such as identifying which clinical or demographic factors are most predictive of malaria infection. This is particularly useful in imbalanced datasets where the signal-to-noise ratio can be low.

3. **Adaptability:** ML models can be designed to adapt to new data over time, learning from further cases to improve their diagnostic predictions. This adaptability is crucial for malaria diagnosis, as the parasite's resistance patterns and epidemiological characteristics can change.

4. **Cost-Effectiveness:** By automating the diagnostic process, ML can help reduce the costs associated with laboratory tests and manual data analysis. This is especially beneficial for resource-limited settings where malaria is most prevalent.

5. **Handling High-Dimensional Data:** ML algorithms are well-suited to manage high-dimensional data (e.g., genetic information, environmental factors) that can be relevant in understanding malaria but challenging to analyze with traditional statistical methods, especially when the data are imbalanced.

6. **Anomaly Detection:** ML models, particularly those designed for imbalanced datasets, are effective at identifying outliers or rare events. This capability is critical for detecting unusual cases of malaria that might not fit the typical infection patterns.

7. **Scalability:** Machine learning models can easily scale to accommodate large datasets, making them suitable for analyzing data from widespread surveillance efforts and population studies on malaria.

8. **Enhanced Decision Support:** By providing more accurate predictions, ML models can serve as valuable decision-support tools for healthcare professionals, aiding in the prioritization of cases based on predicted severity or likelihood of infection.

9. **Cross-Population Generalizability:** Advanced ML techniques can be applied to develop models that generalize across different populations, taking into account the imbalance in malaria prevalence among different demographic groups or geographic regions.

10. **Integration with Other Data Sources:** ML models can integrate diverse data types (clinical records, laboratory results, environmental data) to improve the accuracy of malaria diagnosis, even in cases where direct data may be sparse or imbalanced.

By leveraging these advantages, machine learning can significantly enhance the diagnosis and understanding of malaria, particularly in the face of imbalanced datasets, which pose challenges for traditional analytical approaches.

B.Common Machine learning (ML) models for Modeling and Diagnosis of Malaria

Machine learning (ML) models have become integral tools in the modeling and diagnosis of malaria, offering insights into its spread, risk factors, and treatment outcomes. These models are adept at handling complex, non-linear relationships within large datasets, including those with imbalanced classes often found in disease data. Here are some ML models and techniques commonly used in malaria research:

1. **Decision Trees and Random Forests:** These models are particularly useful for classification tasks, such as diagnosing malaria from clinical or laboratory data. Random Forests, an ensemble of Decision Trees, are effective in reducing overfitting and improving prediction accuracy. They can handle high-dimensional data well and provide insights into the importance of different features for malaria diagnosis.

2. **Support Vector Machines (SVMs)**: SVMs are powerful for binary classification problems, like distinguishing between malaria-infected and non-infected individuals. They work well with high-dimensional data and are effective in cases where the number of features exceeds the number of samples. SVMs can be particularly useful in diagnosing malaria using genetic or proteomic data.

3. **Neural Networks and Deep Learning:** Deep learning models, including Convolutional Neural Networks (CNNs), have shown promise in analyzing medical imaging data, such as blood smear images for malaria parasite detection. These models can automatically learn features from images, eliminating the need for manual feature extraction and significantly improving diagnostic accuracy.

4. **Gradient Boosting Machines (GBMs):** GBMs, including XGBoost and LightGBM, are powerful ensemble learning models that build trees in a sequential manner, where each tree attempts to correct the errors of the previous ones. They have been effective in malaria risk prediction, leveraging epidemiological and environmental data.

5. **Logistic Regression:** Despite being a simpler model, logistic regression can be very effective in binary classification problems, such as predicting the presence or absence of malaria. It's particularly useful when the relationship between the feature variables and the outcome is approximately linear.

6. **K-Nearest Neighbors (KNN)**: KNN can be used for both classification and regression problems in malaria research. For instance, it can classify individuals based on the similarity of their symptoms or genetic makeup to known cases of malaria.

7. **Principal Component Analysis (PCA) and Cluster Analysis:** While not predictive models themselves, PCA and clustering techniques like K-means can be used for dimensionality reduction and to identify patterns or groupings within malaria data, which can then inform more targeted analysis or modeling.

8. **Time Series Analysis and Forecasting Models:** Models like ARIMA (Autoregressive Integrated Moving Average) can be used to forecast malaria incidence or prevalence over time, based on historical data. This model is valuable for understanding seasonal trends and planning public health interventions.

9. **Naive Bayes :** These models can handle uncertainty and incomplete data well, making them suitable for modeling complex epidemiological relationships in malaria, such as the interaction between host, pathogen, and environmental factors. A typical

Each of these models offers unique advantages for malaria modeling and diagnosis, and the choice of model can depend on the specific task at hand (e.g., classification, regression, clustering), the nature of the data available (e.g., images, time series, tabular data), and the computational resources at disposal. Combining multiple models in an ensemble approach can sometimes yield better performance than any single model alone, while taking advantage of the strengths of each.

C. Practical Activity: Applying ML Models to Malaria Data

1. **Activity Description:** Using the malaria dataset provided (Malaria-Data.csv), participants will:

- Apply a Balanced Random Forest model.
- Perform feature selection to identify the most critical features (symptoms) for predicting (diagnosing) malaria.
- Compare the performance of this model to other ML models using precision, recall, F1 score, and AUC-ROC.
- Conduct feature importance using the best-performing model

2. **Goals:**

- Understand how ensemble ML methods can be tailored for imbalanced datasets.
- Learn the impact of feature engineering and selection on model performance in the context of health data.

3. **Expected Outcome:** Participants will gain hands-on experience with advanced strategies for managing data imbalance, enhancing their ability to develop more effective machine learning models for health-related data, specifically focusing on diseases like malaria.

Through these practical notes and activities, participants will have explored a comprehensive set of strategies for addressing data imbalance, with a special focus on health data scenarios. By applying these techniques in R, you'll be better equipped to tackle similar challenges in their future work, ensuring their models are both accurate and sensitive to the intricacies of imbalanced datasets.

For a practical session focused on handling health data imbalance, particularly with malaria data using R, let's engage in some practical activities. These notes will guide you through the process of addressing data imbalance in a machine learning model, including data preparation, applying resampling techniques, and evaluating model performance with a focus on a malaria dataset.

Setting Up the Caret Environment

First, ensure you have R and RStudio installed. Open RStudio and install the necessary packages by running:

```
install.packages(c("caret", "ROSE", "smotefamily"))
```

Loading the Malaria Dataset

Assume we have a dataset named `Malaria-Data.csv` which includes features related to patient symptoms, blood test results, and a binary target variable indicating whether malaria was detected (1) or not (0) called severe_maleria. The data contains 15 malaria symptoms with 0 indicating absence of the symptom while 1 indicates the presence of the symptom.

Load necessary libraries

```
library(caret)
library(ROSE)
library(dplyr)
library(smotefamily)
```

Load the dataset

```
malaria_data <- read.csv("Malaria-Data.csv")
```

Exploratory Data Analysis (EDA)

Perform a quick EDA to understand the imbalance:

Summary of target variable

```
table(malaria_data$severe_maleria)
```

Visualize the imbalance

```
library(ggplot2)
ggplot(malaria_data, aes(x = factor(severe_maleria))) + geom_bar() + labs(x = "Malaria Detected", y = "Count")
```

Handling Data Imbalance using the R workbook provided (Malaria-Resampling.R)

Resampling Techniques

Oversampling Minority Class

```
set.seed(123)  For reproducibility
up_sampled_data <- upSample(x = malaria_data[, -ncol(malaria_data)],
                         y = factor(malaria_data$severe_maleria))
```

Checking the balance

```
table(up_sampled_data$Class)
```

Exercise:
Repeat the model training and evaluation for `up_sampled_data`, ``down_sampled_data’, and `hybrid_sampled_data` to compare results for different machine learning models. Remember to partition each balanced dataset into training and testing sets as shown in the codes for the original data.

Evaluation Metrics

Focus on metrics beyond accuracy, such as Precision, Recall, F1 Score, and AUC-ROC, to truly understand the performance of your model on the imbalanced dataset.

Example for calculating AUC-ROC for original data model

```
library(pROC)
roc_response <- roc(response = test_orig$severe_maleria, predictor = as.numeric(pred_orig))
plot(roc_response)
auc(roc_response)
```

This session would guide you through handling data imbalance with a focus on malaria data, using techniques like oversampling and synthetic data generation with ROSE in R. By comparing the models trained on original and balanced datasets, you would see the impact of addressing data imbalance on model performance. Continue to explore these techniques and apply them to different datasets and machine learning models to gain deeper insights and improve your predictive modeling skills in health data analytics.

Interpreting Your Machine Learning Models

Interpreting machine learning models, especially in the context of imbalanced datasets for tasks like disease diagnosis or fraud detection, requires careful consideration of the metrics that best reflect the model's performance in distinguishing between classes. When precision and F1 score are not available, but specificity is, along with other metrics like sensitivity (recall), accuracy, and the Area Under the Receiver Operating Characteristic curve (AUC-ROC), we can still gain valuable insights into the model's performance.

Understanding Key Metrics

- **Accuracy:** Measures the proportion of true results (correct predictions-both true positives and true negatives) in the total population. While useful, **accuracy can be misleading in imbalanced datasets,** as it may reflect the underlying class distribution more than the model's ability to classify correctly.

Accuracy= (TP + TN)/ (TP+ TN+ FP+ FN)

- **Sensitivity (Recall):** Measures the proportion of actual positive cases that are correctly identified. **High sensitivity is crucial in medical diagnosis to ensure as few false negatives as possible** (e.g., ensuring malaria cases are not missed). It is known as the true positive rate. It is the ability of a test or model to correctly identify patients with malaria. High sensitivity means most people who have malaria are correctly identified by the model.

Sensitivity = TP/(TP+FN)

- **Specificity**: Measures the proportion of actual negative cases that are correctly identified as negative. High specificity means that the model is good at identifying negative cases (e.g., **correctly identifying individuals without malaria),** reducing false positives. It is the ability of a model to correctly identify patients who do not have malaria. High specificity means most people who do not have malaria are correctly identified as not having it.

Specificity = TN/(TN+FP)

What does it mean when Specificity is higher than Sensitivity?

When specificity is higher than sensitivity in the context of malaria diagnosis, it implies a particular emphasis on correctly identifying those who do not have malaria, while potentially sacrificing the ability to identify all those who do have the disease. Let's break down what this means and consider its implications:

In the scenario where specificity is higher than sensitivity for a malaria diagnostic test, the following points are critical:

1. **Avoiding False Positives:** The model is very good at ensuring that those who are diagnosed as not having malaria truly do not have the disease. This is particularly important in areas where malaria treatments might be costly, have significant side effects, or where misdiagnosis could lead to inappropriate use of malaria medications, contributing to drug resistance.

2. **Potential Missed Cases:** A higher specificity often comes at the expense of sensitivity, which means there might be a higher number of false negatives. In practical terms, this implies that

some individuals who actually have malaria might not be diagnosed by the test. This can be dangerous as untreated malaria can lead to severe complications and even death.

3. **Usage in Low Prevalence Areas:** In areas where malaria is rare, a test with high specificity and lower sensitivity might be preferred to minimize the number of false positive results, which can lead to unnecessary anxiety, treatment, and further testing.

4. **Supplemental Testing:** Because of the risk of false negatives, additional tests or follow-ups may be necessary for those who test negative but present strong clinical symptoms of malaria. This is especially important in endemic regions where the probability of malaria is high.

Clinical Decision Making

In clinical settings, the choice between prioritizing sensitivity or specificity depends on the consequences of false positives versus false negatives:

- Prioritizing Specificity: This is often chosen to prevent overdiagnosis and overtreatment, which is crucial in managing resources and avoiding side effects in populations where the probability of malaria is low.

- **Impact of False Negatives**: In a disease like malaria, missing an actual case (false negative) can be life-threatening. Therefore, while a high specificity is valuable, it must be balanced carefully with sufficient sensitivity to ensure effective management of the disease.

Adjusting Diagnostic Strategies

In medical practice, no diagnostic test is perfect. Physicians often use a combination of diagnostic tests, clinical judgment, and epidemiological information to make the best decisions. In the case of malaria:

- Initial screening might use a test with high specificity to rule out malaria in clear cases.
- Subsequent testing with a more sensitive test might be used for borderline cases or in patients with symptoms but a negative initial test.

Overall, the choice of diagnostic tests and the interpretation of their results should be guided by the specific clinical and epidemiological context, aiming to balance the benefits of correct diagnosis with the risks associated with false positives or false negatives.

- **AUC-ROC:** Represents the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative one. AUC-ROC is a powerful aggregate measure of performance across all possible classification thresholds, not affected by the imbalance in the data. A high AUC-ROC value suggests that the model has a good measure of separability; it can distinguish between positive and negative classes effectively across all thresholds. This metric becomes especially important in the absence of precision and F1 scores, as it provides insight into the model's ability to balance sensitivity and specificity.

In the context of malaria diagnosis, a high Area Under the Curve (AUC) value typically indicates a strong diagnostic performance when using a Receiver Operating Characteristic (ROC) curve. The ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

Significance of High AUC in Malaria Diagnosis

1. **Effective Discrimination:** A high AUC value in malaria diagnosis means that the test is very effective at distinguishing between those who have malaria and those who do not. This is crucial in malaria-endemic areas to ensure accurate treatment and management.
2. **Reduced Misdiagnosis:** With high sensitivity and specificity, fewer false negatives and false positives occur. This reduces the likelihood of untreated malaria cases (which can be fatal) and prevents the unnecessary treatment of individuals without malaria, avoiding wastage of resources and potential side effects from antimalarial drugs.
3. **Optimal Threshold Setting:** High AUC also indicates that there might be an optimal threshold at which the test performs well in both identifying true positives and true negatives, providing flexibility in adjusting the test based on the clinical scenario or resource availability.
4. **Comparative Evaluation:** When comparing different diagnostic tests or algorithms for malaria, a higher AUC provides evidence that one test may be superior to another in terms of overall diagnostic accuracy.

Practical Implications

In practice, a high AUC for a malaria diagnostic test, such as a rapid diagnostic test (RDT) or PCR-based methods, supports its deployment in clinical and field settings. This can particularly aid in rapid decision-making, which is vital during malaria outbreaks or in remote areas where laboratory facilities are limited.

Thus, a high AUC is indicative of a reliable and efficient test, crucial for controlling malaria effectively by ensuring that patients receive timely and appropriate treatment. It also helps in the strategic planning of malaria control programs by identifying and deploying the most effective diagnostic tools.

Precision: Precision, also known as positive predictive value (PPV), is a statistical measure used to evaluate the accuracy of a binary classification test, specifically focusing on the performance of the test in identifying positive instances. It reflects the proportion of correct identifications.

The formula for precision is given by: TP/(TP+FP)

Precision is crucial in the diagnosis of malaria due to the serious implications of false positive results. Here's how precision impacts malaria diagnosis and treatment:

1. **Avoiding Unnecessary Treatment:** Malaria treatment typically involves strong antimalarial drugs that can have significant side effects. High precision in diagnostic tests ensures that only patients who truly have malaria receive these medications, thereby minimizing the exposure to potentially harmful side effects in patients who do not need them.

2. **Reducing Healthcare Costs:** Treating malaria can be costly, especially in resource-limited settings where malaria is most prevalent. By ensuring that only patients with true positive results receive treatment, precision helps in conserving valuable healthcare resources and reducing unnecessary expenditures on medications.

3. **Preventing Drug Resistance:** Overuse of antimalarial drugs can lead to the development of drug-resistant strains of malaria, which is a significant public health concern. By maintaining high precision in malaria diagnostics, healthcare providers can help slow the development of resistance by ensuring that treatments are only given to those who truly need them.

4. **Enhancing Patient Trust and Health System Credibility**: False positives can erode trust in health systems. When patients are frequently misdiagnosed, it can lead to a lack of confidence in healthcare services, potentially deterring people from seeking care when they are ill. High precision in diagnostic tests helps maintain patient trust and the credibility of health services.

5. **Improving Disease Monitoring and Control:** Accurate diagnosis is critical for effective disease surveillance and control efforts. High precision in malaria tests allows for better tracking of disease patterns, understanding of transmission dynamics, and evaluation of control measures, which are essential for public health planning and intervention.

Precision in malaria diagnosis is not only about improving individual patient care but also about enhancing broader public health outcomes. Ensuring that malaria diagnostic tests are precise helps in effectively targeting treatment, preserving healthcare resources, preventing drug resistance, and maintaining trust in health services.

F1-Score:

The F1-score is a statistical measure used to evaluate the performance of a binary classification test or model, especially when the classes are imbalanced. It considers both the precision and the recall of the test to compute the score, providing a balance between these two metrics. Precision is the accuracy of positive predictions, and recall (also known as sensitivity) is the ability of the model to find all the relevant cases (true positives).

The F1-score is the harmonic mean of precision and recall, given by the formula:

2 (Precision Recall/Precision + Recall)

Why Use the F1-Score?

The F1-score is particularly useful in situations where:

- False negatives and false positives have different costs: It balances the importance of precision and recall. There is an uneven class distribution (large number of actual negatives): The F1-score can provide a more realistic measure of a model's performance than accuracy alone, which can be skewed by the larger class.

For example, imagine a model designed to identify a rare disease. Most people do not have the disease (negative class is much larger than the positive class). In such a scenario:

- High precision means that when the model predicts the disease, it is correct a high percentage of the time.
- High recall means that the model correctly identifies most people who actually have the disease.

A high F1-score indicates that the model has a robust balance between precision and recall, which is crucial in critical applications like medical diagnostics where missing a true case (high recall) and ensuring not to misdiagnose someone (high precision) are both vitally important.

Interpreting the Model Without Precision and F1 Score

When precision (which measures the proportion of positive identifications that were actually correct) and F1 score (the harmonic mean of precision and recall) are not provided, focusing on sensitivity, specificity, and AUC-ROC can still offer a comprehensive picture:

- **High Sensitivity and Specificity:** This indicates that the model is performing well in identifying both positive and negative cases accurately. However, in imbalanced datasets, achieving high values for both can be challenging. A model that maintains high sensitivity is often preferred in medical contexts to minimize the risk of overlooking true positive cases.

- **High Specificity with Imbalanced Data:** If a model has high specificity but the dataset is imbalanced, it suggests that the model is very effective at identifying the majority class (e.g., non-malaria cases in a predominantly healthy population) but does not indicate how well it identifies the minority class (e.g., actual malaria cases). In such scenarios, examining the sensitivity becomes crucial to ensure that the model is also effective in detecting the less common, but often more critical, positive cases.

Strategies for Model Interpretation

- **Threshold Adjustment:** By adjusting the classification threshold, you can explore the trade-off between sensitivity and specificity. This is crucial for tailoring the model's performance to prioritize either minimizing false negatives (important for critical conditions like malaria) or minimizing false positives (important in situations where the cost of a false alarm is high).

- **Use of Complementary Metrics:** Even if precision and F1 score are not available, other metrics like the Matthews Correlation Coefficient (MCC), AUC_ROC or the Balanced Accuracy can provide additional insights into the model’s performance, especially in the context of imbalanced datasets.

In summary, when analyzing imbalanced datasets with machine learning models, especially in critical applications like disease diagnosis, it’s important to look beyond single metrics. A comprehensive evaluation considering sensitivity, specificity, AUC-ROC, and the context-specific importance of false negatives and false positives is essential for a balanced interpretation of the model's performance.

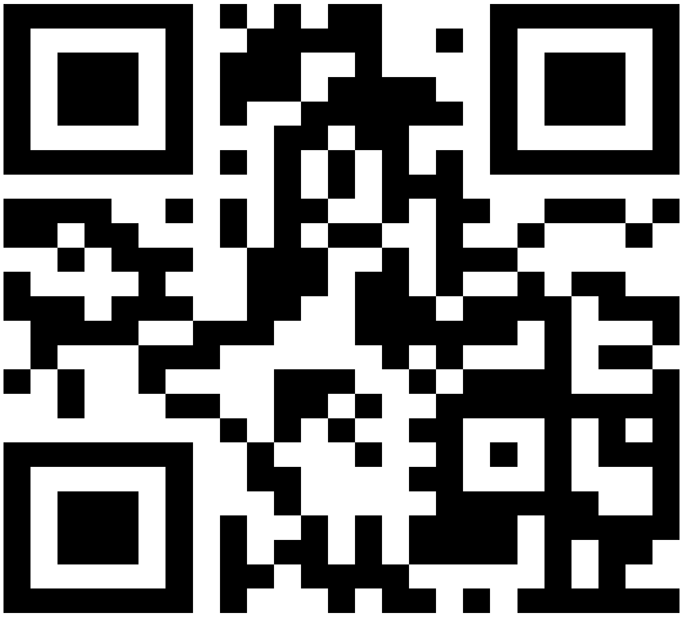
Practical/ Group Exercises

Note: The Assistant Instructors will move around to guide you in running these codes.

1. Each group should familiarize themselves with the Resampling Codebook (15mins)
2. Read and understand this instructional Notes (15 mins)
3. Compute, complete, and discuss different aspects of the resampling codes (15Mins)
4. Discuss how you can apply it to a real-life problem of your choice (15 mins)
5. Each group will make a 3-minute presentation of their findings (15 Mins)

Finally, each group will prepare a concise 3-minute presentation summarizing their key findings, insights from the exercises, and the discussion on applying what they've learned to a real-life problem. Presentations will be shared with all participants, allowing for knowledge exchange and feedback.

Kindly rate this course.



We appreciate your feedback.