Individual Assignment 3: Imbalanced Data

João Gonçalves
Faculdade de Engenharia da Universidade do Porto

1 INTRODUCTION

Machine learning with unbalanced data is one of the main challenges in many areas, especially when it comes to classification problems. In scenarios where one class is significantly more represented than the other, machine learning models tend to lean towards the majority class, which results in poor performance in identifying the minority class.

This work explores data restructuring techniques to tackle the problem of unbalanced data. The aim is to improve the model's ability to correctly identify the minority class, without compromising overall performance. To this end, several resampling approaches are analyzed, including oversampling, undersampling and hybrid combinations of both techniques.

2 INITIAL ANALYSIS

The data set [1] used in this study refers to a direct marketing campaign by a Portuguese bank, with the aim of predicting whether or not customers will sign up for a bank deposit.

2.1 Imbalanced Data

Data imbalance measures were used to assess the distribution of classes in the data set. The analysis revealed that the majority class ("0" class) has 7 times more examples than the minority class ("1" class). This imbalance is significant and can lead to unsatisfactory model performance in predicting the minority class. Furthermore, the minority class represents only 11.52% of the total number of examples in the data set, which confirms the need to apply balancing techniques.

Figure 1 shows the distribution of the classes in the data set, clearly demonstrating the imbalance between the classes. The majority class ("0" class) is mainly concentrated in the bottom left, while the minority class ("1" class) is represented by significantly fewer instances, spread throughout the graph. In addition, it is possible to observe some overlapping of instances of the two classes, particularly in the lower left region, which may indicate difficulties in clearly separating the classes.

2.2 Model Evaluation

For this work, we chose to use a set of specific metrics to evaluate the model's performance, given the challenge of unbalanced data. The metrics chosen are particularly suitable for scenarios in which the minority class is of primary interest, such as:

- **Sensitivity (Recall)** was chosen because it directly measures the model's ability to correctly identify examples of the minority class.
- **F-measure (F1-score)** was chosen to balance accuracy and sensitivity. In unbalanced problems, accuracy alone can be misleading, so using the F1-score combines accuracy and sensitivity.

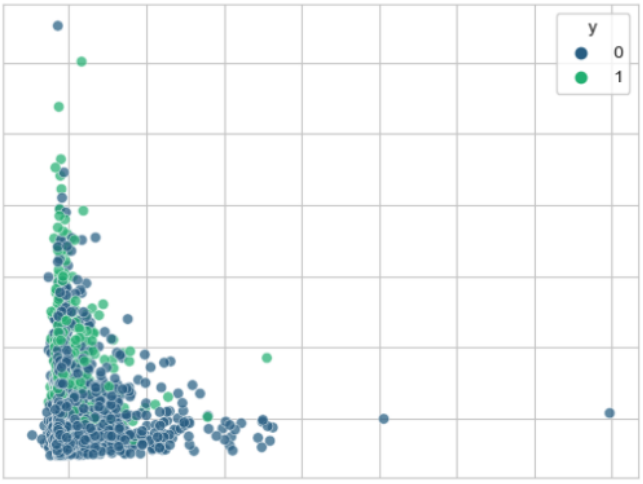

Figure 1: Class distribution visualization

- **G-mean** was chosen to evaluate the balanced performance between the two classes, considering both sensitivity (ability to detect the positive class) and specificity (ability to detect the negative class).
- **AUC-ROC** foi escolhida porque fornece uma visão global do desempenho do modelo em diferentes limiares de decisão.

Table 1: Results

Metric	Value
Sensitivity	0.42
F1-score	0.42
G-mean	0.62
AUC-ROC	0.67

Table 1 shows that the model has trouble correctly identifying instances of the minority class. The **sensitivity (Recall)** indicates that the model has problems recognizing class "1". The **F1-score** suggests that the model is still performing poorly, with a precarious balance between accuracy and sensitivity. The **G-mean** shows a slight improvement on the F1-score, but still reveals a significant imbalance in performance between the classes. Finally, the **AUC** greater than 0.5 indicates that the model can distinguish between classes better than chance, but there is still room for improvement in its ability to separate between classes.

3 DATA IMBALANCE TECHNIQUES

We will focus on the following techniques:

- **Over-Sampling** - Increase the number of examples of the minority class by creating synthetic data or replicating existing examples.
- **Under-Sampling** - Reduces the number of instances of the majority class by removing instances to balance the classes.
- **Hybrid** - It combines over sampling and under sampling to balance the classes while minimizing the risk of overfitting or losing relevant information.

3.1 Over-Sampling

We will explore the following **Over-Sampling** techniques:

(a) **SMOTE** (Synthetic Minority Over-sampling Technique)
(b) **ADASYN** (Adaptive Synthetic Sampling)
(c) **Random Over-Sampling**

(a) **(b)** **(c)**

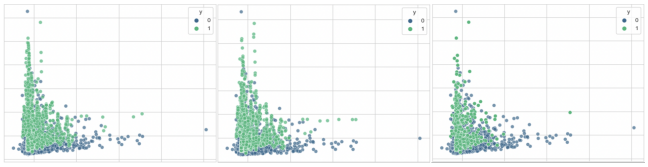

Figure 2: Visualization of class distribution with different oversampling techniques

Figure 2 shows the distribution of classes after the application of over sampling techniques. However, it can be seen that these techniques end up aggravating the problem of overlap between the classes, which can make it even more difficult to separate them and, consequently, the model's performance.

Table 2: Results with Over-Sampling

Metric	SMOTE	ADASYN	ROS
Sensitivity	0.40	0.37	0.42
F1-score	0.35	0.35	0.41
G-mean	0.59	0.58	0.62
AUC-ROC	0.64	0.64	0.67

The results obtained in Table 2 indicate that although the over sampling techniques balanced the classes, the model's performance suffered. SMOTE resulted in worse performance, suggesting that the base model was already benefiting from the imbalance. ADASYN showed worse results than SMOTE, and Random Over-Sampling adjusted the distribution of the classes, but did not bring significant improvements in the metrics, with sensitivity and F1-score still low.

In the case of this example, the over-sampling techniques (SMOTE, ADASYN and Random Over-Sampling) solved the problem of balancing the classes, but aggravated the problem of overlap between the classes. The greater overlap in the training data made it more difficult to learn the model, resulting in poorer performance.

3.2 Under-Sampling

We will explore the following **Under-Sampling** techniques:

(a) **Random Under-Sampling**
(b) **Tomek Links**
(c) **Edited Nearest Neighbors** (ENN)

(a) **(b)** **(c)**

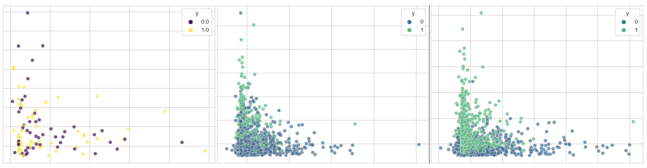

Figure 3: Visualization of class distribution with different undersampling techniques

In Figure 3, we can see three graphs showing the distribution of classes after applying the different undersampling techniques. Each graph illustrates the effects of the techniques mentioned above, showing how each one adjusts the distribution of the classes, reducing the majority class and, in some cases, improving the separation between the classes, albeit with possible loss of information.

Table 3: Results with Under-Sampling

Metric	RUS	TL	ENN
Sensitivity	0.67	0.47	0.66
F1-score	0.37	0.42	0.49
G-mean	0.71	0.65	0.76
AUC-ROC	0.71	0.69	0.76

The undersampling techniques helped to improve the overall performance of the model, as can be seen in table 3, mainly by eliminating problematic and noisy instances, which contributed to better learning of patterns and separation of classes.

The most effective undersampling technique was ENN, which brought substantial improvements in all metrics. This technique was the best at balancing the classes, increasing the value of all the metrics, resulting in a more effective model for discriminating between classes.

3.3 Hybrid

We will explore the following **Hybrid** techniques:

(a) **ENN + SMOTE**
(b) **ADASYN + ENN**

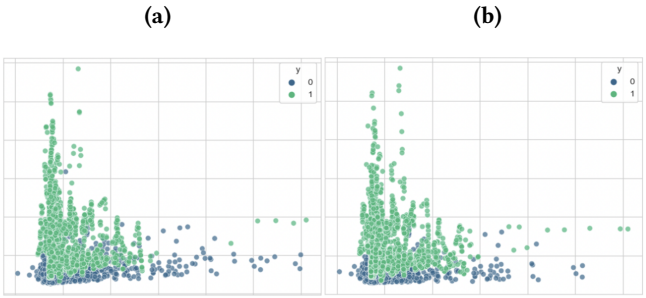

Figure 4: Visualization of class distribution with different hybrid techniques

Figure 4 shows the distribution of classes after applying the hybrid techniques. These techniques aim to balance the classes more effectively, creating a more even distribution of classes, while at the same time trying to reduce the overlap between them.

Table 4: Results with Hybrid

Metric	ENN + SMOTE	ADASYN + ENN
Sensitivity	0.62	0.60
F1-score	0.44	0.40
G-mean	0.73	0.70
AUC-ROC	0.73	0.71

Table 4 shows that the use of ENN + SMOTE and ADASYN + ENN brought significant improvements in the model's performance. Sensitivity increased substantially, indicating better identification of the minority class. However, the F1-score still did not show substantial improvements, indicating that the model continues to struggle to balance precision and recall. The use of hybrid resampling techniques, such as ENN + SMOTE and ADASYN + ENN, has proven to be an effective approach to dealing with the problem of unbalanced data and overlap between classes. These techniques combine the advantages of under-sampling and over-sampling, improving class separability and minority class identification.

4 CONCLUSION AND FUTURE WORK

In this work, we explored various resampling techniques, both over-sampling and under-sampling, to deal with the problem of unbalanced data. The over-sampling techniques, such as SMOTE and ADASYN, helped to increase the representativeness of the minority class, however, they showed a decrease in the model's performance, due to the excessive overlapping of the classes. The under-sampling techniques, such as Random Under-sampling, Tomek Links and ENN, were effective in improving the separability of the classes, especially ENN, which proved to be the best technique for this example, helping to remove noisy instances, promoting a better balance between the classes. Hybrid techniques combining over-sampling and under-sampling, such as ENN + SMOTE and ADASYN + ENN, resulted in improvements in overall performance, providing a balance between the creation of synthetic instances and data cleaning. However, they faced some difficulties in improving the f1-score. This is a topic to be explored in the future, which could bring a significant increase in model performance and improve the use of hybrid techniques.

REFERENCES

[1] P. Rita S. Moro and P. Cortezt. 2014. Bank Marketing. *UC Irvine* (2014). https://archive.ics.uci.edu/dataset/222/bank+marketing

MORE INFO

- Imbalance Learn
- Smote Variants
- Hybrid Technique
- Hybrid Technique
- Imbalance Classification