“Upsampling” Module Transcript

Chapter 1

Intro, Topics Covered, & Learning Outcomes

Hey, my name is Hayley, and I'm on the One AI team here at One Model. In previous modules, we discussed how classification models can encounter imbalanced datasets where classes are not equally represented. This imbalance can cause the model to favor the majority class, which is the more represented outcome, and hinder its ability to accurately predict the minority class, which is the less represented outcome.

In this module, we will continue the Advanced Configuration Modules series and talk about upsampling, which is a common technique to address this.

We will cover an overview and significance of upsampling, the different upsampling methods and ratios available in One AI, the default upsampling settings in One AI, and how to configure upsampling for One AI models.

After watching, you will understand how upsampling addresses imbalanced model datasets and improves model performance by ensuring that both majority and minority classes are adequately represented. You will differentiate between available upsampling methods to determine when to use each method based on the characteristics of your dataset. And you will configure upsampling settings in One AI, including selecting appropriate methods and ratios, and understand how to save and apply these settings to optimize your models.

Chapter 2

Overview & Significance of Upsampling

Section 2 - Overview and Significance

Upsampling, which is also known as oversampling, is a technique used in machine learning to address the issue of imbalanced datasets. An imbalanced dataset happens when the classes are not represented equally. Most of the popular classification One AI recipes result in imbalanced datasets.

For example, if you're predicting high performance and only 20% of your workforce are high performers and 80% are not, this would result in an imbalanced dataset where high performers make up the minority class and not high performers the majority class.

Upsampling involves increasing the number of instances in the minority class to balance the class distribution. This is important because if the dataset has a majority of instances from one class, the model may favor that class during training because there was just so many more examples of it, and that would result in it neglecting the minority class. This bias can lead to poor or inaccurate predictions for the minority class, because the model has not had enough examples to learn from. By creating more balanced datasets, the model can learn from both classes more effectively, leading to improved overall performance.

Classes do not need to be perfectly balanced in upsampling. The goal is to reduce the imbalance to a more manageable level, which can improve model performance. But achieving a perfect 50-50 balance is not always necessary or practical. Instead, the focus should be on reducing the imbalance sufficiently so that the model can learn effectively from both classes.

The degree of balance is determined by the upsampling ratio, which specifies how much to increase the minority class instances. This ratio guides how many synthetic or duplicated instances of the minority class will be added. Keep in mind that while upsampling ratio helps reduce class imbalance, creating too many synthetic instances can lead to overfitting. It's essential to strike a balance that improves model performance without overfitting.

Now that you understand the upsampling ratio, let's discuss the different types of upsampling methods. First, we have random oversampling. This method involves randomly duplicating instances from the minority class until the desired balance is achieved. And then there's also synthetic data generation. This method generates new synthetic instances of the minority class by interpolating between existing instances. Interpolating means creating synthetic instances that lie between existing ones. One AI only offers synthetic data generation methods as they are often more effective than random oversampling.

We will discuss the available options in detail in the next section.

Chapter 3

Methods & Ratios Available in One AI

Section 3 - Methods and Ratios Available in One AI

If you wish to perform upsampling configuration, you will be prompted to select an upsampling method and or ratio.

There are three options for method. First, we have ADASYN, which is short for adaptive synthetic sampling. This generates synthetic data points for the minority class focusing on harder to learn examples. It adapts the distribution of synthetic samples according to the learning difficulty of each instance, creating more synthetic samples in regions where the minority class is sparsely represented. Therefore, ADASYN is great for datasets where class imbalance is more complex and certain minority class regions are more challenging for the model to learn.

Next we have SMOTE, which is short for Synthetic Minority Oversampling Technique. This is a popular method that generates synthetic instances for the minority class by interpolating between existing minority instances. This method selects a random sample from the minority class and its nearest neighbors and creates new instances by blending features from the instance and its neighbors. SMOTE should be used when you have a generally imbalanced dataset where you want a balanced approach to generating synthetic instances evenly across the minority class space. It works best when your dataset primarily consists of continuous features.

And finally, we have SMOTENC, which is short for SMOTE for Nominal and Continuous Features. This is a variant of the SMOTE method I just described, and it's designed to handle datasets with both categorical and continuous features. It uses SMOTE for continuous features while handling categorical features separately to ensure meaningful interpolation, creating synthetic instances that respect the nature of both feature types. Therefore, it's best used for datasets with mixed feature types, especially where you need to account for the different behaviors of categorical and continuous features during synthetic data generation.

There are several different upsampling ratio options. Accepted values can be numerical or text string. Numerical values should be entered with a decimal value for the percentage of the minority class-majority class split desired. For example, entering 1.0 upsamples the minority class to achieve a 50-50 balance with the majority class. Entering 0.5 means that half as much upsampling would be performed for the minority class, resulting in a less balanced ratio.

The acceptable text strings are as follows: "auto", which upsamples all classes except the majority class; "minority", which upsamples only the minority class; "not minority", which upsamples all classes but the minority class; "not majority", which does the same thing as auto; and "all", which upsamples all classes.

You can select multiple methods and ratios. One AI will try each combination and select the one that results in the best model performance and fit.

Chapter 4

Default Settings in One AI

Section 4 - Default Settings in One AI

If you do not wish to perform upsampling configuration, One AI will select between none and SMOTE for the upsampling method by default. If upsampling improves performance and SMOTE is selected over none as the method, One AI will default to auto for the upsampling ratio. These are both discussed in the previous section in detail.

Chapter 5

Configuring Upsampling in One AI

Section 5 - Configuring Upsampling in One AI

To configure upsampling, click the 'Edit' button for the appropriate model and scroll down to "Upsampling" under "One AI Configuration". Toggle the override slider to 'On' and click the caret next to "Upsampling" to reveal your options.

If you would like upsampling disabled, slide the none toggle to 'On' and do not select any other methods or ratios. Different upsampling methods can be selected with the methods dropdown. As mentioned in the last section, you can select multiple methods, and One AI will try them all and select the one that performed the best. If you want One AI to try no upsampling in addition to other methods, slide the none toggle to 'On' and then select the other methods that you would like tried.

For ratios, you can either insert a numerical value or one of the accepted text strings. If you want to try multiple values, enter them in the designated fields separated by commas, like so.

Once you are happy with your upsampling configuration, click the 'Save' button at the bottom of the screen and then rerun your model.

Remember, you can always view the selected upsampling configuration in the Estimation Details section of the Results Summary Report for any completed model run.

Chapter 6

Conclusion & Thanks

Upsampling is a powerful technique to address the challenges of imbalanced datasets, ensuring your machine learning models can learn from both majority and minority classes. By understanding and applying different upsampling methods and configurations in One AI, you can significantly enhance model performance and accuracy. Remember to balance your approach to avoid overfitting and leverage One AI's capabilities to find the optimal settings for your models. Happy modeling!