geoconvention
Calgary • Canada • May 15-17 2023

Rock Facies Imbalanced Classification with Over-Sampling and Under-Sampling Techniques

Ryan A. Mardani, Mohammad Mardani, Daniel Trad
Data Energy, University of Zanjan, University of Calgary

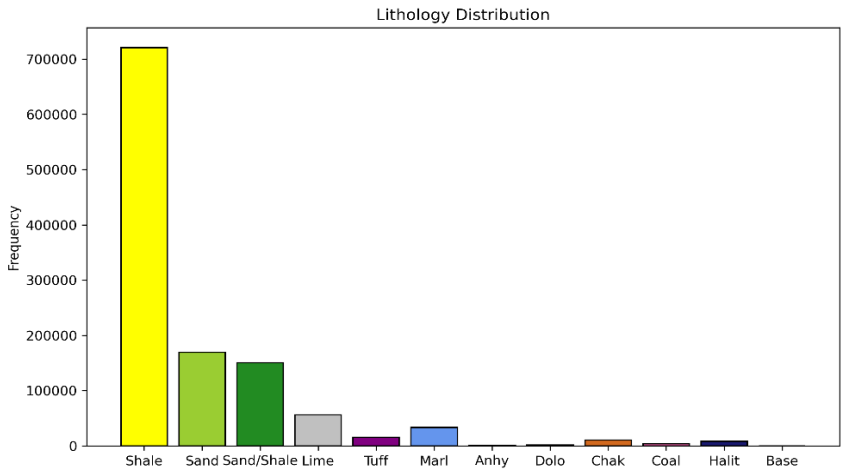
Summary

In classification problems, if the dataset is skewed, most machine learning algorithms produce poor prediction results for minor classes. In this study, we used different resampling techniques to balance the dataset that includes well logs and rock facies. XGBoost model is employed for this multi-class classification problem. We found that oversampling can improve prediction results in minor classes. Overall, Synthetic Minority Oversampling Technique (SMOTE) is a better candidate for oversampling, though for some classes Adaptive Synthetic Sampling (ADASYN) could compete with the SMOTE performance.

Introduction

Sometimes real-world datasets are imbalanced. If these classes have special importance, we need to approach such skewed data with imbalanced classification considerations. The dataset that we used for this research is open data published as FORCE 2020 for lithology prediction from well logs in Norwegian offshore. It contains more than 100 wells with common wireline logs and interpreted lithology as facies classes.

Figure 1 shows the histogram of target classes on the left as well as numerical information in the table on the right. While shale is the major class in this dataset, the others can be categorized as minority classes with different levels of severity. Sand, Sand/Shale, Lime, Tuff, and Marl are moderately imbalanced. The rest of them are extremely imbalanced.



Class	Count	Percent
Shale	720803	61.58
Sand	168937	14.43
Sand/Shale	150455	12.85
Lime	56320	4.81
Tuff	15245	1.30
Marl	33329	2.85
Anhy	1085	0.09
Dolo	1688	0.14
Chak	10513	0.90
Coal	3820	0.33
Halit	8213	0.70
Base	103	0.01
Sum	1170511	1

FIG1: Lithology distribution shows shale is the major class, and the other classes are either moderately or extremely imbalanced.

Methodology

There are two resampling methods available to handle the imbalanced data:

Under-Sampling: this technique helps balance data by eliminating some samples from major classes. This method should be used with caution as reducing skewness can lead to information loss during the training process. The methods that we used in this study are: 1- Random Under Sampling, 2- Edited Nearest Neighbor Under Sampling (ENN), 3- Neighborhood Cleaning Under Sampling (NHC).

Over-Sampling: unlike the under-sampling method that focuses on decreasing major classes, over-sampling increases minority class samples. In this work, we will use different methods of over-sampling to handle imbalanced data. These methods are: 1- Random Over Sampling, 2- Synthetic Minority Oversampling Technique (SMOTE), 3- Adaptive Synthetic Sampling (ADASYN).

We need to emphasize that although the over-sampling technique balances the data but does not add new/additional information to the dataset. It increases the chance that minority groups can be seen by the model more than whenever it is imbalanced. To study details of these algorithms, you may refer to the references.

Results

We employed the XGBoost algorithm with optimized hyper-parameters to implement multi-class classification. To evaluate the model performance, there are several evaluation metrics. Here we elaborate on the f1-score which is the harmonic mean of precision and recall.

Let's look at Figure 2. For the majority class, we have a drop in model performances when balancing happens. F1 is improved for marl and dolostone when we balance the dataset but still noticeably low. Except for coal, the other minor classes' F1 scores are improved if we used oversampling techniques. Anhydrite, as expected from low recall, didn't receive an acceptable f1-score. Overall, SMOTE performed well enough to be the candidate for all minor class's resampling techniques.

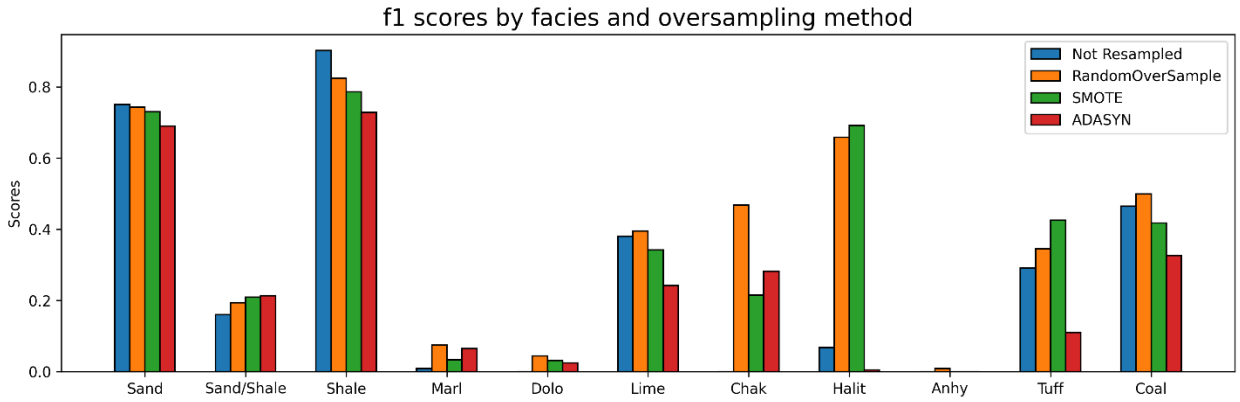


FIG 2: F1 scores for different facies predictions using over-sampling methods.

In Figure 3 we see that Random Over Sampling shows strong fluctuation through the different facies. F1-score strongly picked for Halite, Chalk, and Anhydrite. The other methods of under-sampling show similar performance to over-sampling methods.

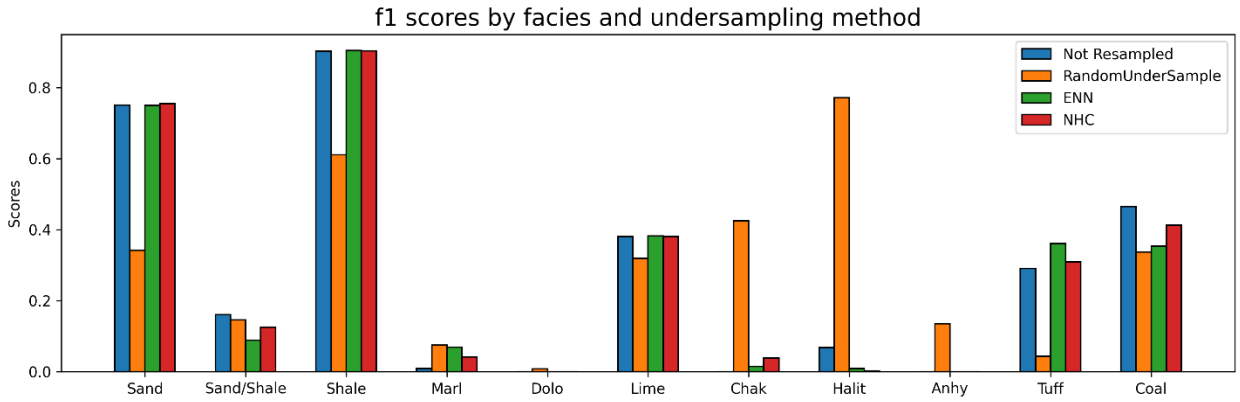


FIG 3: F1 scores for different facies predictions using under-sampling methods.

In the last step of model performance and prediction results based on balancing methods, let's plot the predicted facies along the true classes in the one well (Figure 4). In the first three tracks of the logs, we plotted features like depth, formation tops, and composite logs of GR and SP. Other tracks are the predicted facies. Track four is the XGBoost's prediction result on the original dataset (not balanced). The other tracks are the prediction result on the dataset balanced by Random Under Sampling, ENN, NHC, Random Over Sampling, SOMTE, and ADASYN, respectively. True lithology is plotted at the last facies track.

The main key points in this plot are:

- Major classes are predicted perfectly by both resampling methods except random under-sampling approach.
- Random Under Sampling shows the weakest result for both minor and major classes
- SMOTE and ADASYN predicted shales at shallow intervals as limestone
- Tuff and coal could be detected successfully by SMOTE and ADASYN resampling methods while under-sampling methods fail to predict these classes.

As you may notice, some minor classes could be predicted better by oversampling techniques, especially SMOTE did an effective job. Some other classes, on the other hand, have been over-predicted like limestone. Here the concept of business problem importance comes to play. Suppose Tuff is a very important facies and when you drill a well, you don't want to miss that interval. In such cases, we prescribe running oversampling techniques on the imbalanced dataset before feeding it to intelligent systems. Or suppose marl is a dangerous zone for drilling and there is a chance of drilling bit stuck in such intervals. In such case, we don't want to miss any marl interval (high recall) though some other rock could be predicted as marl and we need to screen the model result manually. This will cost the operator to supervise the drilling operation to manually help the model by live lithology interpretation during operation.

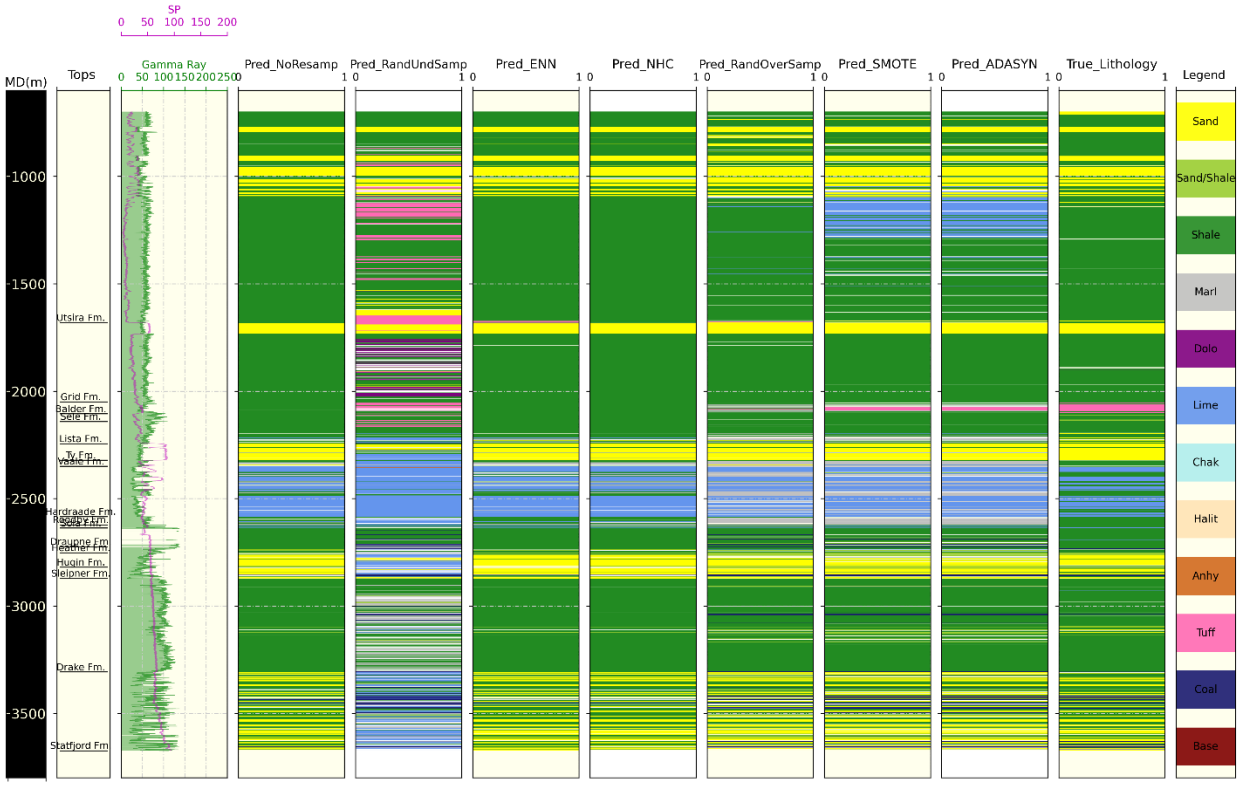


FIG 4: plot of some well logs (predictors) and interpreted facies using various resampling methods along with actual lithology (well: 26/4-1).

Conclusions

This was a multiclass classification problem in that we employed the XGBoost algorithm to predict lithology from well logs. The objective of this research was to examine how resampling methods can affect the prediction result on the imbalanced dataset. We observed that oversampling techniques could help the classifier to predict minority classes with higher accuracy than when it is imbalanced. Overall, SMOTE is the better candidate for oversampling, though for some classes ADASYN could beat the SMOTE performance.

References

Geron, A., 2019, Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems (2nd ed.). O'Reilly.

He, H. and Ma, Y., 2013, Imbalanced Learning: Foundations, Algorithms, and Applications. Wiley-IEEE Press, Hoboken, NJ, USA.

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

https://imbalanced-learn.org/stable/auto_examples/over-sampling/plot_illustration_generation_sample.html