The effect of the imbalanced training dataset on the quality of classification of lithotypes via whole core photos

Daria Makienko
Schlumberger Moscow Research
Moscow, Russia
dmakienko@slb.com

Ilya Seleznev
Schlumberger Moscow Research
Moscow, Russia
iseleznev@slb.com

Ilia Safonov
Schlumberger Moscow Research
Moscow, Russia
isafonov@slb.com

***Abstract*—Nowadays machine learning methods play an important role in many industries. However, the effectiveness of the predictive models depends on the quality of data sets used to train the model. In practice, the imbalanced datasets are quite common. For example, in the problems of lithotypes classification via whole core photos, some lithotypes often predominate the training dataset while some of the other lithotypes can be underrepresented. The significant imbalance in the dataset can affect the quality of the classification. In this case it is difficult to obtain good generalization for poorly represented classes. First, some characteristics of a given minor lithotype may be absent. Second, some features of a minor class can be ignored due to imbalance. In this paper, we analyze the oversampling of a minor class as one of the possible options to obtain the balanced dataset within the framework of the problem of speeding-up the geological core description. We considered examples with different dataset sizes and imbalance characteristics to study the effect of applying the oversampling approach on the quality of predictive models.**

Keywords—imbalanced dataset, oversampling, classification of lithotypes, geological core description

I. Introduction

The lithological description of whole core specimens is a time-consuming process. Using whole core photos to classify rocks and mark depth intervals corresponding to these rock classes can significantly reduce the time required for such description. Modern methods for automating the description of rocks by core photographs are based on machine learning. The most informative features for machine learning are the color characteristics of core image fragments [1,2]. In this paper we build predictive machine learning-based models using color characteristics of whole core photos. We consider an important factor that largely determines the quality of rock classification, namely, the influence of data imbalance, on which the predictive model is trained, and one of the approaches to compensate the imbalance.

The aim of our study is to determine the parameters of data samples that can significantly affect the quality of predictive models, as well as to assess the degree of such influence. Analyzing characteristics of sample imbalances in a wide range of values, we want to understand the limitations of the dataset parameters at which such imbalance can be corrected to improve the quality of predictive models.

II. Overview of techniques for processing imbalanced data

Real datasets often lack any data due to the difficulty of obtaining them. Different methods are used to compensate for missing data depending on the data type and the type of task [3-11]. We consider the case of imbalance of classes in the classification problem, when the data are presented in the form of numerical features.

In the classification problem, it is preferable that the training examples are evenly distributed among the classes. Some classifiers take into account the errors for different classes with same weights and in case of imbalance they become more focused on the overrepresented classes. The reason for such behavior of classifiers is that identifying the characteristics of the majority class contributes stronger to the target value (quality functional or error function) than identifying the characteristics of the minority class. However, the imbalanced classification data sets are often observed in applied problems [4-11]. Data sets for the lithological description of core are no exception. The imbalance of classes is associated with different rock occurrence. The following methods can be used to train a model on imbalanced data [9-11]:

1. Balancing, that is changing the ratio of classes in the sample by increasing the number of instances of the minority class (oversampling) or reducing the number of instances of the majority class (undersampling).

2. Making adjustments to the learning algorithm. For example, setting different penalties for classes in the support vector machine, changing the probability threshold for classifying an example as a class in trees.

3. Establishing different error rates for classes. The cost of errors can be taken into account both when changing the ratio of classes in the sample, and when making adjustments to the learning algorithm.

4. The use of boosting. Several classifiers that correct each other's errors can improve the quality of model predictions based on examples of a minority class.

For lithological description based on full-size core photographs, we investigate the oversampling. This approach balances samples by increasing the number of examples of the minority class. Some of the existing oversampling techniques are as follows:

1. Random oversampling: Copies of randomly selected elements of the minority class are created until the required ratio is reached.

2. SMOTE (Synthetic Minority Oversampling Technique) [12]: New examples are generated by interpolating the examples of the minority class — some *i*-th example and one of its k-nearest neighbors. There are several options for choosing the *i*-th example. One can make a random selection (Regular SMOTE), select an example depending on the classes to which the surrounding examples (Borderline SMOTE) belong, depending on the constructed support vectors or on the constructed clusters.

3. ADASYN (Adaptive Synthetic) [13]: It works similarly to the SMOTE method but selects the *i*-th example

of a minority class depending on the coefficient r_i, which shows the proportion of examples of other classes around the *i*-th example. The greater the coefficient r_i, the more examples are generated in the vicinity of the *i*-th example.

III. Setting up an experiment

There are a lot of lithotypes. In general, lithology classification is a multiclass problem. For our study we simplify the problem at this stage, considering a binary solver that is the one-vs-rest classifier. To study the influence of the imbalance on the quality of classification the depth intervals were selected in the manner to obtain balanced and variously represented target lithotype data, reflecting the typical color features of this lithotype as well. We denote the data obtained after processing all of the images the initial sample. To study the effect of imbalance on the quality of classification, we form different size subsamples of the initial sample, which act as minority class with different imbalance. We train predictive models on such subsets and try to compensate the imbalance.

We tested 4 initial data sets with different sizes of minority and majority classes: 2330:4075, 1165:2038, 583:1019, 292:510, where the first value is the number of examples of the minority, the second is the number of examples of the majority class (other lithotypes). To create subsamples with different class ratios and study the influence of the initial ratios on the further complement of the sample, the minority class is reduced by randomly choosing a subset of it of size *m*. The value of *m* corresponds to some new proportion *p* relative to the size of the majority class. Such subsamples are denoted as *p(m)*. To reduce the influence of a random factor on the classification results, for each subsample *p(m)*, examples are selected 10 times and the results are averaged.

While training sets have different levels of imbalance, the test set is not changed and has the class ratio inherited from the initial sample. To assess the quality of models, a 5 folds cross-validation is used [14]. The training data sets consist of 4/5 of the initial data sets and have sizes of minority and majority classes: 1864:3260, 932:1630, 466:815, 234:408. Testing is performed 5 times on each of the folds and the results are averaged.

To balance the training set, we use SMOTE with random selection of examples. After balancing and training, the classification accuracy is estimated. We apply the linear classification algorithm (logistic regression) and the tree-based algorithms (gradient boosting and random forest) to train the classifier. We employ F1 score to evaluate the quality of models. F1 is the harmonic mean of *Precision* and *Recall*:

$$Recall = \frac{TP}{TP + FN}; \; Precision = \frac{TP}{TP + FP};$$

$$F1 = \frac{2TP}{2TP + FP + FN},$$

where TP (True Positive), TN (True Negative) are the number of correctly predicted objects of the positive and negative classes correspondingly; FN (False Negative), FP (False Positive) are the number of objects incorrectly assigned to negative and positive classes correspondingly. The positive class is the minority class corresponding to the target rock, and the negative class is the majority class, corresponding to other rock types.

IV. Results

The quality of models for determining the "silty-clay rock" lithotype, trained on imbalanced subsamples without using oversampling, increases with the growth of the proportion *p* and their number *m* of examples of the minority class (Fig. 1). At the same time, the quality reaches an acceptable level only in subsamples where the level of imbalance is very small. Therefore, it becomes necessary to correct the imbalance, as well as to study the influence of parameters *p* and *m* on the operation of the classifier. After applying oversampling to balance classes, the quality of the models improves.

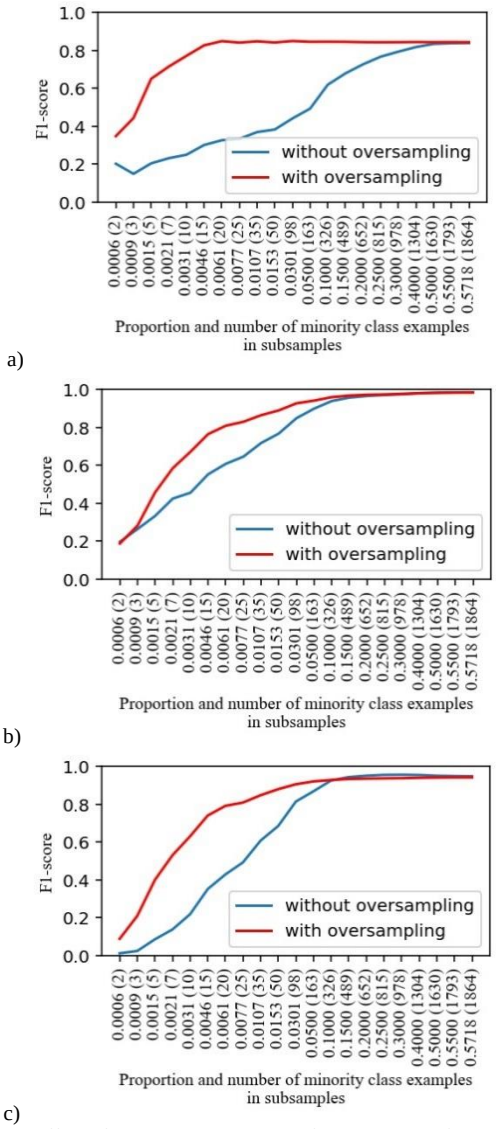

Fig. 1. The effect of oversampling on classification quality for training set 1864:3260: (a) logistic regression, (b) gradient boosting, (c) random forest

Fig. 2 shows plots of the dependence of the F1 score on the proportion of minority class examples after oversampling for the two training sets. The plots for training sets 932:1630 and 466:815 are not shown, because they look similar. One can see the proportion *p* and the number *m* in the legend of plots.

By comparing of Fig. 2(a) and Fig. 2(b), for the classification of the target lithotype, we conclude the following:

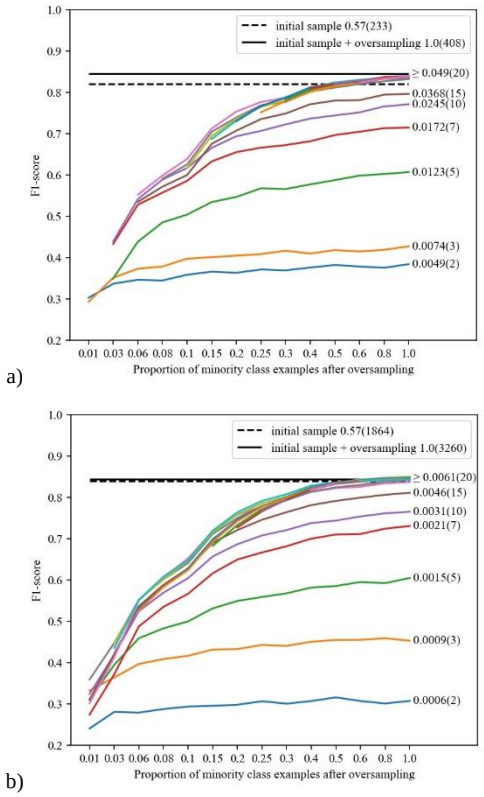

Fig. 2. The effect of proportion of minority class examples after oversampling on classification quality for training sets (a) 233:408, (b) 1864:3260.

1. The quality of the model depends on the number of examples *m* in the minority class before oversampling. The dependence on *p* is not significant.

2. The quality of the model grows with the increasing number of examples representing the minority class before the oversampling.

3. The quality of the model increases when the fraction of minority class examples increases due to the use of oversampling.

4. There is the threshold for the number of examples *m*, at which the quality of the initial sample can be reached if oversampling is applied.

Fig. 3 shows plots of the F1 score dependence on the proportion of minority class samples after oversampling for random examples extracts from the initial sample at $p = 0.002$ ($m = 5$), $p = 0.005$ ($m = 15$), and $p = 0.015$ ($m = 50$). With a small number of examples of the minority class, the random factor in choosing these examples has a significant impact on the accuracy of classification. If the initial training set 1864:3260 is trimmed to an imbalance $p = 0.002$ ($m = 5$), then when the minority class is oversampled to balance with the majority class, the average F1 score is 0.62, but the deviation from the average reaches 0.09. As the number of examples increases, the average value of the F1 score increases, and its dispersion decreases. This is not true for all *m*, but in general this trend persists. For a subsample with $p = 0.015$ ($m = 50$), the average value of the F1 score after balancing is 0.85 and the deviation is 0.01.

To assess the similarity of the subsamples balanced by the SMOTE method with the initial sample, we used histograms and cross-plots constructed for the most significant features.

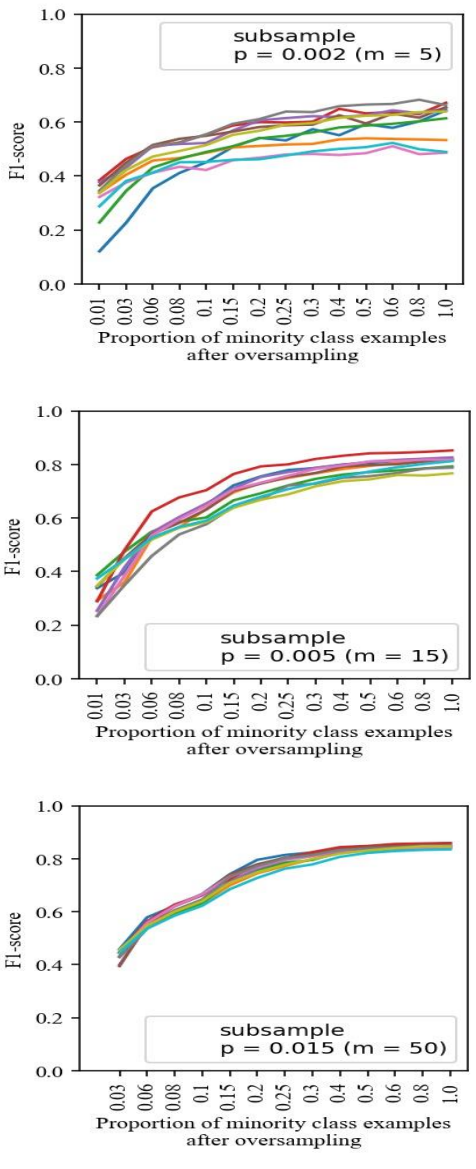

Fig. 3. Graphs showing the scatter of the F1 score for 10 random versions of subsamples from the training set 1864:3260 with parameters (a) *p = 0.002 (m = 5)*, (b) *p = 0.005 (m = 15)*, (c) *p = 0.015 (m = 50)*.

For 10 versions of subsamples with the parameter $m = 50$, balanced to an equal ratio of classes, the distribution of features on histograms and cross-plots is visually similar to the distribution of the initial sample (Fig. 4 (c), (d)). For subsamples with the parameter $m = 5$, balanced to an equal ratio of classes, the feature distributions may be close to the distribution of the initial sample, but in most cases, they have significant differences (Fig. 4 (a), (b)).

V. Imbalance in the Multiclass Classification Problem

We consider a multiclass lithology classification and try to verify if the approach we applied for the binary classification can also improve predictive models for the multiclass problem. As well as for binary classification, to study the effect of imbalance, we change the size of the target class until the equality with the largest class is achieved. We use random forest classifier and consider nine class model. One of these nine classes - carbonate sandstone, is underrepresented in our dataset and we consider it as a target minority class.

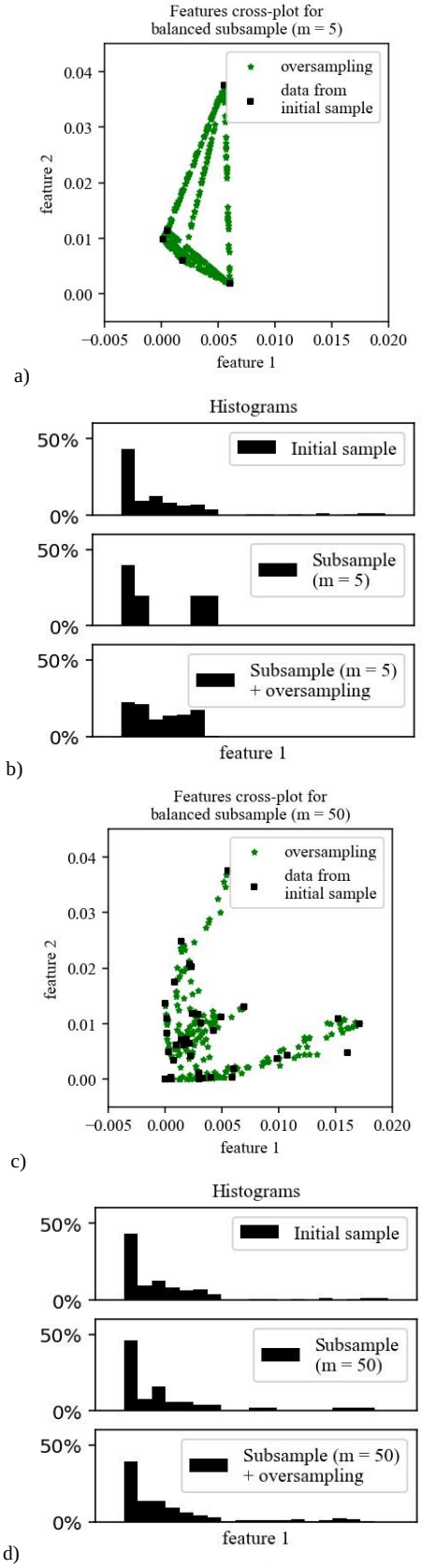

Fig. 4. Scattering diagrams and histograms for (a) - (b) balanced subsample at m = 5, (c) - (d) balanced subsample at m = 50. Subsamples are balanced to equal class sizes.

To establish different levels of imbalance, we randomly select 10, 30, and 100 examples from 1249 labeled ones. Similar to our previous experiments, an increase in the number of examples selected from the initial sample and increase of the minority class fraction after oversampling led to growing the F1 score (Fig. 5). The quality of the model is evaluated by cross-validation.

Fig. 6 (a, b) illustrates the difference of prediction confidence of the two predictive models – before and after oversampling. The first 3 columns show day light (DL), ultraviolet (UV), and gamma corrected ultraviolet (UV gamma corrected) photographs. Fig. 6 (a) contains depth intervals which were involved in training, and Fig. 6 (b) contains depth intervals which were not be involved in the training process.

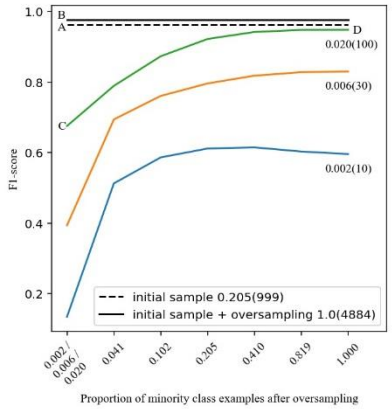

Fig. 5. Assessment of the classification of the target lithotype relative to the rest.

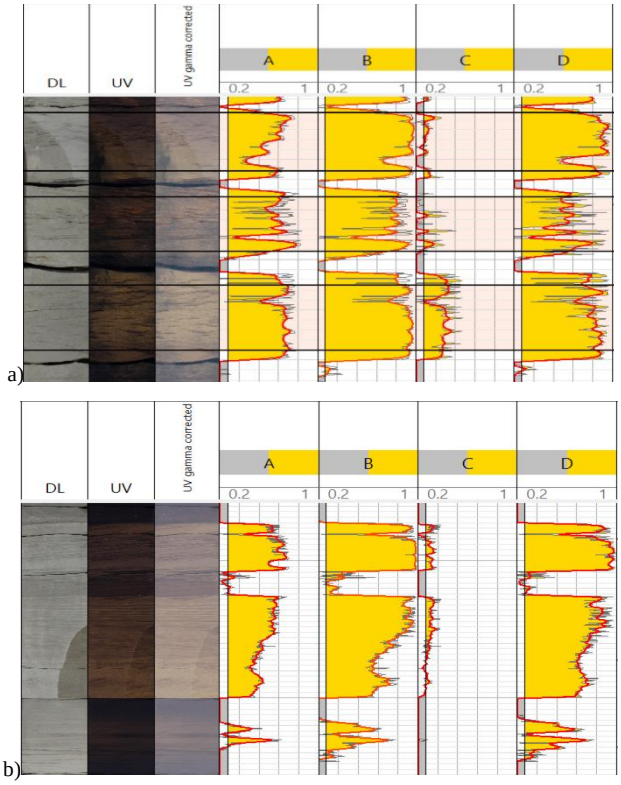

Fig. 6. The predicted probabilities of the presence of the target lithotype in the corresponding core sections (a) areas involved in training, (b) areas that were not involved.

The cases A, B, C, and D correspond to the following:

A: The predictive model is trained on the initial sample, containing 1249 target examples;

B: The predictive model is trained on the initial sample oversampled to equality with the majority class and contained 4884 target examples;

C: The predictive model is trained on a sample of 100 randomly selected target examples;

D: The predictive model is trained on a sample of 100 randomly selected target examples, oversampled to equality with the majority class.

Cases A, B, C, and D contain in red the gaussian smoothed curves of probability (confidence level) for the core specimens to belong to the target class.

Fig. 5 as well contains labels A, B, C, and D which are related with corresponding cases

Thus, it is seen that applying the oversampling technic can improve the quality of the predictive model for the multiclass problem, both in terms of F1 score and in terms of confidence of the prediction.

VI. Conclusion

The paper considers the influence of data imbalance on the quality of lithotypes classification by the whole core photographs. It is shown that the quality of predictive models trained on imbalanced data may depend on the degree of imbalance and for some samples the imbalance can dramatically affect the quality of classification.

The level of imbalance at which it is possible to obtain a predictive model that is close in quality to the model trained on a balanced sample is not constant and depends on the size of the data sample, as well as on the quality of the data sample. Quality here refers to how fully the sample reflect the characteristics of the target lithotype.

Applying the oversampling technic of data balancing by SMOTE method can increase the quality of the lithology classification for binary problem (detection of silty-clay rocks), and for the multiclass problem.

The quality of predictive models, close to the quality of the model built on the entire balanced data set, was achieved for those imbalanced samples which let us restore the distribution of the entire data set with the least influence of the random factor.

There is a minimum acceptable number of specimens, weakly depending on the size of the entire sample, at which we can claim the reproducible quality of model training (with an acceptable variance of the quality criterion). As the number of specimens available for training decreases, the variance of the model quality criterion increases.

References

[1] E.E. Baraboshkin, L.S. Ismailova, D.M. Orlov, E.A. Zhukovskaya, G.A. Kalmykov, O.V. Khotylev, E.Y. Baraboshkin and D.A. Koroteev, “Deep convolutions for in-depth automated rock typing,” Computers & Geosciences, vol. 135, 104330, 2020.

[2] A. Thomas, M. Rider, A. Curtis and A. MacArthur, “Automated lithology extraction from core photographs,” First Break, vol. 29, no. 6, pp. 103-109, 2011.

[3] G.R. Vorobeva, “Approach to the recovery of geomagnetic data by comparing daily fragments of a time series with equal geomagnetic activity,” Computer Optics, vol. 43, no. 6, pp. 1053-1063, 2019. DOI: 10.18287/2412-6179-2019-43-6-1053-1063.

[4] V.I. Shakhuro and A.S. Konushin, “Image synthesis with neural networks for traffic sign classification” Computer Optics, vol. 42, no. 1, pp. 105-112, 2018. DOI: 10.18287/2412-6179-2018-42-1-105-112.

[5] M.F. Sohan, M.I. Jabiullah, S.S.M.M. Rahman and S.M.H. Mahmud, “Assessing the Effect of Imbalanced Learning on Cross-project Software Defect Prediction,” 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT, 8944622, pp. 1-6, 2019.

[6] S. Huda, K. Liu, M. Abdelrazek, A. Ibrahim, S. Alyahya, H. Al-Dossari and S. Ahmad, “An ensemble oversampling model for class imbalance problem in software defect prediction,” IEEE Access, vol. 6, pp. 24184-24195, 2018.

[7] R. Shimizu, K. Asako, H. Ojima, S. Morinaga, M. Hamada and T. Kuroda, “Balanced mini-batch training for imbalanced image data classification with neural network,” 1st IEEE International Conference on Artificial Intelligence for Industries, vol. AI4I, 8665709, pp. 27-30, 2018.

[8] N.B. Paklin, S.V. Ulanov and S.V. Tsar'kov, “The construction of classifiers on imbalanced samples by the example of credit scoring,” Artificial Intelligence, no. 3, pp. 528-534, 2010.

[9] Y. Sun, A.K. Wong and M.S. Kamel, “Classification of imbalanced data: A review,” International Journal of Pattern Recognition and Artificial Intelligence, vol. 23, no. 04, pp. 687-719, 2009.

[10] C. Seiffert, T.M. Khoshgoftaar, J. Van Hulse and A. Napolitano, “Building Useful Models from Imbalanced Data with Sampling and Boosting,” FLAIRS conference, pp. 306-311, 2008.

[11] G.M. Weiss, K. McCarthy and B. Zabar, “Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?” International Conference on Data Mining, vol. 7, pp. 35-41, 2007.

[12] N.V. Chawla, K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” Journal of artificial intelligence research, vol. 16, pp. 321-357, 2002.

[13] H. He, Y. Bai, E.A. Garcia and S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322-1328, 2008.

[14] S. Raschka, “Python machine learning,” Packt Publishing Ltd, 2015.