An Effective Methodology for Imbalanced Data Handling in Predictive Maintenance for Offset Printing

Alexandros S. Kalafatelis, Nikolaos Nomikos, Angelos Angelopoulos, Chris Trochoutsos, and Panagiotis Trakadas

Abstract The printing industry is one of the largest manufacturing industries in the world, being characterized by high production volumes, where continuous maintenance of machine performance is key. Predictive Maintenance (PdM) enables the use of a maintenance policy based on novel Machine Learning (ML) algorithms, in or-der to provide valuable insights for both diagnostics and prognostics. However, real-world data used for PdM model training are characterized by great class imbalances, as failure events have a significant lower rate of happening compared to the normal no failure operations. Furthermore, ML models that are subjected to imbalanced datasets, are prone to be highly biased while having misleading accuracy scores. This can prohibit systems to accurately predict machine failure, leading to excessive costs while affecting the safety of the workers. This work proposes a data sampling methodology for predictive maintenance algorithms used mainly in Offset Printing environments, aiming to improve model performance. Based on a historical dataset extracted by an Offset Printing manufacturer, a methodology consisting of multiple classification algorithms utilizing different sampling techniques (SMOTE,

A. S. Kalafatelis (✉) · N. Nomikos · A. Angelopoulos · P. Trakadas
Department of Port Management & Shipping, National and Kapodistrian University of Athens, 34400 Psachna, Evia, Greece
e-mail: akalafat@core.uoa.gr

N. Nomikos
e-mail: nomikosn@pms.uoa.gr

A. Angelopoulos
e-mail: a.angelopoulos@uoa.gr

P. Trakadas
e-mail: ptrakadas@pms.uoa.gr

C. Trochoutsos
Pressious Arvanitidis, 15232 Athens, Greece
e-mail: chtrox@pressious.com

G. L. Conte and O. Sename (eds.), *Proceedings of the 11th International Conference on Mechatronics and Control Engineering*, Lecture Notes in Mechanical Engineering,
https://doi.org/10.1007/978-981-99-6523-6_7

ADASYN, and RUS), was trained and evaluated using cross-validation. The evaluation outcomes demonstrated the ability of the proposed methodology to effectively handle data imbalances while significantly enhancing model performance, outperforming other state-of-the-art techniques.

Keywords Predictive maintenance · Machine learning · Industry 4.0 · Offset printing

1 Introduction

Predictive maintenance (PdM) has been gaining prominence recently in multidisciplinary sectors, enabling the use of a maintenance policy based on novel Machine Learning (ML) algorithms. In essence, PdM works by estimating and foreseeing failures in deteriorating systems around manufacturing environments, in order to optimize maintenance efforts [1].

The printing industry is one of the largest manufacturing industries in the world, having high production volumes, where continuous maintenance of machine performance is key. Possible breakdown events will automatically result in production stop, disturbing thus not only the production process, but also burdening financially the manufacturers. Offset Printing enables the production of large quantities, as the variable production costs are deemed small compared to the setup costs, thus having a greater risk in case of machine breakdown. Possible failures found in Offset Printing, include but are not limited to: (i) defective offset rubbers, (ii) wear of ink rollers, (iii) incorrect bending or damaged printing plates, (iv) insufficient pressures on the printing machines, (v) non-conformity issues in the sheet delivery unit, and (vi) random failures, which are found at every manufacturing environment [2].

According to Haarman et al. [3], maintenance procedures are shown to represent a total of 15–60% out of the total costs of operating of all manufacturing, thus showcasing the importance of a PdM solution. In detail, a PdM solution aims to not only prevent possible failures but to also optimize operations, affecting thus different aspects of manufacturing, including safety, product quality, reliability, and minimization of operational costs.

PdM data provide valuable insights for both diagnostics and prognostics information, enabling maintenance work to become proactive. ML assumes that data used for training and testing purposes are under the same feature space, having similar distribution and comparable proportion of training instances belonging to each class. However, this is not always the case in real world applications, where ML have to face complex challenges in which these assumptions are not always satisfied [4].

Furthermore, ML models that are subjected to imbalanced datasets, are prone to be highly biased while having misleading accuracy scores. This phenomenon can be attributed due to the lack of information coming from the minority class of a given dataset and to ML models in general, as they tend to classify every test sample into the majority class, in order to improve the accuracy metric [5, 6].

This phenomenon is predominated in cases where anomaly detection is of prime importance, such as in PdM, prohibiting the systems to accurately predict machine failure, leading not only to excessive costs for the manufacturers, but also possibly affecting the safety of the workers.

To mitigate this issue, sampling techniques such as under-sampling and oversampling are used either to create more instances of the minority class to increase its population or to minimize the data instances found on the majority class.

In this paper, the occurrence of machine failure is determined on a predictive maintenance dataset, implementing SMOTE, ADASYN and RUS methods to generate balanced datasets of machine failure instanced found in Offset Printing. The efficiency of the proposed oversampling and undersampling methodologies are analyzed with the help of various machine learning classifiers, with the aim to improve predictive maintenance accuracy scores.

This paper is structured as follows. In Sect. 2, we suggest the details of the utilized dataset and of the proposed methodology of handling imbalanced datasets, alongside with the classification algorithms. In Sect. 3, the experimental results used to assess the performance of the different sampling methods and of the classification models, are presented. Finally, in Sect. 4 the results are summarized and discussed.

2 Materials and Methods

2.1 Dataset Description

The original dataset consisted of features and labels based on historical measurements collected during a 4-month trial period (03/07/2022–31/10/2022) from Pressious Arvanitidis, an Offset Printing manufacturer based in Greece. Each of the collected parameters and features, follows the process of a particular printing order (i.e., from the sales department to the quality assessment department). The order and factory related characteristics used in this paper are presented in Table 1.

Table 2 summarized the descriptive statistics of the independent and dependent variables of the complete dataset.

2.2 Data Processing Methodology

Due to the high-class imbalance in the initial raw dataset regarding the failure events (containing only 145 events of some type of machine failure out of the 4205 total printing runs), data preprocessing was performed to facilitate the training and testing processes of the ML models with high-quality data.

Particularly, to avoid a scenario where a particular variance dominates the objective function of the learning algorithms (making it unable to learn from other features

Table 1 Parameters used for the training and testing procedures for the ML models

Parameter	Description
Unique order ID	Unique identifier varying from 1 to 10000
Quality	The requested end product paper type in a particular order. It is a categorical variable that takes values 'Velvet', 'Uncoated' and 'Illustration'
Quantity	The number of printing pieces requested in a particular order
Type	The outcome type of a particular order, taking values of 'Book', 'Poster' and 'Journal'
Color	The specific color requirements of an order. Categorical variable taking values between 'typical' 4-color printing, '4 + 1' color printing or 'grayscale' printing
Machine	The specific ID of the machine that a particular order was forwarded for printing, ranging from 1 to 5
Humidity	Water vapor relative to air temperature
Temperature	Air temperature at the factory ranging from 292 to 298 (K)
Tool Wear	The time required and used by a machine for each printing run
Failure	Indicates whether the machine has failed or not

Table 2 Parameters and attributes of the input and target variables

Parameter	Mean	Standard deviation	Minimum	Maximum
Unique Order ID	2969	1214.02327	867	5071
Quality	1.600238	0.761683	1	3
Quantity	2331.809750	1319.670624	206	9956
Color	3.680856	0.947442	1	5
Machine	2.676100	1.337525	1	5
Humidity	55.007498	3.391050	45.070	69.940000
Temperature	294.327795	1.041168	292	300.010
Tool wear	7.772699	4.398902	0.686667	33.186667
Failure	0.034483	0.182487	0	1

correctly as expected), data scaling was performed initially, using the Log Transformation methodology. The method was used as it enables data measurements to become more symmetric to a normal distribution. After the scaling the dataset was divided into a training set (80%) and test set (20%) (Step 1).

Furthermore, high data dimensionality has shown to have a direct effect on classification accuracy, increasing the rate of misclassification and thus reducing the overall accuracy of a classification algorithm. Therefore, dimensionality reduction was also performed using Principal Component Analysis (PCA). Specifically, the dataset PCA enables the conversion of correlated features found in the high dimensional space into a series of uncorrelated features in the low dimensional space, that

depict the linear combination of existing variables, and for that reason it has become a necessity before applying any data sampling approach [7] (Step 2).

To effectively deal with the class imbalance, three different sampling techniques were employed, namely, Random UnderSampling (RUS) [8], Synthetic Minority Oversampling Technique (SMOTE) [9] and the Adaptive Synthetic sampling approach (ADASYN) [10] (Step 3).

These techniques operate in a feature space aiming either to under-sample the majority class data or oversample the minority one. On the one hand, under-sampling techniques such as RUS, are used to improve imbalance levels of the classes to the desired target, by reducing the number of majority instances. However, the removal of instances from the majority class is performed without replacement, meaning that useful information might be permanently lost. In addition, due to the randomized nature of RUS, an unclear decision boundary may be resulted, affecting classifiers performance [11].

On the other hand, over-sampling approaches intent to improve imbalance levels of the classes to the desired target, by generating synthetic instances and adding them to the minority class. Unlike approaches such as random oversampling, SMOTE generates artificial instances in the minority class, based on the feature space, rather than the data space, considering linear combinations between existing minority samples. Moreover, derived from SMOTE, the ADASYN approach gives different weights to different minority samples of a given dataset, while it automatically determines the number of samples required to produce in order to achieve data balance.

The aforementioned methodology is depicted in Fig. 1.

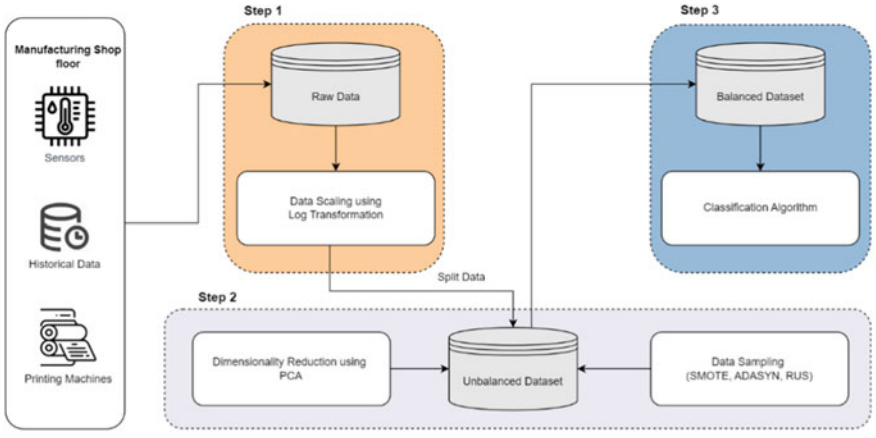

Fig. 1 Proposed methodology for imbalanced data handling in predictive maintenance

2.3 Machine Learning Models

To create the proposed framework, stratified 5-fold cross validation was used for all the experiments in this study. The base ML models were trained, using the scikit learn package [12], including:

- Logistic regression (LR) is a standard probabilistic statistical classification model that has been extensively used for classification problems across disciplines. Different from linear regression, logistic regression analyzes the relationship between multiple independent features and estimates the probability of occurrence of an event, by fitting the data onto a logistic curve. LR is affected by outliers, which greatly skews parameter estimation, reducing classification performance [13].
- k-Nearest Neighbors (kNN) [14] enables a low-power computational classification through the identification of the nearest neighbors given by a query example and using those neighbors to determine the class of the query [15].
- Decision Tree (DT) is a learner which repeatedly splits the dataset according to a cost criterion that maximizes the separation of the data, resulting in tree-like branches. In detail, the algorithm attempts to select the most important features to split branches and iterate through a given feature space. Compared with the other machine learning methods, decision trees have the key advantage, that are not characterized as black-box models and can be easily expressed as rules [16].
- Random Forest (RF) algorithms fall under the broad umbrella of ensemble learning methods. The key principle underlying the algorithm is the decision tree. Specifically, every data instance is initially classified by every individual DT, and then classified by a consensus among the individual DTs. The diversity among these individual DTs can thus further improve the overall classification performance, and so bagging is introduced to promote diversity. The advantages of using RF include its robustness to overfitting and its stability in the presence of outliers [17].

2.4 Evaluation

To compare the performance of the candidate models, the most frequently used metrics for classification are utilized, including accuracy (ACC), precision (P), recall (R), and F1-score values, calculated as [18]:

$$\mathrm{ACC} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}) \tag{1}$$

$$P = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \tag{2}$$

$$R = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \tag{3}$$

$$F1 = (2 \cdot \text{Precision} \cdot \text{Recall})/(\text{Precision} + \text{Recall}) \quad (4)$$

3 Results

Table 3 showcases the overall performance comparison between the different classification algorithms, each utilizing different sampling methodologies with the aim of achieving higher accuracy and f1-scores, for more accurate classification of machine failures in the field of Offset Printing.

In detail, the experiment results demonstrate that both Random Forest and Decision Trees algorithms performed significantly better than the rest of the base models, while Logistic Regression performed the least accurate scores. Moreover, both SMOTE and ADASYN sampling methods, showed to improve classification accuracy throughout the models, while Under-sampling had the least effect on improving classification accuracy.

Furthermore, as showcased in Fig. 2, the implementation of SMOTE and ADASYN indicated similar results, with models under SMOTE slightly outperforming the rest of the methods using ADASYN.

Table 3 Overall performance evaluation of classification algorithms under different sampling methodologies

Model	Data sampling method	Accuracy	Precision	Recall	F1-score
Logistic regression	Under sampling	0.468304	0.924963	0.468304	0.609441
	SMOTE	0.563391	0.929364	0.563391	0.692212
	ADASYN	0.543582	0.928011	0.543582	0.675927
k-nearest neighbors	Under sampling	0.541204	0.933273	0.541204	0.673145
	SMOTE	0.751189	0.925741	0.751189	0.827805
	ADASYN	0.755151	0.924929	0.755151	0.830226
Decision trees	Under sampling	0.496830	0.939275	0.496830	0.633158
	SMOTE	0.841521	0.930260	0.841521	0.882692
	ADASYN	0.828051	0.929606	0.828051	0.874811
Random forest	Under sampling	0.516640	0.931664	0.516640	0.652170
	SMOTE	**0.882726**	**0.930196**	**0.882726**	**0.905511**
	ADASYN	0.874802	0.928736	0.874802	0.900716

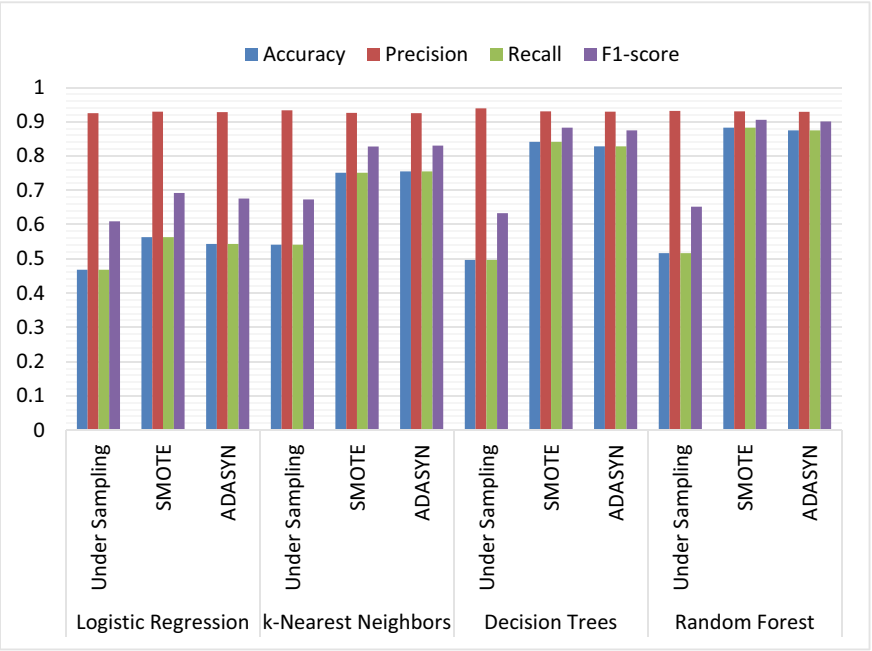

Fig. 2 Performance comparison of the proposed algorithms

4 Conclusions

Predictive Maintenance systems are utilized to predict trends, behavior patterns, and correlations by ML models in order to anticipate pending machine failures in a proactive manner, thus avoiding downtime and production stop. Machine maintenance has therefore attained critical importance for manufacturing industries such as the ones found in Offset Printing, due to the current growth in complexity of the manufacturing ecosystems.

In this study, we proposed a data sampling methodology for predictive maintenance algorithms for Offset Printing environments, which aims to effectively balance data classes and improve the performance of PdM models to accurately identify the minority classes using binary classification. The methodology consisting by the SMOTE, ADASYN, and RUS techniques, and the classification algorithms (DT, LR, KNN, RF), was generated based on a dataset from an Offset Printing company.

Overall, the results of this study indicate that the proposed methodology effectively handles data imbalances while enhancing model performance in classification accuracy, by outperforming other state-of-the-art techniques. Moreover, to the best of our knowledge, ours is the first study to explore PdM systems and data handling approaches for the Offset Printing domain.

Finally, in future work, the proposed methodology can be further extended for multi-class classifications, as well as the evaluation of further ML and DL techniques.

Acknowledgements This work has been partially supported by the PDS project, under the open call of the AI REGIO (Regions and Digital Innovation Hubs alliance for AI-driven digital transformation of European Manufacturing SMEs) project, funded by the European Commission under Grant Agreement number 952003 through the Horizon 2020 program (https://www.airegio-project.eu/) and by the ICOS (Towards a functional continuum operating system) project, funded by the European Commission under Grant Agreement number 101070177 through the Horizon 2020 program (https://www.icos-project.eu/).

References

1. Selcuk S (2016) Predictive maintenance, its implementation and latest trends. Proc Inst Mech Eng Part B: J Eng Manuf 231(9):1670–1679
2. Bălan E, Berculescu L, Răcheru RG, Piţigoi DV, Adăscăliţa L (2021) Preventive maintenance features specific to offset printing machines. In: MATEC web of conferences, vol 343
3. Haarman M, Mulders M, Vassiliadis C (2021) Predictive maintenance 4.0: predict the unpredictable. PwC Mainnov 4(30)
4. Gazzah S, Hechkel A, Amara NEB (2015) A hybrid sampling method for imbalanced data. In: 2015 IEEE 12th international multi-conference on systems, signals & devices (SSD15), pp 1–6
5. Kalafatelis A, Panagos K, Giannopoulos AE, Spantideas ST, Kapsalis NC, Touloupou M, Kapassa E, Katelaris L, Christodoulou P, Christodoulou, Trakadas P (2021) ISLAND: an interlinked semantically-enriched blockchain data framework. In: International conference on the economics of grids, clouds, systems, and services, pp 207–214
6. Spantideas ST, Giannopoulos AE, Kapsalis NC, Angelopoulos A, Voliotis S, Trakadas P (2022) Towards zero-defect manufacturing: machine selection through unsupervised learning in the printing industry. In: Proceedings of the workshop of I-ESA, Valencia, SP
7. Kotekani SS, Velchamy I (2020) An effective data sampling procedure for imbalanced data learning on health insurance fraud detection. J Comput Inf Technol 28(4):269–285
8. Saripuddin M, Suliman A, Syarmila Sameon S, Jorgensen BN (2021) Random undersampling on imbalance time series data for anomaly detection. In: 2021 the 4th international conference on machine learning and machine intelligence, pp 151–156
9. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority oversampling technique. J Artif Intell Res 16:321–357
10. He H, Bai Y, Garcia EA, Li S (2008) ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), pp 1322–1328
11. Zuech R, Hancock J, Khoshgoftaar TM (2021) Detecting web attacks using random undersampling and ensemble learners. J Big Data 8(1):1–20
12. scikit-learn. https://scikit-learn.org/stable/. Accessed 17 Nov 2022
13. Feng J, Xu H, Mannor S, Yan S (2014) Robust logistic regression and classification. Adv Neural Inf Process Syst 27
14. Zhang Z (2016) Introduction to machine learning: k-nearest neighbors. Ann Transl Med 4(11):218
15. Angelopoulos A, Giannopoulos AE, Kapsalis NC, Spantideas ST, Sarakis L, Voliotis S, Trakadas P (2021) Impact of classifiers to drift detection method: a comparison. In: International conference on engineering applications of neural networks, pp 399–410

16. Myles AJ, Feudale RN, Liu Y, Woody NA, Brown SD (2004) An introduction to decision tree modeling. J Chemom Soc 18(6):275–285
17. Angelopoulos A, Giannopoulos A, Spantideas S, Kapsalis N, Trochoutsos C, Voliotis S, Trakadas P (2022) Allocating orders to printing machines for defect minimization: a comparative machine learning approach. In: IFIP international conference on artificial intelligence applications and innovations, pp 79–88
18. Flach P, Kull M (2015) Precision-recall-gain curves: PR analysis done right. Adv Neural Inf Process Syst 28