Rare Event Classification for Multivariate Time Series Data

Navaneeth Shanavasan, T V Nithin, Ans Baby, Jithin G P, Abhinave M S, Anagha M V, and Mohammed Shaheem C

National Institute of Technology, Calicut

Abstract. A real world dataset is provided from a pulp-and-paper manufacturing industry. The dataset includes a rare event of paper break, a common occurrence in the industry. It contains sensor readings at regular time intervals (x's) along with the event label (y). The goal is to attain satisfactory levels of precision and recall metrics for a classification model.

1 Introduction

In pulp-and-paper manufacturing industry, the occurrence of paper breaks is a significant concern, often leading to production delays, increased costs, and decreased efficiency. This challenge necessitates the development of predictive models capable of anticipating such events to enable proactive maintenance and process optimization. The provided dataset offers a unique opportunity to delve into this problem domain, leveraging multivariate time series data and machine learning techniques to predict paper breaks accurately. The dataset comprises sensor readings collected at regular intervals from a pulp-and-paper manufacturing process, along with timestamps and a binary label indicating the occurrence of a paper break event. With 61 anonymized features (x1 to x61) capturing various aspects of the manufacturing process, the dataset offers a rich source of information for predictive modeling.

2 Problem Formulation

The classification task can be approached in two ways. Firstly, we can view it as a normal classification problem, where the goal is to build a model that can accurately identify data points that causes paper break. This approach focuses on predicting the occurrence of a paper break based on the available data and features.

Alternatively, we can view the task as an early classification problem, which introduces a temporal aspect to the prediction. In this approach, we move a specific column k rows up in the dataset. For example, if k is 1, we need to predict the label for a row one row before the current row. If k is 2, we need to predict the label for a row two rows before the current row. Early detection

of a failure, such as a paper break, is crucial for preventing it and minimizing downtime. This approach aims to predict the occurrence of a paper break before it actually happens, allowing for proactive maintenance and intervention.

By considering both approaches, we can explore different strategies for predicting paper breaks and evaluate their effectiveness in preventing downtime and minimizing the impact of failures in the paper manufacturing process.

3 Methodology

3.1 Data Exploration

As shown in Fig. 1, the attribute x28 is categorical and has 8 distinct values: 51, 82, 84, 93, 96, 112, 118, and 139. The value 96 appears the most. The binary attribute x61 has a highly skewed distribution, with only 19 observations (0.1%) having a value of 1 (Fig. 2). Rest of the attributes are continuous with different scales. These have to be standardized before passing them to linear classification models. The rare event of paper break (y = 1) occurs only 124 times in the dataset containing 18398 records (Fig. 3). This means that the dataset is highly unbalanced and this must be taken in to account when using basic machine learning models for classification.

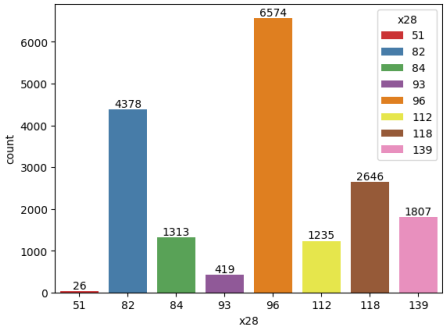

Fig. 1. Distribution of x28

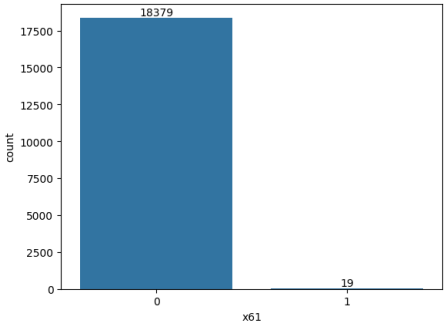

Fig. 2. Distribution of x61

3.2 Feature Engineering

Principal Component Analysis (PCA): PCA is a dimensionality reduction technique which works by transforming the original features of a dataset into a new set of orthogonal variables called principal components, which are linear combinations of the original features. These principal components capture the maximum variance present in the data. By performing PCA on the given dataset, it was observed that 33 principal components were sufficient to retain 95% of the

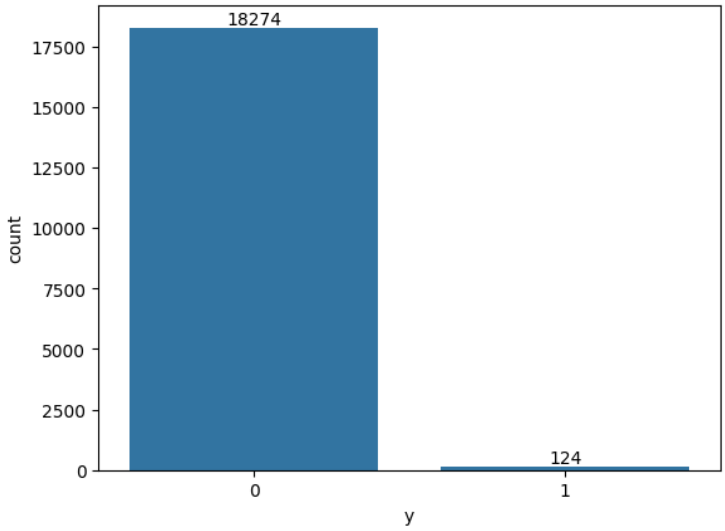

Fig. 3. Distribution of y

variance present in the original dataset. This implies that the majority of the information present in the original dataset is preserved within these 33 principal components.

Correlation Analysis: Correlation analysis was conducted to identify highly correlated attributes in the dataset. A threshold of 0.85 was set, and attributes with a correlation coefficient of 0.85 or higher were considered highly correlated. If two attributes have a correlation coefficient above the threshold, one of them was dropped to reduce the dimensionality of the dataset. By eliminating redundant attributes, we aimed to improve the efficiency of the model and reduce the risk of multicollinearity. The heatmap given in Fig. 4 illustrates the correlation matrix for the dataset.

3.3 Initial Modeling Approaches and Observations

1. The provided dataset is divided into training and testing sets,ensuring that data points up to the cutoff time of "5/21/99 23:58" are selected for the training sample set, while the remaining data is allocated to the testing sample.Various models have been trained using the training sample, both with and without early detection.
2. During the initial phase, some of the models that exhibited notably favorable results include Support Vector Classifier, Random Forest, and Adaboost.
3. Without incorporating early classification, nearly all models demonstrated a fairly good F1 score.
4. Incorporating early classification yielded somewhat underwhelming results, particularly with Adaboost and Random Forest. However, the Support Vec-

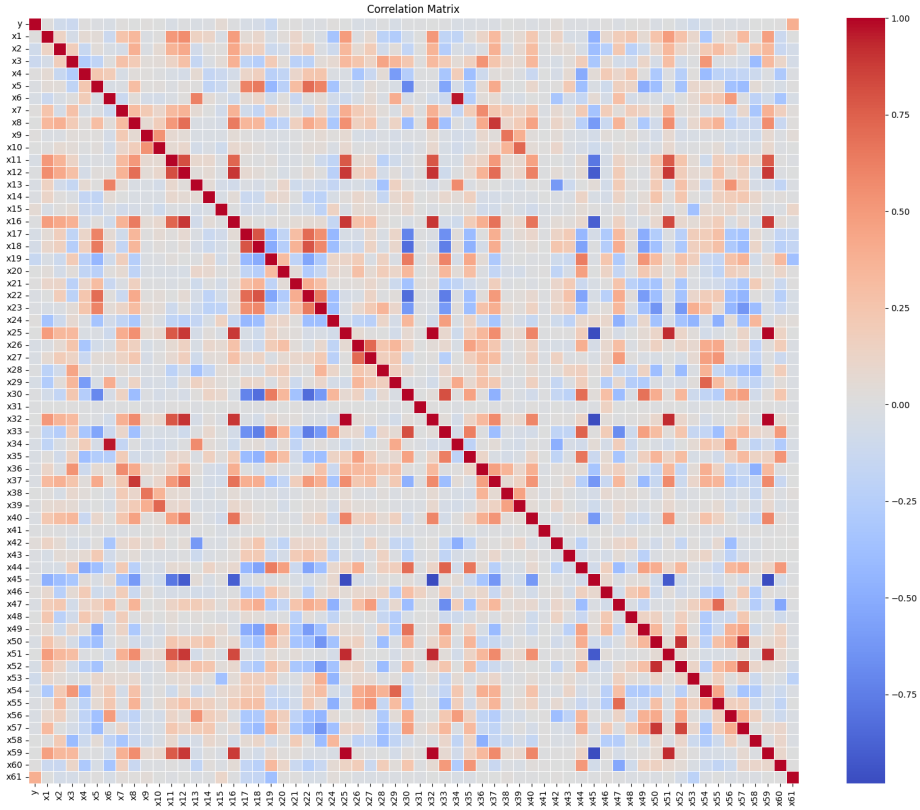

Fig. 4. Heatmap Showing Correlation Matrix

tor Classifier exhibited relatively positive outcomes compared to the other ones.

5. In the preliminary phase, the observed results indicate the potential for significant enhancements through the meticulous fine-tuning of model parameters. A comprehensive exploration of hyperparameter tuning across different models promises to unlock further performance improvements, refining the accuracy and efficacy of the predictive models under consideration. This strategic approach to optimization is poised to contribute substantially to the refinement of outcomes as the project progresses.