Project Description:

Classification is a big part of machine learning, and we have been able to improve our classification capabilities on many practical tasks with sophisticated mathematical models. Recently, deep learning has gained tremendous attention worldwide with its state-of-the-art performance on tasks we never thought a machine could do better on than a human.

In this project, we deal with a more difficult real-world dataset from a paper mill. Namely, we have a paper mill producing reels of paper at one end and ingredients flowing in to the other. The issue is that this machine breaks down occasionally, and these breaks are incurring a significant loss to the company. It is reported that even a small improvement over these breaks, as little as 5%, can lead to a significant financial gain. Hence, we wish to be able to predict when a break is about to occur and take measures to prevent it in advance.

The data consists of 61 anonymous fields(x1 to x61) that contain information about both raw materials(for e.g. amount of pulp fiber) and process variables.(e.g. motor speed). There are sensors placed throughout the machine to gauge these figures. Note variables x28 and x61 are the only categorical variables and others are continuous, numerical variables. It is known that for each variable, their center and scale could have been adjusted. Typically, each data point was recorded on a 2-minute interval, with some exceptions in cases of long breaks. Some breaks last more than 3 hours while others are short. As can be expected, the proportion of time the machine is broken is small(about 6%), and our job is to detect these rare events in advance.

Note that once a break occurs, the machine remains broken for some time, which yields a sequence of consecutive 1's in the y-label column.

For the purpose of prediction of breakdown, data is preprocessed by the following procedure:

1. Labels are shifted by one time-step to the past so that we say y=1 when we think a breakdown is about to happen in 2 minutes.
2. All consecutive 1's in the label are removed except for the first 1

In summary, we have a multivariate time series data with binary labels that are extremely unbalanced. You are free to use any models for the task and techniques for training including data augmentation.

The efficacy of your model will be evaluated based on F1 score on a held-out test dataset. Note that the final model should be able to sequentially process incoming data and output predictions for a future break in real time.

References:

1. Ranjan, C., Mustonen, M., Paynabar, K., Pourak, K. (2018). Dataset: Rare Event Classi_cation in Multivariate Time Series. arXiv preprint arXiv:1809.10717.