Data Lab: Paper Mill Challenge from INFORMS

ISyE, Georgia Institute of Technology

Illustration of a Working Machine

Ranjan et al., “Dataset: Rare Event Classication in Multivariate Time Series”

A Snapshot of Raw Data

- Real-world dataset from a paper mill
- Variables on raw materials (e.g. amount of pulp fiber) and process variables (e.g. motor speed)
- 2-minute interval
- Want to predict the breakdown some time before it does

time	y	x1	x2	x3	x4	x5
5/1/1999 0:00	0	0.36018343	8.91635343	10.2989914	332.475535	125.067515
5/1/1999 0:02	0	0.45923843	8.97028643	10.3763884	335.208506	125.057612
5/1/1999 0:04	0	0.34736643	8.83139443	10.0416004	333.105845	125.047709
5/1/1999 0:06	0	0.28510843	8.75385443	10.3711354	332.139414	125.038203
5/1/1999 0:08	0	0.24909643	8.76286043	10.0615974	334.245188	125.031031
5/1/1999 0:10	0	0.36477143	8.90104243	10.3096754	333.121043	125.023844
5/1/1999 0:12	0	0.29684343	8.98257843	10.2738164	337.659312	125.016673
5/1/1999 0:14	0	0.37991943	8.81374143	10.3205544	340.285502	125.009486
5/1/1999 0:16	0	0.32620643	8.95928243	10.2058624	341.869883	125.002299
5/1/1999 0:18	0	0.44177043	8.87322743	10.1482004	342.138804	124.995127
5/1/1999 0:20	0	0.37727443	8.70079043	10.2782924	346.528696	124.98794
5/1/1999 0:22	0	0.38846443	8.89498943	10.2072854	344.013499	124.980769
5/1/1999 0:24	0	0.45051843	9.05673843	10.0782074	349.233256	124.973582
5/1/1999 0:26	0	0.41694843	9.05803943	10.1705144	351.716532	124.966395
5/1/1999 0:28	0	0.45314943	9.11472343	10.3662874	348.835123	124.959223
5/1/1999 0:30	0	0.41008143	9.16660443	10.2131864	351.163889	124.952036
5/1/1999 0:32	0	0.39049943	9.16450443	10.3095974	351.587626	124.944849
5/1/1999 0:34	0	0.39916343	9.11866043	10.4412944	348.761332	124.937678
5/1/1999 0:36	0	0.38719243	9.11964543	10.4148124	348.717692	124.930491
5/1/1999 0:38	0	0.33244343	9.37483143	10.5315324	352.87086	124.923319
5/1/1999 0:40	0	0.38642943	9.34605443	10.4557064	353.8986	124.916132

Data Statistics

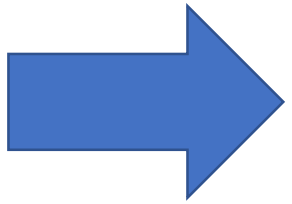
- 19,461 records
- Columns
  - Time Stamps
  - Y-label: Indicator (0,1 encoded) of machine breakdown
  - Predictor variables: 61 types, continuous except for two (x28, x61)
- Y-label is highly imbalanced (machine breaks down seldomly)
- Given records up to time t, can we predict y-label at t+1 or even t+2?
- For this session, we only try to predict 1 time-unit (2 mins) ahead

Data Preprocessing

- Y-labels shifted by 1 time-unit
- Consecutive 1's removed
- End up with 16,446 training samples (in "train.csv") and 1827 test samples (in "test.csv").

Raw Data

5/2/1999 3:42	0
5/2/1999 3:44	0
5/2/1999 3:46	0
5/2/1999 3:48	0
5/2/1999 3:50	0
5/2/1999 3:52	0
5/2/1999 3:54	0
5/2/1999 3:56	0
5/2/1999 3:58	0
5/2/1999 4:00	0
5/2/1999 4:02	1
5/2/1999 4:04	1
5/2/1999 4:06	1
5/2/1999 4:08	1
5/2/1999 4:10	1
5/2/1999 4:12	1
5/2/1999 4:14	1
5/2/1999 4:16	1
5/2/1999 4:18	1
5/2/1999 4:20	1
5/2/1999 4:22	1
5/2/1999 4:24	1
5/2/1999 4:26	1
5/2/1999 4:28	0
5/2/1999 4:30	0
5/2/1999 4:32	0
5/2/1999 4:34	0
5/2/1999 4:36	0
5/2/1999 4:38	0
5/2/1999 4:40	0
5/2/1999 4:42	0
5/2/1999 4:44	0
5/2/1999 4:46	0



Preprocessed Data

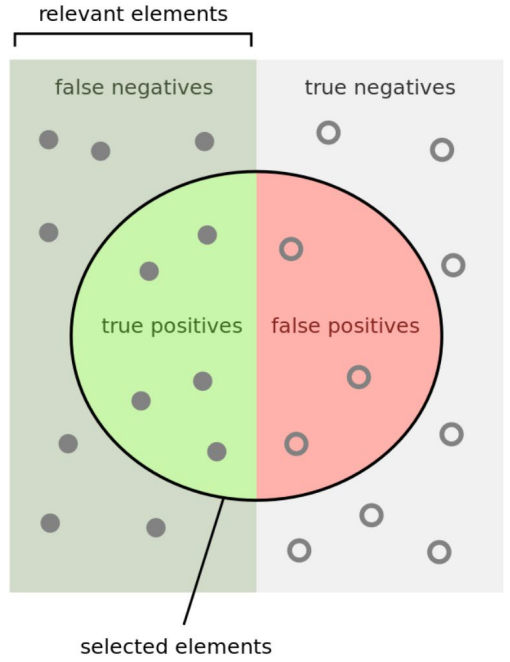
5/2/1999 3:42	0
5/2/1999 3:44	0
5/2/1999 3:46	0
5/2/1999 3:48	0
5/2/1999 3:50	0
5/2/1999 3:52	0
5/2/1999 3:54	0
5/2/1999 3:56	0
5/2/1999 3:58	0
5/2/1999 4:00	1
5/2/1999 4:28	0
5/2/1999 4:30	0
5/2/1999 4:32	0
5/2/1999 4:34	0
5/2/1999 4:36	0
5/2/1999 4:38	0
5/2/1999 4:40	0
5/2/1999 4:42	0
5/2/1999 4:44	0
5/2/1999 4:46	0

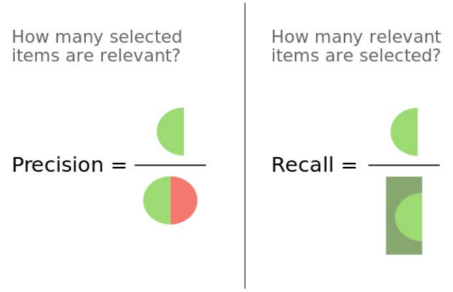
Evaluation Metric

- F-1 score: Harmonic mean of recall and precision

$$F_1 = \left(\frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} \right) = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

- https://en.wikipedia.org/wiki/F1_score





Basic Suggested Approaches

- Logistic Regression
- Gradient Boosting
- Recurrent Neural Network
- Benchmark F1-score: 0.114

Practical Implementation

- Resource Access from:

https://github.com/susain/Data_Lab_paper_mill

- Anaconda:

https://www.anaconda.com/distribution/

- Jupyter Notebook
- F1-measure

Live Code Example