Multi-Machine Disruption Challenge by ITU 2023: Gradient Boosting Estimators for Fusion Energy Prediction
Final report of team WolfPack

SIaw-Darko Emmanuel
Immanuelsiawdarko@gmail.com

ABSTRACT

This research paper seeks to highlight the significance of deep learning in the field of multi-machine disruption fusion systems, its shortcoming in terms of data availability to improve the model performance and proposes a new data preprocessing approach and an ensemble of gradient boosting estimators trained with three distinctive tokmaks (Alcator C-Mod, J-TEXT, and HL-2A), to classify C-Mod as a future device.

SIGNIFICANCE OF STUDY

Significance of this study is to show an alternative to deep learning techniques used modelling fusion systems when data availability becomes an inherent issue and exemplifies that improved performance of over 0.8 percent and over is guaranteed when modeled with the proposed model.

RESEARCH OBJECTIVES

1. Proposes a new data preprocessing approach with three distinctive tokmaks (Alcator C-Mod, J-TEXT, and HL-2A).
2. Illustrates an ensemble of gradient boosters its performance when trained on three distinctive tokmaks to universally classify c-mod as a future device.

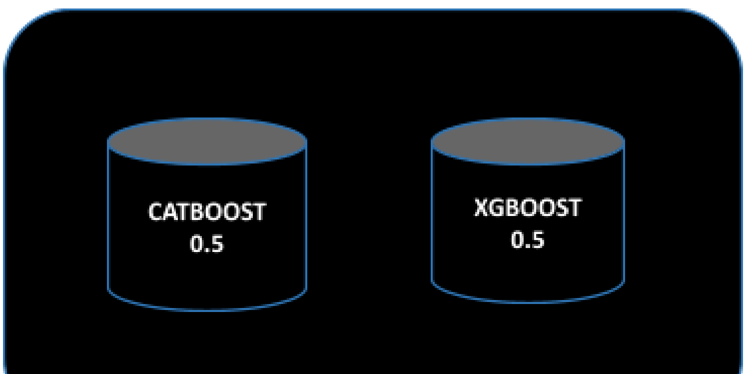
INTRODUCTION

In our current generation the global challengers related to energy pose a significant challenge for us. Nuclear fusion involves the merging of two light nuclei, a process that releases a substantial amount of energy. This fusion occurs within a plasma, a state of matter consisting of ions and free electrons, and represents a clean, safe, and accessible energy source. Although fusion energy is not yet commercially viable source of energy, scientists worldwide are collaborating to make this a reality. The use of deep learning in the field of fusion energy systems over the years has been highly sought after in the process seeking a better, low cost and highly improved model. When training deep learning systems we may end facing a data availability challenge which might result in a less impressive model. To avoid this, we exploit the use of gradient boosting estimators paired in an ensemble.

METHODOLOGY

Data used involves three distinctive tokmaks Alcator C-Mod, J-TEXT, and HL-2A containing 20, 2136, 975 shots respectively. While preprocessing shots, we realized a few inconsistencies with the training and testing data, such shots not having the same number tags in the CMOD and HL_2A dateset. Data preprocessing approach involves first retrieving tags in a shot, using the retrieved tags to retrieve the signals their related to and finally employing A sum of Squares approach that is, squaring all signals and summing them up as single value to represent a tag as feature in a dataframe for a shot. This process is done in a sequentially for all shots.

Below is a diagram of the employed model is referred to as a pipeline.



Ensembling is employed because can often improve the precision and recall of the model thus with a ratio of 0.5 on each gradient boosting estimator, we end of reducing the FPR all while increasing the TPR. As stated earlier these datasets have different operational scales but this wont affect the performance of the gradient boosting models used in the overall pipeline. Also our choice of estimators wont be affected with the the addition of inf values to represent features of shots which were not available. For the above model, the following final features or diagnostics were selected, plasma_current, torodial magnet field, horizontal displacement, vertical displacement, line integral density(centre chord), c3 radiation, loop voltage, axuv 09-22 , soft X_ray 02-38, poloidal mirnov probes 01 - 18, toloidial mirnov probes (01,02,03,04,05,06,07,09,10) , density measurment1, density measurment2, density measurment3, density measurment4, greenwald_fraction, lm_proxy, q95_proxy, ip_error_fraction, radiation_fraction , rotating_mode_proxy, unknown_bp_ka_top, unknown_bp_ef_top, unknown_bp_ka_bot, 'unknown_aminor', unknown_betan, unknown_kappa, unknown_li, unknown_wplasm. Features with the unknown as an identifier are features whose diagnostic names weren't known but were used since they existed in both train and test samples.

RESULTS

Performance metrics collected over each ensemble StratifiedKFold

	train_auc	val_auc	test_auc	train_f1	val_f1	test_f1	train_log	val_log	test_log	train_recall	val_recall	test_recall	train_acc	val_acc	test_acc	train_precs	val_precs	test_precs
0	0.995404	0.962402	0.537501	0.995818	0.965483	0.595131	0.053004	0.107903	0.903750	0.993631	0.949495	0.350365	0.995819	0.965517	0.600484	0.995442	0.959184	0.387097
1	0.998807	0.974472	0.544827	0.998955	0.973252	0.599771	0.036951	0.099060	0.875415	0.998179	0.980000	0.372263	0.998955	0.973180	0.602906	0.999088	0.951456	0.395349
2	0.997787	0.960683	0.542989	0.997909	0.965414	0.598918	0.044142	0.096542	0.965042	0.997268	0.940000	0.364964	0.997909	0.965517	0.602906	0.997268	0.969072	0.393701
3	0.997961	0.945683	0.511663	0.997910	0.953733	0.580500	0.040329	0.116597	0.869204	0.998179	0.910000	0.218978	0.997909	0.954023	0.610169	0.996364	0.968085	0.357143
4	0.999545	0.956366	0.528417	0.999652	0.957894	0.586239	0.030741	0.130832	1.024139	0.999089	0.950000	0.343066	0.999652	0.957854	0.590799	1.000000	0.940594	0.373016
5	0.998698	0.966894	0.541204	0.998607	0.973041	0.595725	0.040856	0.114529	0.858670	0.999089	0.940000	0.372263	0.998606	0.973180	0.598063	0.997273	0.989474	0.389313
6	0.997505	0.933789	0.539339	0.997561	0.945756	0.596032	0.045282	0.119626	0.925535	0.997268	0.880000	0.357664	0.997561	0.946360	0.600484	0.996360	0.977778	0.388889
7	0.994664	0.953789	0.541151	0.995121	0.961444	0.598036	0.059750	0.112974	0.873524	0.992714	0.920000	0.357664	0.995122	0.961686	0.602906	0.994526	0.978723	0.392000
8	0.995966	0.980683	0.544801	0.996514	0.980861	0.600929	0.048983	0.079166	0.927299	0.993625	0.980000	0.364964	0.996516	0.980843	0.605327	0.997258	0.970297	0.396825
9	0.994491	0.969472	0.544801	0.995120	0.969405	0.600929	0.057339	0.106961	0.887311	0.991803	0.970000	0.364964	0.995122	0.969349	0.605327	0.995430	0.950980	0.396825
10	0.996030	0.953944	0.546639	0.996168	0.950480	0.601791	0.055108	0.160488	0.854955	0.995446	0.970000	0.372263	0.996167	0.950192	0.605327	0.994540	0.906542	0.398438
11	0.995687	0.963486	0.537501	0.996168	0.969102	0.595131	0.048862	0.103055	0.933310	0.993631	0.939394	0.350365	0.996169	0.969231	0.600484	0.996350	0.978947	0.387097

CONCLUSION

We briefly described what nuclear fusion was, how multi-machine predictive model is beneficial, thoroughly explained the data preprocessing steps and how the model was set up and trained, and finally showed a performance metrics sheet for training, validation and testing dataset.