Data-driven approach to Tokamak: 30ms before disruption.

Enrico Trotti

3rd September 2024

Overview

The problem

The models

The data

Results

Conclusions

Approaches to plasma disruptions

Plasma disruptions: sudden and unexpected plasma terminations (only in Tokamak) devices → can damage the whole fusion device, with expensive repairs and safety risks.
Early detection is crucial to mitigate these risks.

High non linearity → predictions mainly through 2 approaches

- **data-driven**: a flexibility and high accuracy with sufficient data, but often not cross-device predictive.
- **model-based**: good interpretability, but limited adaptability to new scenarios.

Our approach: Data-driven approach with experimental databases from Alcator C-Mod, DIII-D, and EAST Tokamaks.

Key contributions

- Dimensionality reduction on high-dimensional plasma data.
- Deep learning framework that enables cross-device knowledge transfer.
- Data integration from different Tokamak (generalization without overfitting).
- The three Tokamaks have distinct characteristics:
  - **EAST:** Medium size (R=1.85m, a=0.45m) superconducting Tokamak with a hybrid first wall (C, Mo, W).
  - **DIII-D:** Mediumsize (R=1.67m, a=0.67m) Tokamak with a carbon wall; many disruptions are preceded by a locked mode.
  - **C-Mod:** Small (R=0.68m, a=0.22m) Tokamak with high energy density, high magnetic field, and a wall made of a high Z metal (Mo).
- These combined features cover a substantial fraction of ITER's characteristics.

Our dataset: C_Mod_for_nn.mat

Some of the signals in the dataset, together with their description

Signal Description	**Symbol**
Electron density / Greenwald density	Greenwald_fraction
Distance between the plasma and the lower divertor	lower_gap
Current centroid vertical position error	z_error
Plasma elongation	kappa
Normalized plasma pressure (ratio of thermal to poloidal magnetic pressure)	beta_p
Radiated power/Input power	radiated-fraction
Loop voltage V$_{loop}$	v-loop
Safety factor at the 95% flux surface	q95
Normalized internal inductance	li

Model Comparison

- **Support Vector Machine (SVM):**
  - **Pros:** Effective in high dimensional spaces, good overfitting.
  - **Cons:** Computationally expensive; sensitive to parameter.
- **Neural Networks:**
  - **Pros:** Flexible and scalable with large datasets.
  - **Cons:** Requires large datasets and computational resources; tend to overfit without careful regularization.
- **Logistic Regression:**
  - **Pros:** Simple and interpretable; works well with linearly separable data.
  - **Cons:** Limited capacity for complex patterns; not suitable for high dimensional or nonlinear problems.
- **Random Forest:**
  - **Pros:** Handles high dimensional data well, robust to overfitting and works well with both linear and nonlinear data.
  - **Cons:** Slower for large datasets; less interpretable compared to simple models like logistic regression.

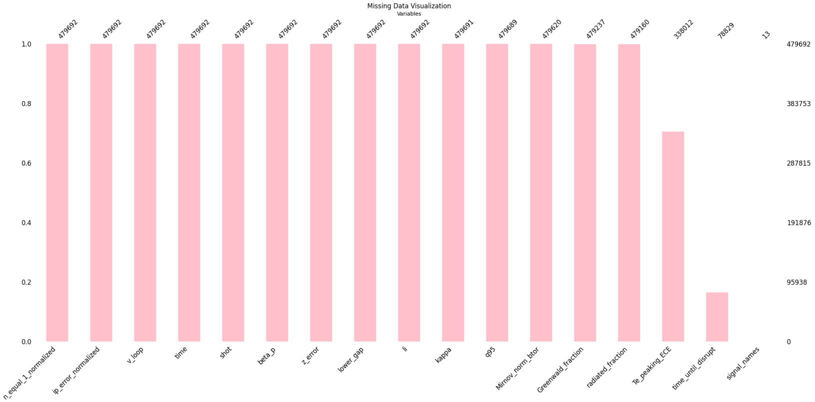
Why Random Forest?

Random Forest (RF) has several important features:
1. Can rank features by importance (which mostly contributes) → better refining the model
2. Good vs overfitting: combination of multiple decision trees
3. Deal well with rare events (Here interruption are much less frequent than other cases)

Other methods:
SVC (part of SVM): Computationally expensive.
NN: Too complex for this problem.
Logistic Regression: simple, but could not be efficient in this case, due to non linear interaction between != features.

Missing values

One should always starting with a clear idea about the data in the dataset. To do it, one could search for the missing values in the dataset (see graph).



Results: the ROC curve

The ROC (Receiver Operating Characteristic) curve is an index of how good the model is. Parameters:

- **Area Under the Curve (AUC)**: how good are the predictions: perfect (=1), equal performation than the random guessing (=0.5), worse than random (<0.5)
- **True Positive Rate (TPR)**
- **False Positive Rate (FPR)**

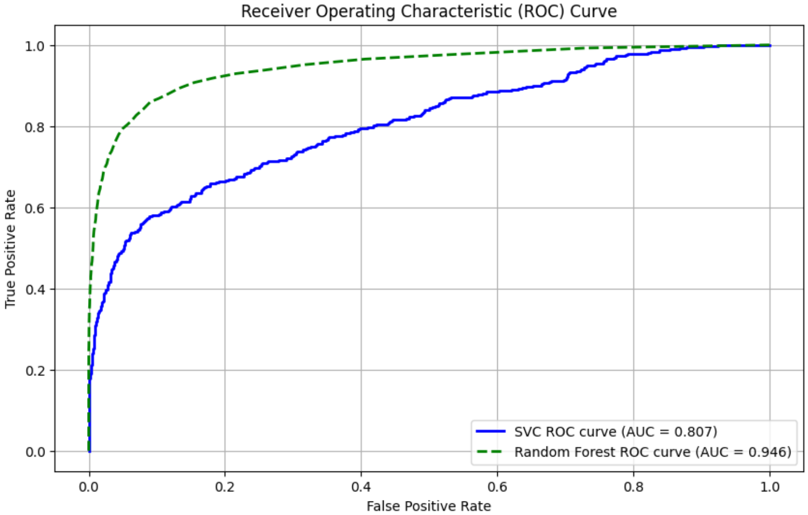
$$TPR = \frac{TP}{TP+FN} \text{ and } FPR = \frac{FP}{FP+TN},$$

where TN, TP, FN, FP are True Negative, True Positive, False Negative, False Positive.
To have AUC =1, the curve should be a step function touching the value (TPR,FPR)=(1,0).

Example

The AUC should always refer to a desired threshold. This threshold
Let us consider an example:
If my AUC = 0.9 means that the general probability of having good guess is 0.9, but this value is the combination of all the evaluations: some can have a probability of 0.98 to be correct, and other of 0.82,0.88, 0.92 and so on...
If I would like to have almost perfect guessing, I put my threshold at 0.95. Then the model will return only the points whose probability to be correct is higher than 0.95, so all the other with lower probability will not be considered.
In this case, if I have a disruption predicted with 0.85 probability, this will be returned like if it was safe. Thus a too high threshold is not always good.

Results: the ROC curve



Conclusions

- Random Forest works well (AUC>0.9) and, in comparison, better than SVC model.
- Random Forest is also faster than SVC model (it required half of the time).
- The results for C-Mod Tokamak seem promizing, even if there is always room for improvement (RF can find the most relevent features).
- Next step is to do the same thing with the other 2 Tokamaks and to compare their results after a proper normalization.
- For future: This can be done considering more and more Tokamaks with some required characteristic.

Benefits of the solution

- Higher safety in fusion reactor operations
- Reduced costs (build a model is cheaper than repairing a Tokamak)
- Better efficiency in the experiments (less time used to repair → more time used for testing)
- (Next step) Scalability to different reactor designs
- Contribution to future fusion research

Thank you for your attention