Explainable Boosting Classifier for malware detection and local explanation

Alexandre R de Mello, Vitor Gama Lemos, Emilio Simoni - PSafe Cyberlabs

Motivation

We propose using Explainable AI to identify malware in Portable Executable (PE) files and to understand the prediction decision by listing the features that contribute most to the model's decision. We train an Explainable Boosting Classifier (EBC), which is a generalized additive model from InterpretML (https://interpret.ml/), on 20.000 files and to each file from test set that we want to further investigate we can use the local explanation to check the most relevant features given a prediction. The lack of explicability increase the challenge on understanding why a model fail on classifying certain files, and does not provide precise information regarding the model decision making.

Data Collection

- 28.000 portable executable files (53,4GB size) in the .exe or .dll format
- 14.000 malware, 14.000 benign files
- Collected from January to March 2022.
- The malware were downloaded from the Malware Bazaar website (https://bazaar.abuse.ch/)
- The benign files were randomly selected from trusted sources
- To avoid data leakage during tests, the first 10.000 files acquired from each class are the the training set, and the newest 4.000 files are the test set.
- The hash function (sha256) of all files are available at https://github.com/areeberg/OSG.

Work in Progress

- Feature engineering and selection is under development to improve EBC performance
- An ensemble classifier composed of EBC and other non-Explainable tree-based models

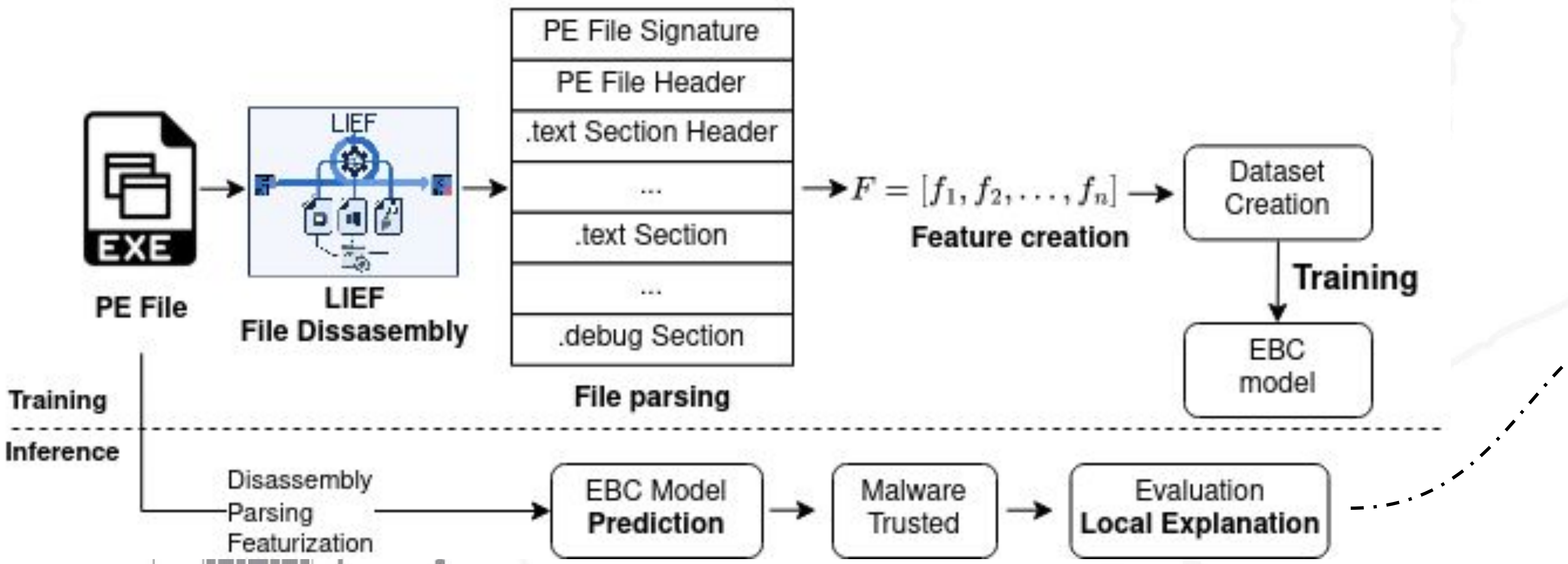
Challenges

- The EBC local and global explanation comprehension is time consuming and requires a cybersecurity specialist]
- There are many possible explanations to the local and global (most relevant features to a model) features:
  - A benign file may have a behaviour similar to a malware, e.g. a file encrypter software
  - The file may have the wrong label
  - The model may be over or underfitting

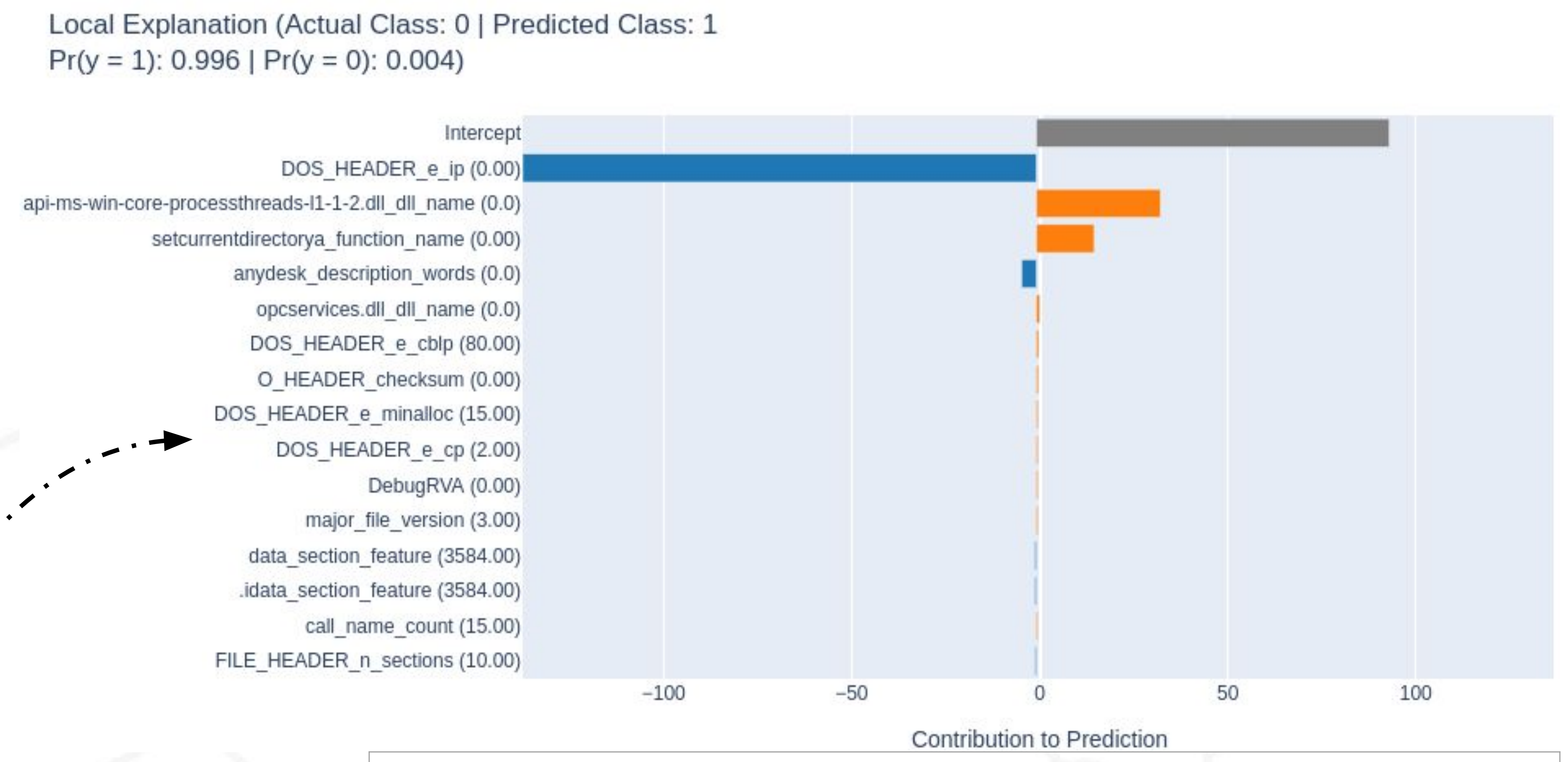
Conclusion

- To know the most relevant features given a prediction provides a good hint on where to look in a file to confirm if it is a malware or a benign.
- The most important features may indicate the malware family or the file behaviour

Machine Learning Pìpeline



Local Explanation



EBC classification scores			
	Precision	Recall	F1-Score
Benign	**1**	**0.97**	**0.98**
Malware	**0.8**	**0.98**	**0.89**
Accuracy	**0.97**		
Weighted Avg	**0.97**		

10TH ANNUAL HOT TOPICS
in the SCIENCE OF SECURITY
HOTSOS 2023
April 3 - 5, 2023 | Virtually Hosted
THE NATIONAL SECURITY AGENCY