Universidade de Brasília

Faculdade UnB **Gama**

Feature selection using SHAP: An Explainable AI approach

Author: Miguel Pimentel
Advisor: Phd. Nilton Correia da Silva

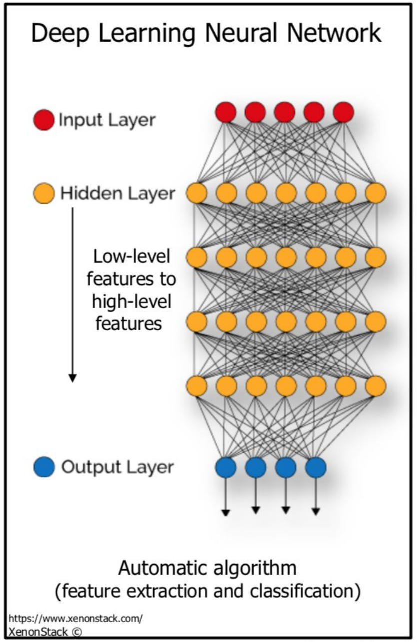
Reserach Goals

- Understand how SHAP works as feature selection tool by measuring performance metrics, training time, accuracy
- We apply SHAP as study of case

Background: **Black Box**

- Black boxes are models usually complex that present comprehension gaps, impairing people’s understanding.
- Brings questions such as:
  - Why did you do that?
  - When Can I trust you?
  - When do You succeed?
  - How do I correct an error?

Background: **Black Box**

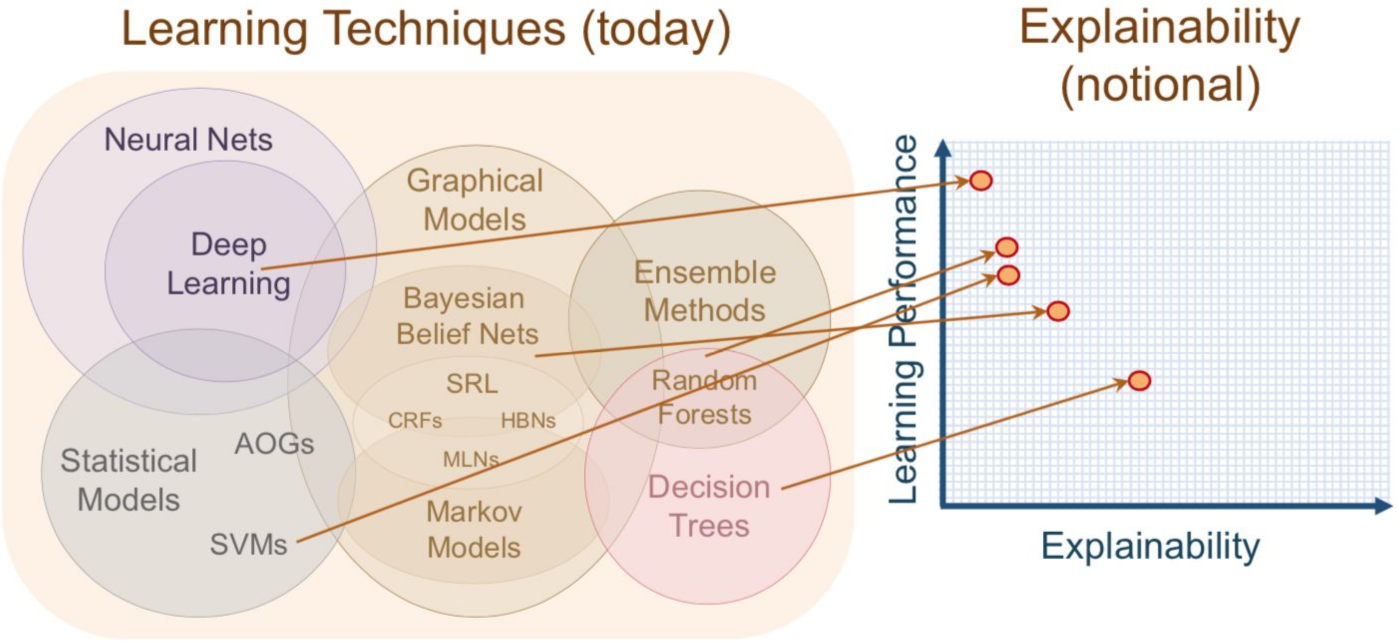

Deep Learning Neural Network
Input Layer
Hidden Layer
Low-level features to high-level features
Output Layer
Automatic algorithm
(feature extraction and classification)
https://www.xenonstack.com/
XenonStack ©

E.g Black Box - Source: XenonStack

Background: **Explainable AI**

- Explainable AI (XAI) is an area of artificial intelligence research related to the ability in which humans can understand AI solutions.
- XAI contrasts with the concept of "black box"
- There was a common belief that a trade off must be done in favour of interpretability or accuracy

Background: **Explainable AI**



XAI Initial Concept - Source: DARPA XAI

Background: **Explainable AI**

- XAI presents other relevant characteristics:
    - Verification of the system
    - Improvement of the System
    - Learning from the system
    - Compliance to legislation

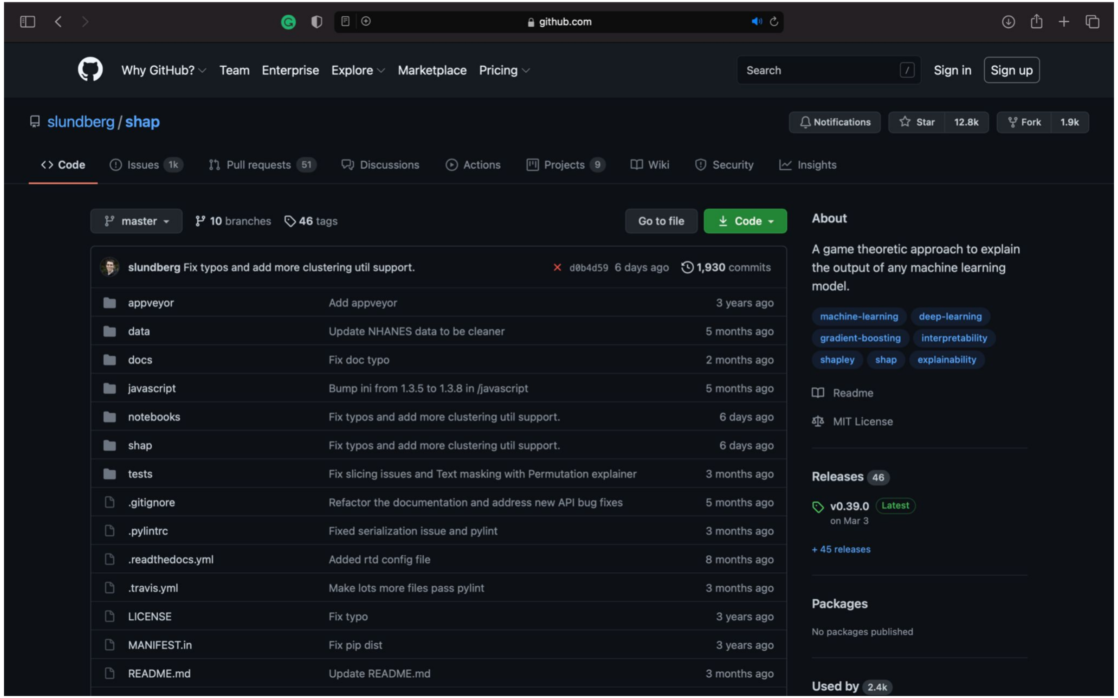
Background: **SHAP - General Idea**

Suppose you had to explain a machine learning model that calculates the value of an apartment. There are several attributes that can set your price, for example, covered parking, swimming pool, pets friendly, size, location, etc.

Background: **SHAP**

- Shapley Values
- Understand the impact of each feature in the final prediction
- Post-hoc
- Model Agnostic
- Optimized library for Shapley Values
- Python Library

Background: **SHAP**



slundberg Fix typos and add more clustering util support. ✕ d0b4d59 6 days ago 1,930 commits

appveyor	Add appveyor	3 years ago
data	Update NHANES data to be cleaner	5 months ago
docs	Fix doc typo	2 months ago
javascript	Bump ini from 1.3.5 to 1.3.8 in /javascript	5 months ago
notebooks	Fix typos and add more clustering util support.	6 days ago
shap	Fix typos and add more clustering util support.	6 days ago
tests	Fix slicing issues and Text masking with Permutation explainer	3 months ago
.gitignore	Refactor the documentation and address new API bug fixes	5 months ago
.pylintrc	Fixed serialization issue and pylint	3 months ago
.readthedocs.yml	Added rtd config file	8 months ago
.travis.yml	Make lots more files pass pylint	3 months ago
LICENSE	Fix typo	3 years ago
MANIFEST.in	Fix pip dist	3 years ago
README.md	Update README.md	3 months ago

About

A game theoretic approach to explain the output of any machine learning model.

machine-learning deep-learning gradient-boosting interpretability shapley shap explainability

Readme

MIT License

Releases 46

v0.39.0 Latest on Mar 3

+ 45 releases

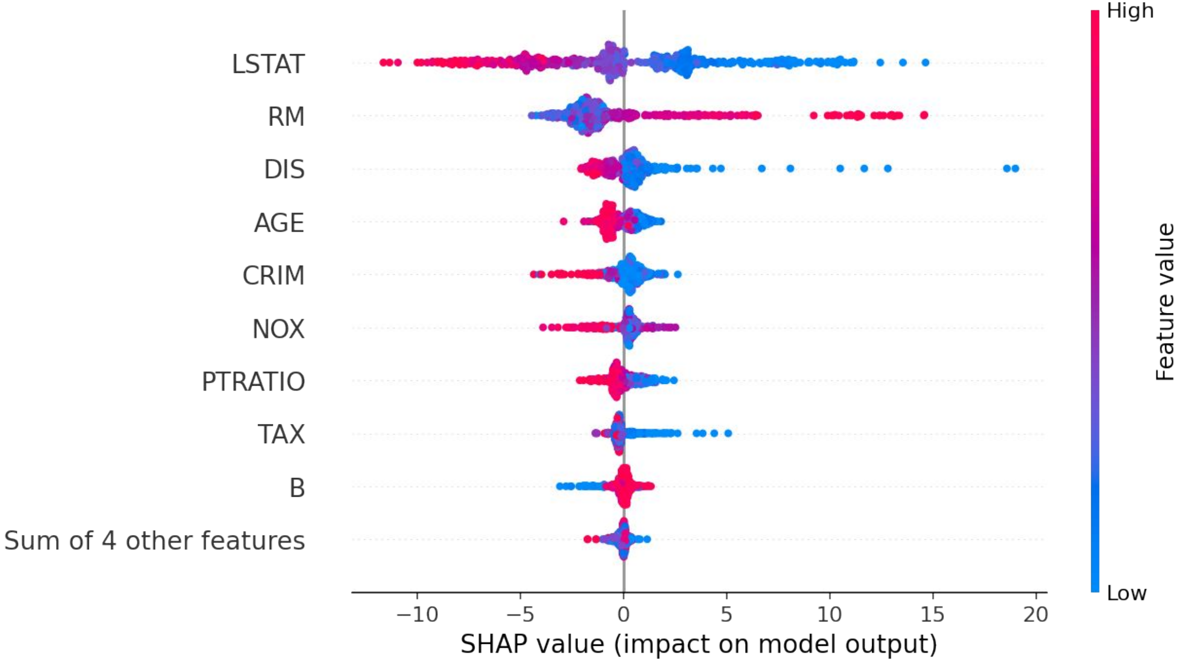
Packages

No packages published

Used by 2.4k

SHAP Repository - Source: SHAP

Background: **SHAP**



Example SHAP Output - Source: SHAP

Background: **Feature Selection**

- This concept of selecting features that are relevant to an AI model is called feature selection
- Feature Selection can be classified as: Filter; Wrapper; and Embedded.
- Feature Selection could bring some benefits, such as: Reduces Overfitting; Improves Accuracy; and Reduces Training Time.

Background: **Modelos**

- In the experiments were used the following models:
  - Random Forest
  - Catboost
  - LightGBM
  - XGBoost

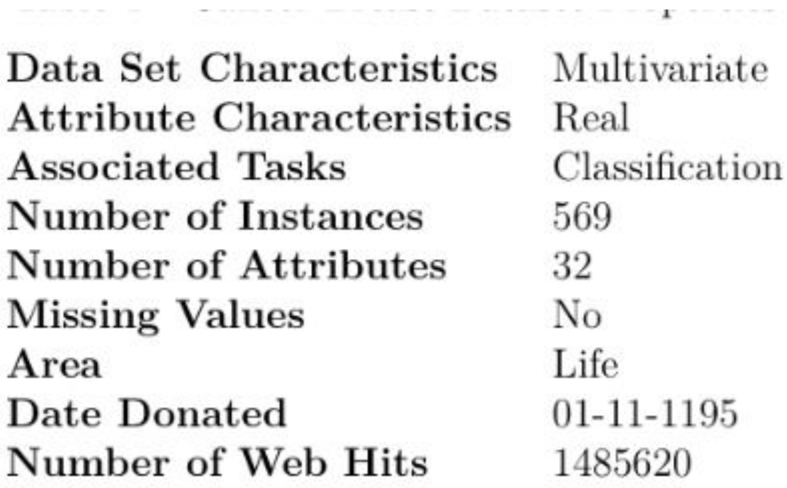
Materials & Methods: **Metrics**

- Performance: Accuracy; Precision; Recall; and F1 Score
- Training time
- Storage

Materials & Methods: **Hardware**

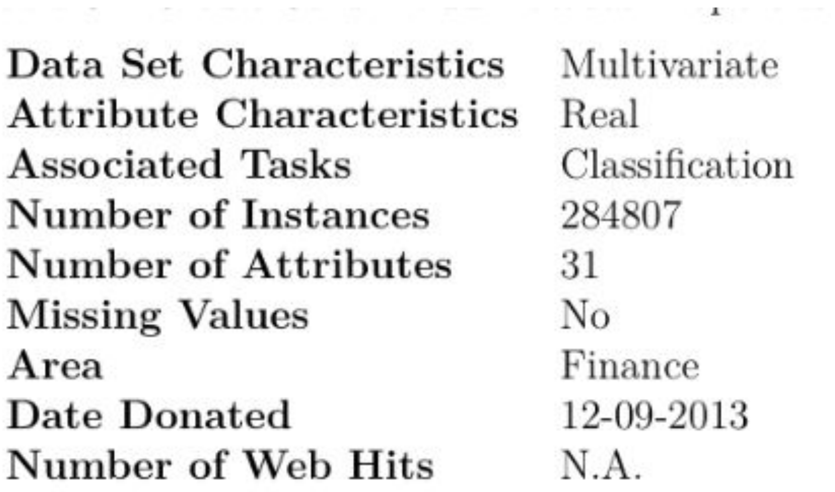
- **Processor:** Intel Core i5 (10th generation), 4 cores and 2.0 GHz, Turbo Boost up to 3.8 GHz, with 6 MB shared L3 cache
- **RAM memory:** 16GB LPDDR4X integrated memory with 3733 MHz
- **Graphics Chip**: Intel Iris Plus Graphics
- **Storage:** 512 GB SSD
- **Operating System:** macOS Big Sur 11.2.3

Dataset: **Cancer Breast Dataset**



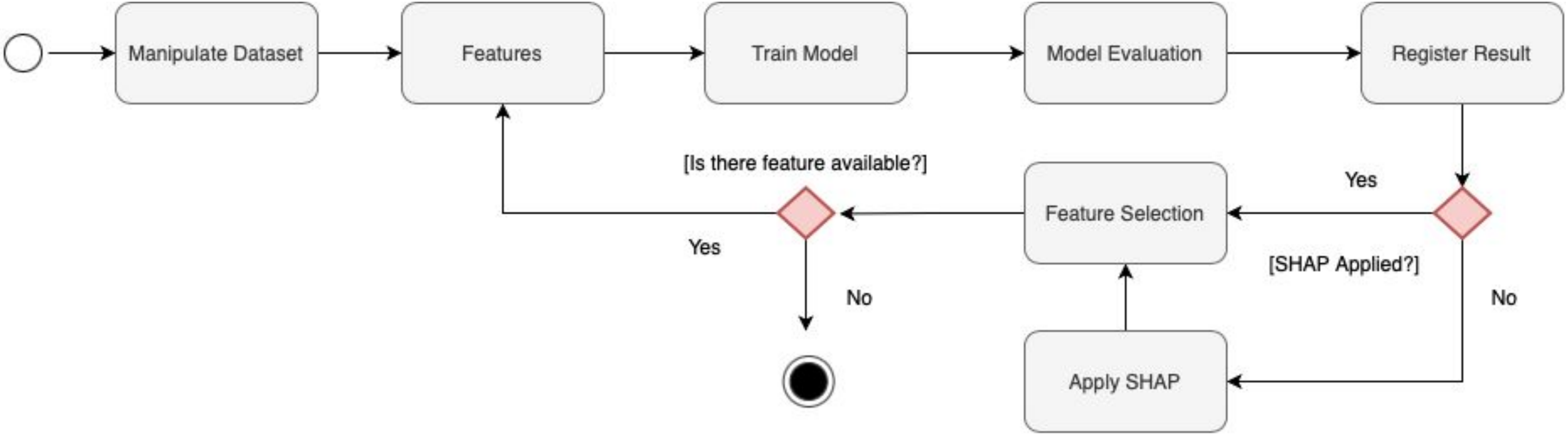
Data Set Characteristics	Multivariate
Attribute Characteristics	Real
Associated Tasks	Classification
Number of Instances	569
Number of Attributes	32
Missing Values	No
Area	Life
Date Donated	01-11-1195
Number of Web Hits	1485620

Dataset: **Credit Card Fraud Dataset**

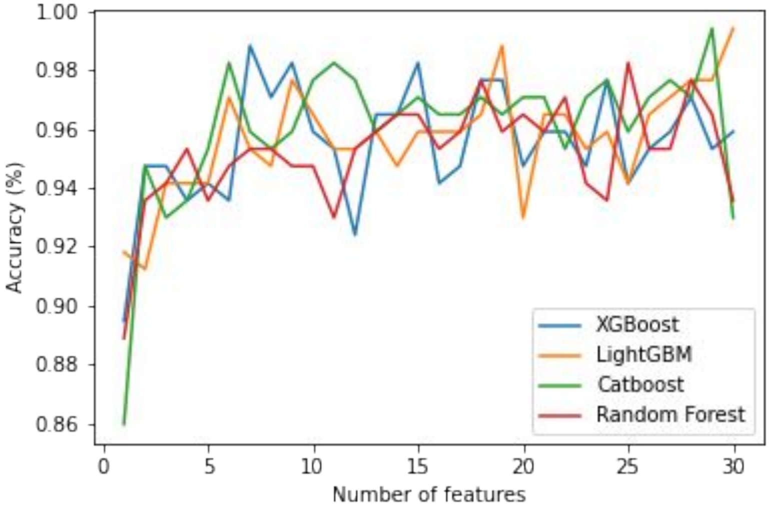


Data Set Characteristics	Multivariate
Attribute Characteristics	Real
Associated Tasks	Classification
Number of Instances	284807
Number of Attributes	31
Missing Values	No
Area	Finance
Date Donated	12-09-2013
Number of Web Hits	N.A.

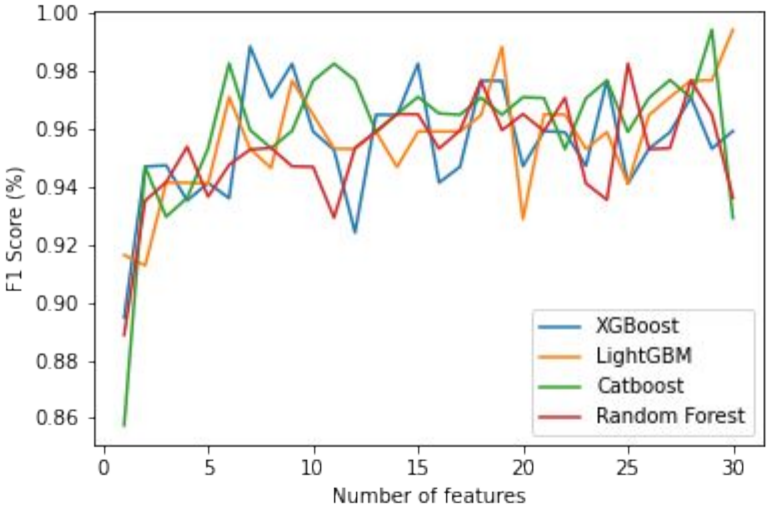
Materials & Methods: **Experiment**



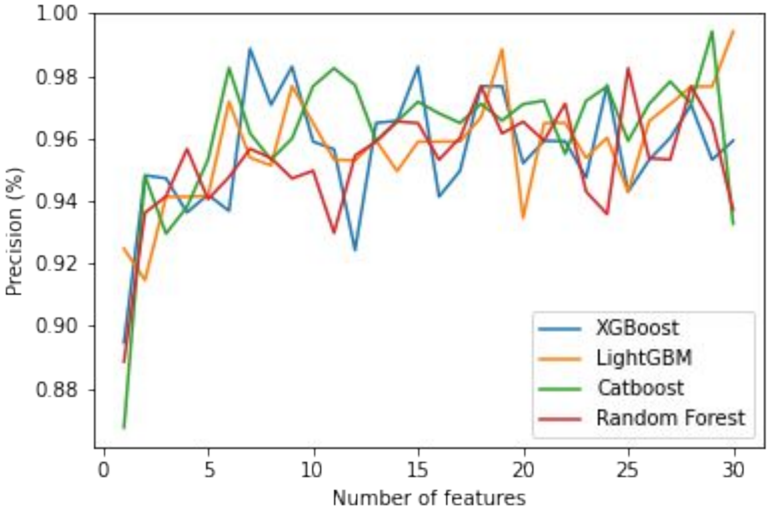
Breast Cancer Dataset - Accuracy



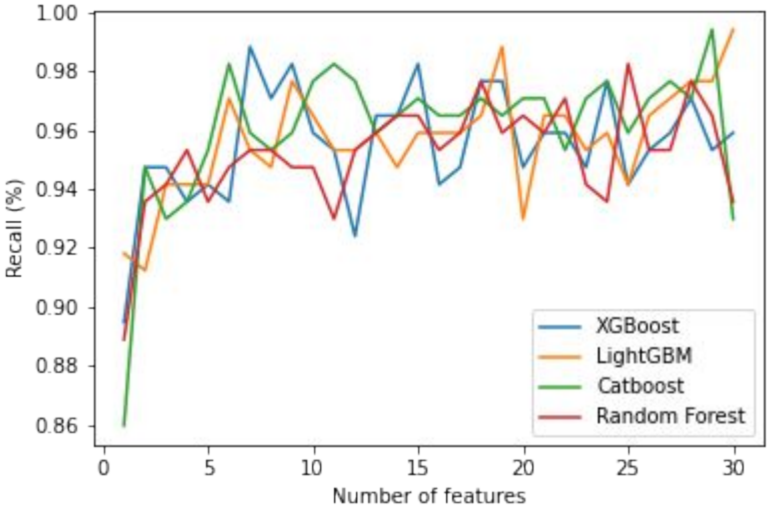
Breast Cancer Dataset - F1 Score



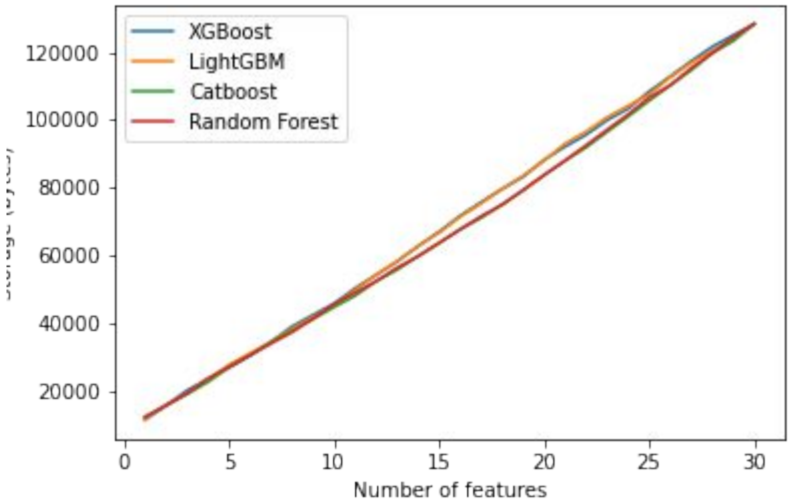
Breast Cancer Dataset - Precision



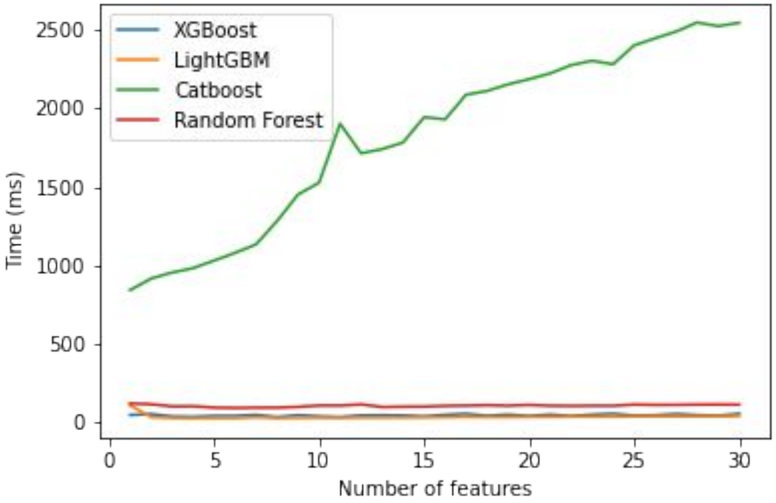
Breast Cancer Dataset - Recall



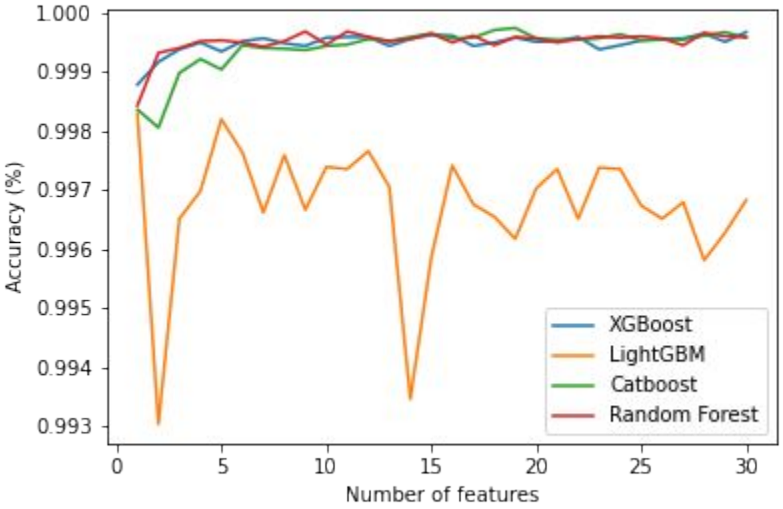
Breast Cancer Dataset - Storage



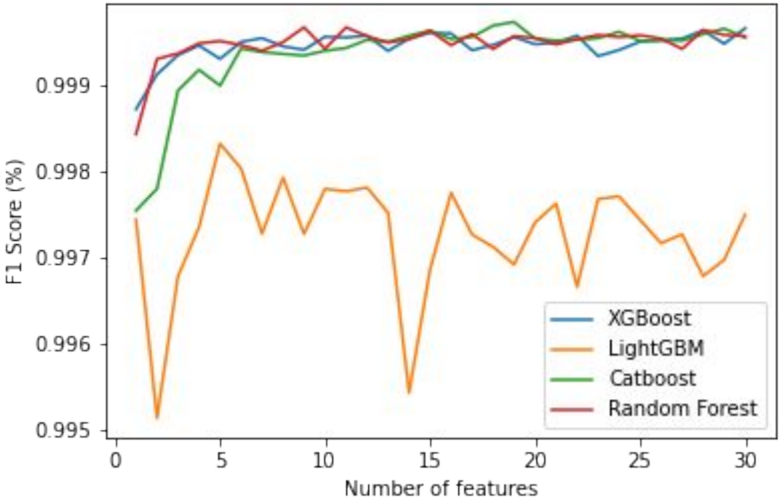
Breast Cancer Dataset - Training Time



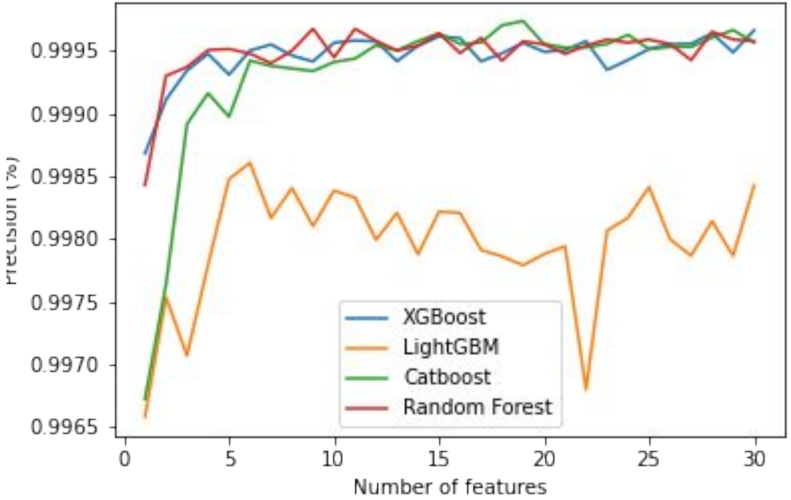
CC Fraud Detection - Accuracy



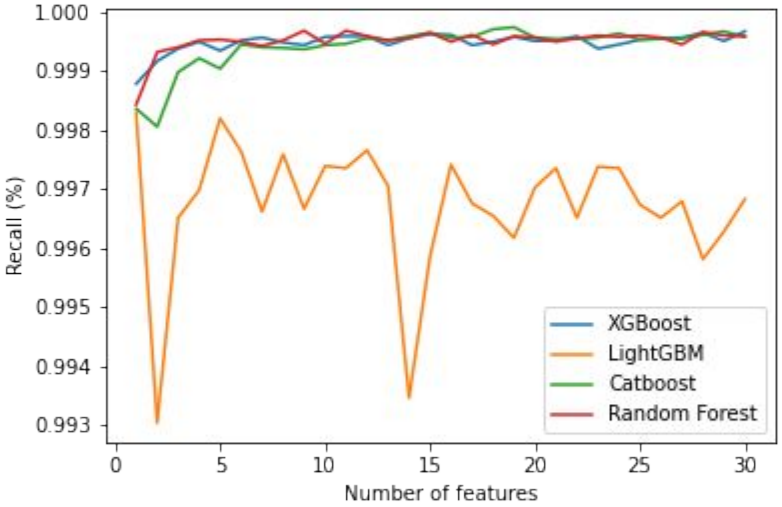
CC Fraud Detection - F1 Score



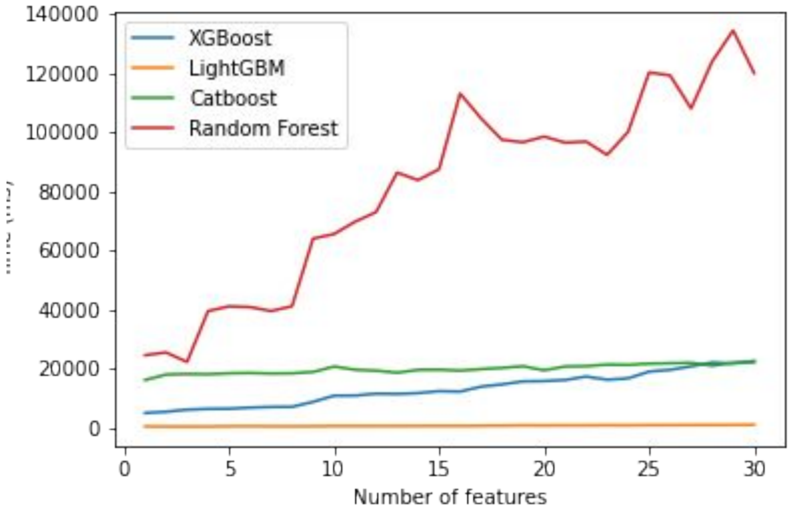
CC Fraud Detection - Precision



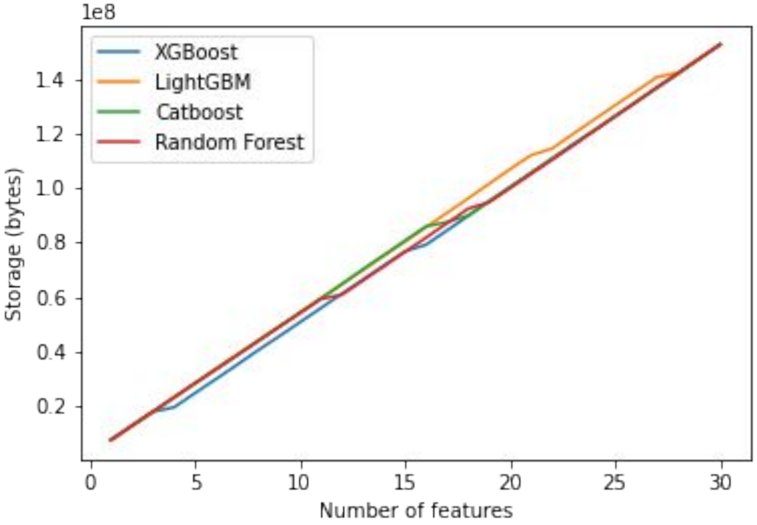
CC Fraud Detection - Recall



CC Fraud Detection - Training Time



CC Fraud Detection - Storage



Conclusion

- SHAP allows to understand how relevant each feature is.
- In some cases, with a small group of features are possible to obtain great results (storage, training time, and performance metrics)
- In future works, studies about how we can development machine learning models based on SHAP could be performed, since SHAP can be used in different ways.

Thank you!

Any questions?