SIS 2022 - The 51ST Scientific Meeting Of The Italian Statistical Society

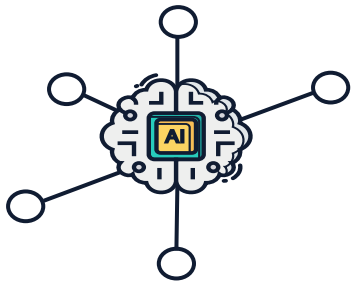
University of Campania "Luigi Vanvitelli", Caserta, Italy, IT
June 22, 2022 – June 24, 2022

eXplainable AI for Finance

Dr. Branka Hadji Misheva
ZHAW, Switzerland

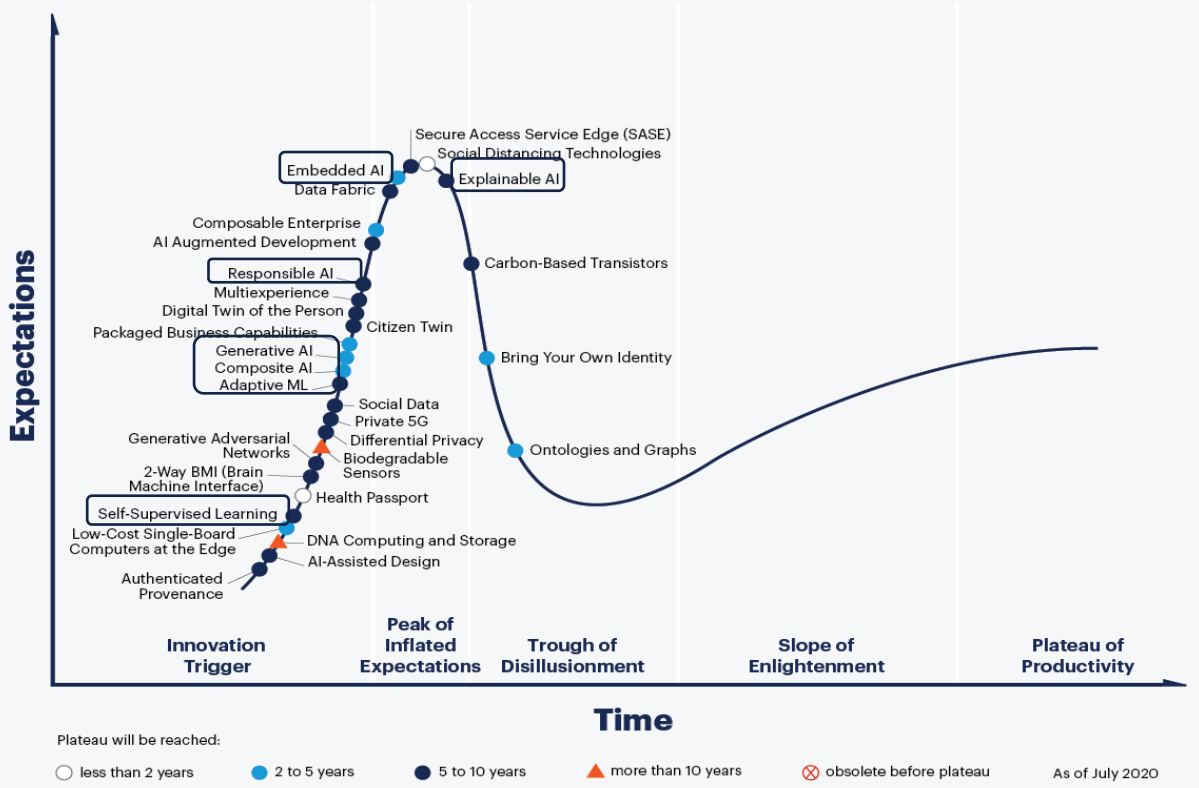

The Need for XAI

Hype vs Real?



AI in finance?

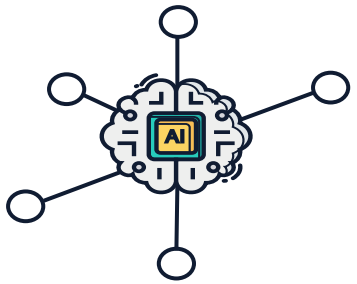
Hype Cycle for Emerging Technologies, 2020



gartner.com/SmarterWithGartner

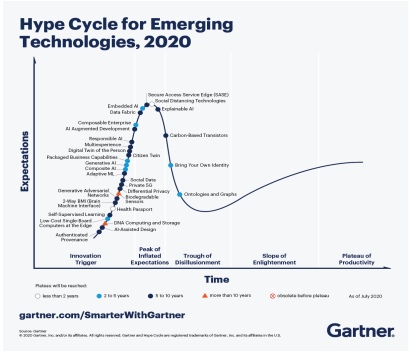
Gartner

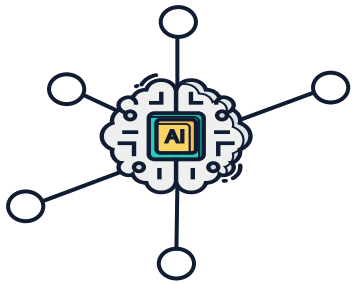
Hype vs Real?



AI in finance?

Hype vs Real?



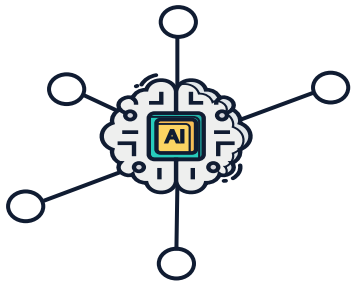


AI in finance?

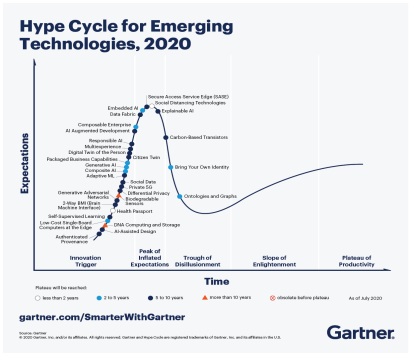
VOLUMES OF DATA

QUANTITATIVE PROBLEMS

Hype vs Real?



AI in finance?

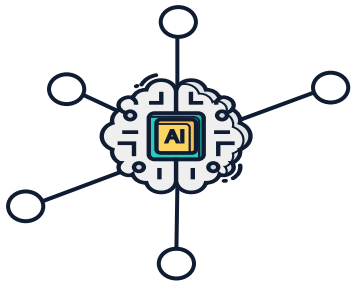


VOLUMES OF DATA

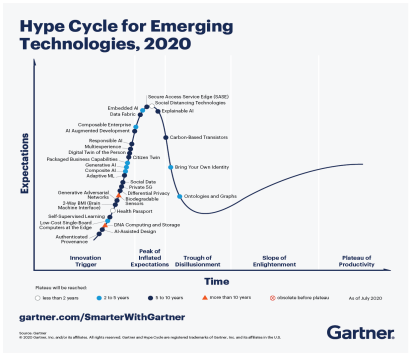
QUANTITATIVE PROBLEMS

Where is the progress?

Hype vs Real?



AI in finance?



VOLUMES OF DATA

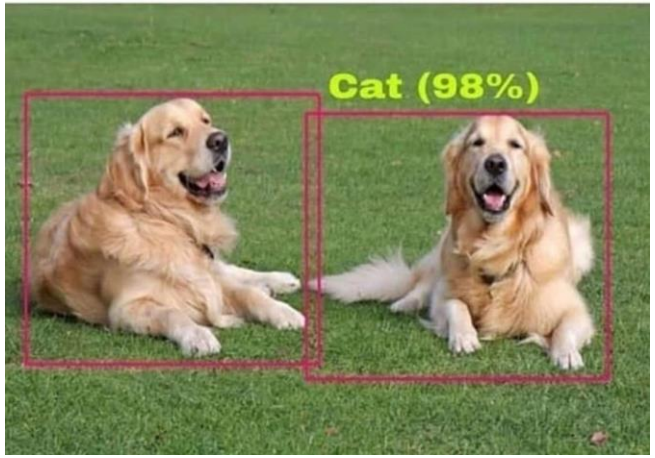
QUANTITATIVE PROBLEMS

Where is the progress?

Well …

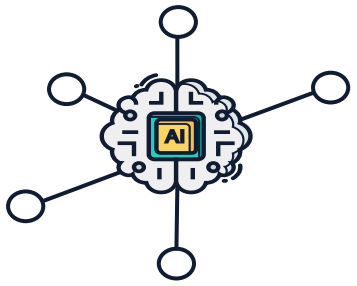
People: *fearing*AI takeover

AI:

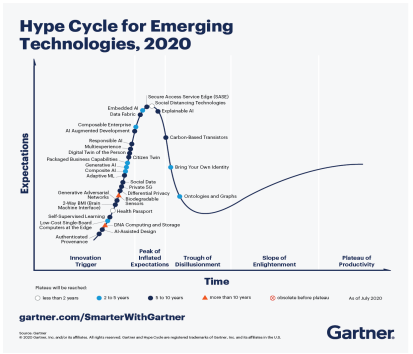


AI in practice is difficult

Hype vs Real?



AI in finance?

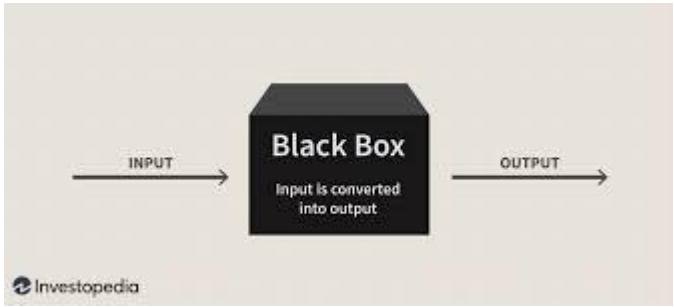


VOLUMES OF DATA

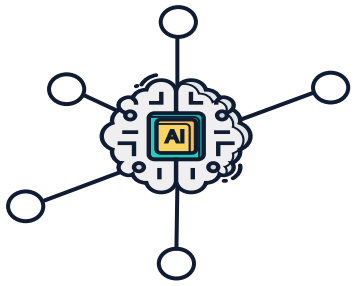
QUANTITATIVE PROBLEMS

Where is the progress?

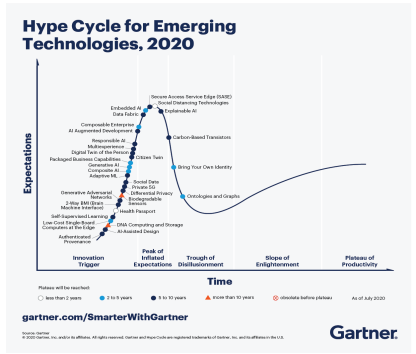
Well …



Hype vs Real?



AI in finance?

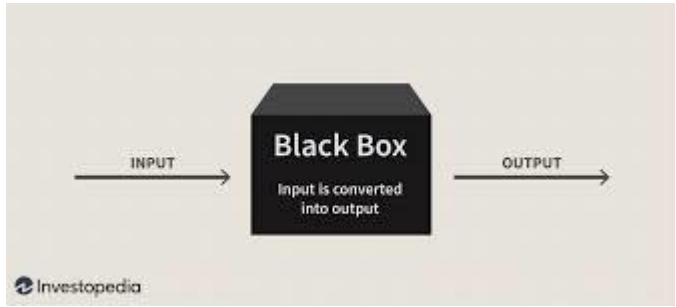


VOLUMES OF DATA

QUANTITATIVE PROBLEMS

Where is the progress?

Well …



Hype vs Real?

AI in finance?

Well …

Explainability is the name of the game!


Deploying
Explainability

POST-HOC Explainability

- There are many models that are highly interpretable.
- Regardless, for some ML models, **post-hoc explaianbility is required**!
- Post-hoc explainability techniques → understandable information about how an already developed model produces its predictions for any given input!
- Methods are considered in view of two main criteria (Linardatos et al. 2021):
  - the **type of algorithm** on which they can be applied (model-specic vs. model-agnostic)
  - **the unit being explained** (if the method provides an explanation which is instance-specic then this is a local explainability technique and if the method attempt to explain the behavior of the entire model, then this is a global explainability technique).

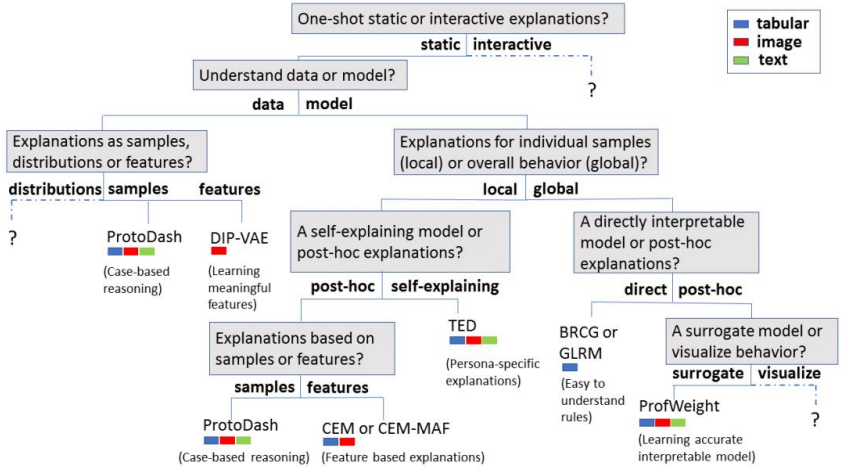


Figure. Arya et al. (2019) proposed taxonomy based on questions about what is explained, how it is explained and at what level

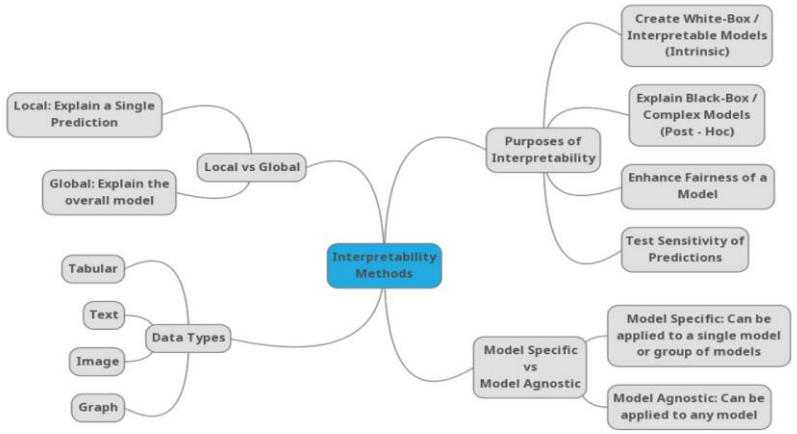


Figure. Linardatos et al. (2021) taxonomy mind-map of Machine Learning Interpretability Techniques.

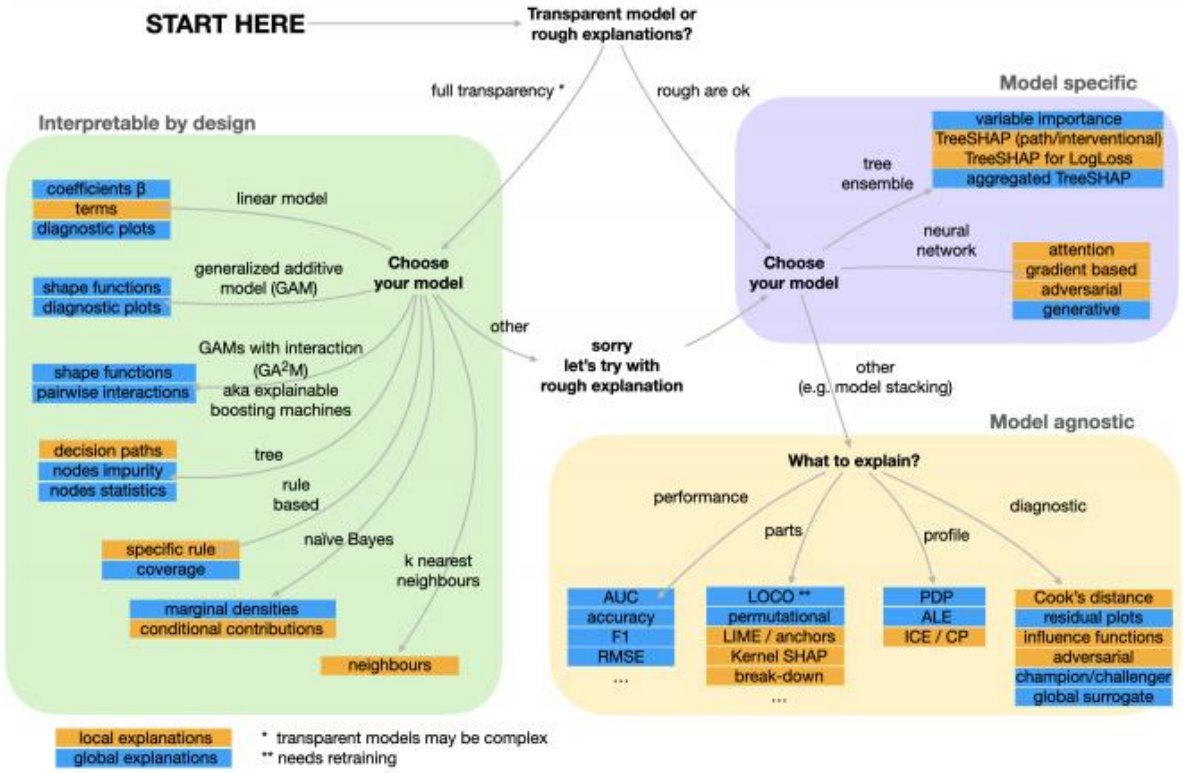


Figure. Maksymiuk et al. (2021) model-oriented taxonomy for XAI method

LIME & SHAPLEY: Details

LIME

- The explanation provided by LIME for each observation:

$$\xi(x) = argmin_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

where G is the class of potentially interpretable models (i.e. linear models)

$g \in G$: An explanation considered as a model

$f: \mathbb{R}^d \rightarrow \mathbb{R}$: The main classifier being explained

$\pi_x(z)$: The proximity measure of an instance z from x

- The goal is to **minimize the locality aware loss** $\boldsymbol{L}$ without making any assumptions about f, since a key property of LIME is that it is model agnostic.
- L is the measure of how unfaithful g is in approximating f in the locality defined by π_x.

SHAPLEY

- Given a model

$$f(x_1, x_2, x_3 \dots x_n)$$

with feature 1 to n being payers in a game in which the payoff v is the measure of importance of the subset.

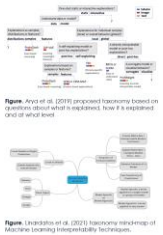
- Marginal contribution $\Delta_v(i, S)$ of a feature i:

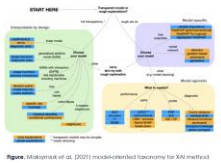
$$\Delta_v(i, S) = \mathrm{v}(\mathrm{S} \cup i) - v(S)$$

- Let $\prod$ be the set of permutations of the integers up to N, and given $\pi \in \prod$ let $S_{i,\pi} = \{j: \pi(j) < \pi(i)\}$ are the players preceding payer i in π, then:

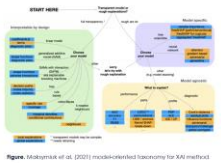
$$\Phi_v(i) = \frac{1}{N!} \sum_{\pi \in \prod} \Delta_v(i, S_{i,\pi})$$


XAI in Finance



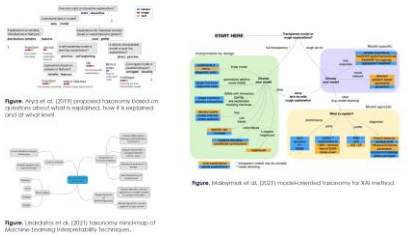


XAI Research



Match explainability needs of stakeholders with the XAI methods

XAI Research

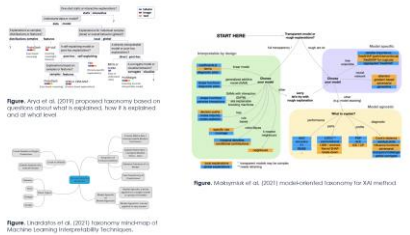


Match explainability needs of stakeholders with the XAI methods

Performance of XAI methods in view of the unique features of financial data

XAI Research

XAI research in **FINANCE**



Match explainability needs of stakeholders with the XAI methods

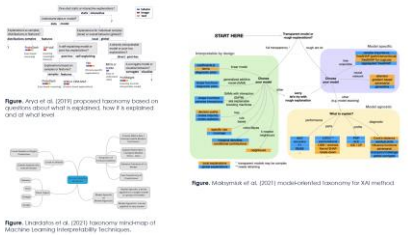
Performance of XAI methods in view of the unique features of financial data

How stable and robust are the explanations provided?

XAI Research

XAI research in **FINANCE**

Deployment



Match explainability needs of stakeholders with the XAI methods

Performance of XAI methods in view of the unique features of financial data

How stable and robust are the explanations provided?

Wider adoption of AI-based use cases in finance

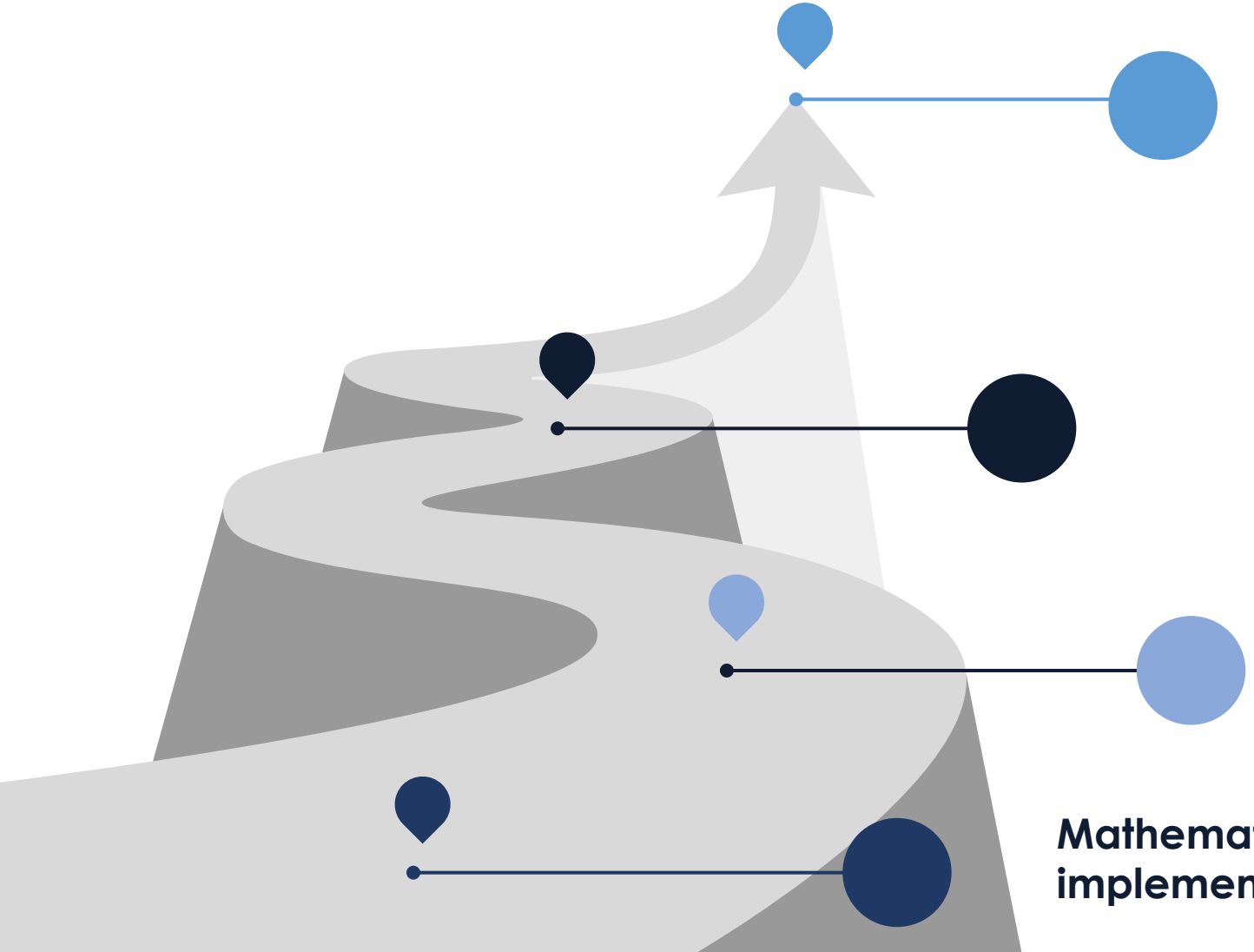
XAI Research

XAI research in **FINANCE**

Deployment

Productivity

Use Case: **XAI IN FINANCE**



To explore the **utility of classical XAI frameworks in the context of financial applications**

Applicability for time series data

Stability and robustness of explanations

Mathematical & implementation issues

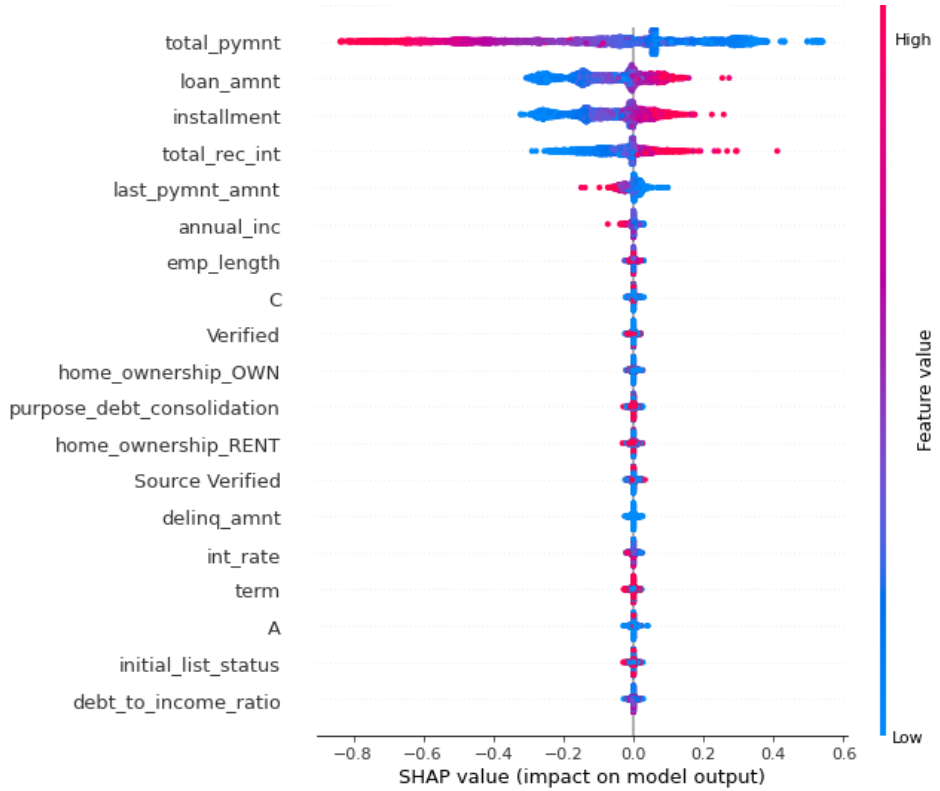
Data & Models: **Credit Risk Use Case**

- 2GB of data and containing information [160 features] on **2.2 million loan contracts**
- Processing:
  - In order to deal with the missing values, in the first instance, all columns which had "NaN" values in more then 90% of the records, were cancelled.
  - Highly correlated features were also eliminated from the input space
  - One hot encoding and combining levels
  - Balanced target
  - Boruta algo

Table 1. Performance

Model	Parameter Space	Performance on Test Data
Logistic Regression	penalty='l2' solver='lbfgs'	Accuracy: 0.9978 , Precision: 0.9960 Recall: 0.9932, F1 score: 0.9946
XGBOOST	scoring = 'roc_auc', cv = 5, n_jobs = -1, verbose = 3, n_estimators = 100, max_depths = 4	Accuracy: 0.9971 , Precision: 1.00 Recall: 0.97, F1 score: 0.99
Random Forest	n_estimators: 500, max_depth: 20	Accuracy: 0.9932, Precision: 1.00 Recall: 0.96, F1 score: 0.98
SVM	gamma='auto', C=1.0, kernel='rbf', probability=False/True	Accuracy: 0.99487, Precision: 1.00 Recall: 0.96, F1 score: 0.98
Neural Networks	n_hidden = 2, neurons = [35,35], activations = RelU, sigmoid loss = binary_crossentropy , Optimizer = adam	Accuracy: 0.9998, Precision: 0.9999 Recall: 0.9985, F1 score: 0.9992

Overall feature importance



Most important features that drive the output are:

1) Total payments
2) Loan amount
3) Instalments
4) …

Figure. SHAP value for RF Classifier [2000 loan contracts, TreeExplainer]

TECHNICAL Issues

- **SHAP** depends on background datasets to infer a baseline/expected value.
- For large datasets, it is computationally expensive to use the entire dataset and we have to rely on approximations.
  - Sampling from the **marginal distribution** means ignoring the dependence structure between present and absent features.
  - The **conditional value function** induces other difficulties:
    - the exact computation of the Shapley value is computationally intensive - requires knowledge of 2^n different multivariate distributions, hence significant amount of approximation is needed.
    - Feature selection can be crucial
      - The choice of features that count as players can affect the resulting explanations

SHAP Kernel

- Sample different coalitions, k: $z'_k = \{0,1\}^m$ where 1=feature is present and 0=feature is absent
- Get the prediction got each coalition
- Compute the weight of each with SHAP kernel
- Fit linear model
- Return SHAPLEY values

→ Feature is absent == feature is replaced by a random feature value from the data

TECHNICAL Issues

- **SHAP** depends on background datasets to infer a baseline/expected value.
- For large datasets, it is computationally expensive to use the entire dataset and we have to rely on approximations.
  - Sampling from the **marginal distribution** means ignoring the dependence structure between present and absent features.
  - The **conditional value function** induces other difficulties:
    - the exact computation of the Shapley value is computationally intensive - requires knowledge of 2^n different multivariate distributions, hence significant amount of approximation is needed.
    - Feature selection can be crucial
      - The choice of features that count as players can affect the resulting explanations

Feature selection

Consider adding a third redundant (C) feature to dataset of two features (A and B).

$$E[f(X)|[X_i, X_e] = E[f(X)|[X_i] = E[f(X)|[X_e]$$

$$E\left[f(X)|[X_p, X_i, X_e\right] = E\left[f(X)|[X_p, X_i\right] = E\left[f(X)|[X_p, X_e\right]$$

Therefore, for any data instance x:

$$\phi_v(A) = \frac{1}{3}\triangle_v(A, \emptyset) + \frac{2}{3}\triangle_v(A, BC)$$

$$\phi_v(B) = \phi_v(C) = \frac{1}{3}\triangle_v(B, \emptyset) + \frac{1}{6}\triangle_v(B, A)$$

Now let's consider a case with two features:

$$\phi'_v(A) = \frac{1}{2}\triangle_v(A, \emptyset) + \frac{1}{2}\triangle_v(A, B)$$

$$\phi'_v(B) = \frac{1}{2}\triangle_v(B, \emptyset) + \frac{1}{2}\triangle_v(B, A)$$

Notice

$$\phi_v(B) \neq \phi'_v(B)$$

The relative apparent importance of A and B thus depend on whether C is added as a third feature.

Kumar et al. 2020

Stability & Robustness of Explanations though **GRAPH THEORY**

- **Hypothesis:** Similar data points/loan contracts should have similar explanations.

- **Two approaches:**
  - We can see whether similar loan contracts have the same top explanatory features as per the specific XAI method
  - We can investigate the dependence between the spatial distance and the explanation distance between loan contracts

- **Graph theory** applicability:
  - **Visualization**
  - **Modelling:** we can define a **special distance** between loan contracts using different distance measures (e.g. Euclidian, HEOM distance etc.) and we can estimate the relationship between the spatial and **explanation distance**.

Stability & Robustness of Explanations: **Spatial Distance of Loans**

- We exploit information derived from the numerical features collected in a vector x_n representing the different loan contacts n.

- We define a metric **D - standardized Euclidean distance** between each pair $(x_j; y_j)$ loan feature vectors:

$$D_{x,y} = \sqrt{\sum_{j=1}^{J} \left(\frac{x_j}{s_j} - \frac{y}{s_j} \right)^2}$$

- **Visualization**: For a **Graph G**, the goal is to find a tree T which is a spanning subgraph of G, i.e. every node is included to at least one edge of T and has minimum total weight.
  - Pick some arbitrary start node u. Initialize $T = u$
  - At each step add the lowest-weight edge to T (the lowest-weight edge that has exactly one node in T and one node not in T);
  - Stop when T spans all the nodes.

Stability & Robustness of Explanations: **Explanation Distance of Loans**

- We investigate the dependence between the special distance and the explanation difference between similar contracts.
- For this purpose, we define an explanatory distance measure:

$$expDis = \sqrt{\underbrace{\sum_{z=1}^{10}(SHAP_{n_{zx_i}} - SHAP_{n_{zx_j}})^2}_{for\ n_{x_i} \cap n_{x_j}} + \underbrace{\sum_{z=1}^{n}(SHAP_{n_{zx_i}} + SHAP_{n_{zx_j}})^2}_{for\ n_{x_i}\ \emptyset\ n_{x_j}}}$$

where,

- n are the top 10 features provided by XAI method
- x_i and x_i are the different pairs of loan feature vectors
- SHAP are the specific SHAP contributions

Visualization: Stability of Explanations

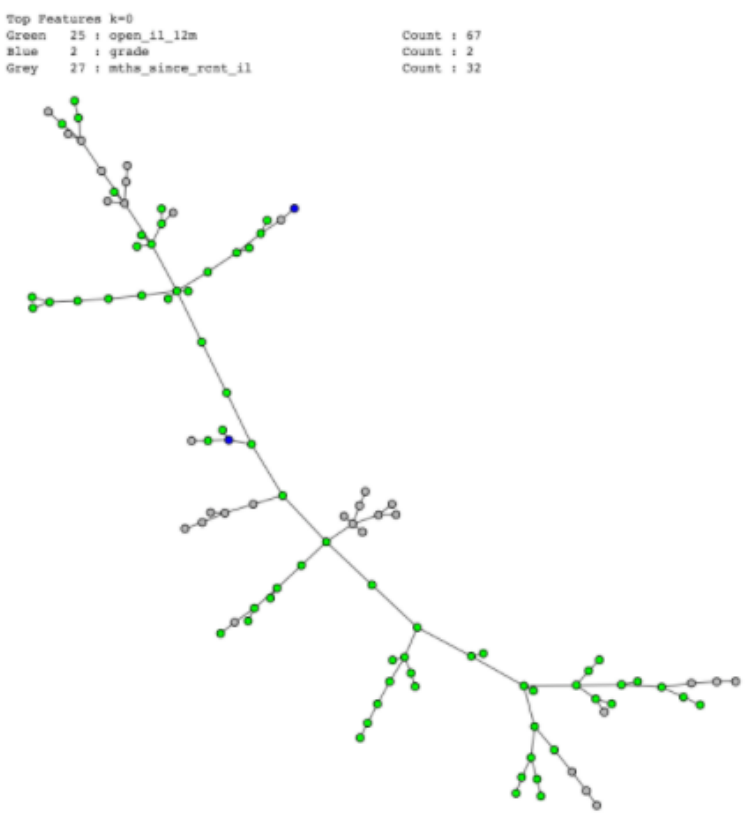


Figure 1. MST tree representation of 100 random data points. Coloring based on the top explanatory feature [green = "Number of instalment accounts opened in past 12 months"; grey = "Months since most recent instalment accounts opened" ; blue = "Grade"]

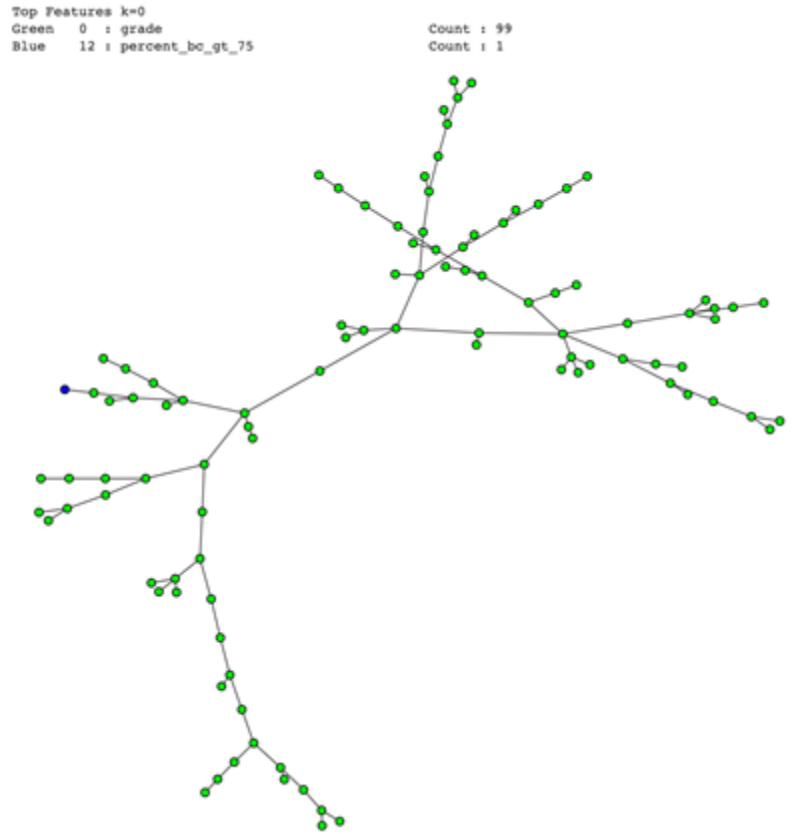


Figure 2. MST tree representation of 100 random data points. Coloring based on the top explanatory feature [green = "Grade", blue = "Percent of trades never delinquent"]

Stability of Explanations though **GRAPH THEORY**

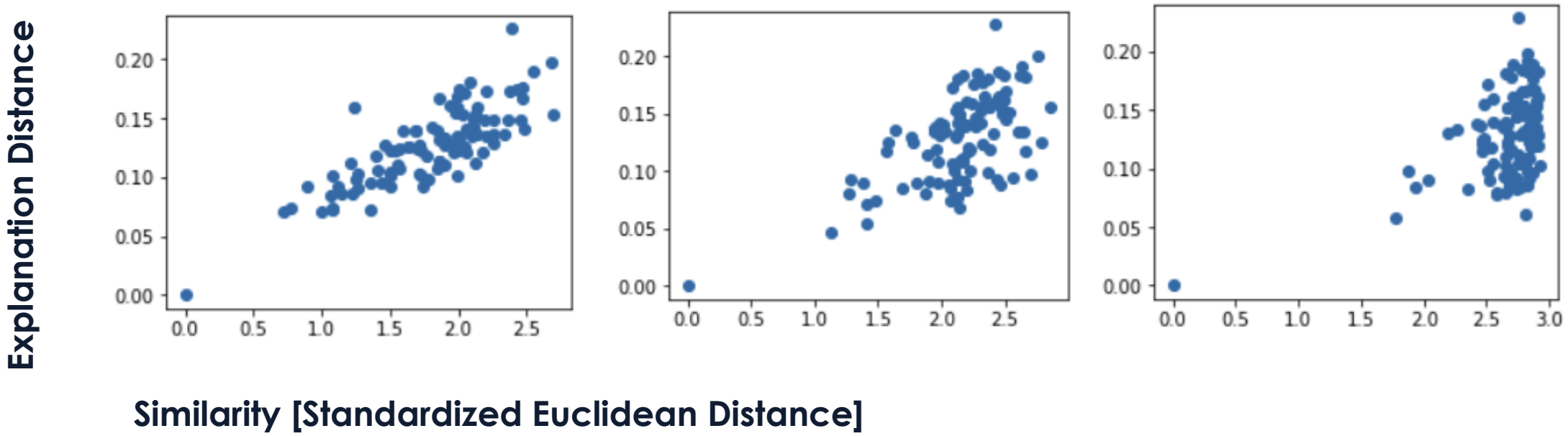


Figure 3. Explanation Distance vs Spatial Distance for ref_i = 1000, n = 100 for 5, 10, and 20 Features.

*Remember: the explanation difference formula takes the top n features of two points, adds up the squared difference of the contributions of each feature in common, and for each feature that is not common, adds up the square of each contribution then finally take the square root of the sum.

Applicability for **Time Series Data**

- Classical approaches and their current implementation are not tailored for **financial time series** (subject to trends, vola-clusters,...).

- Specifically, **perturbation-based methods are fully dependent on the ability to perturb samples in a meaningful way**. In the context of financial data:

    - if features are correlated, the artificial coalitions created will lie outside of the multivariate joint distribution of the data,

    - if the data are independent, coalitions/generated data points can still be meaningless;

    - generating artificial data points through random replacement disregards the time sequence hence producing unrealistic values for the feature of interest.

Applicability for **Time Series Data**

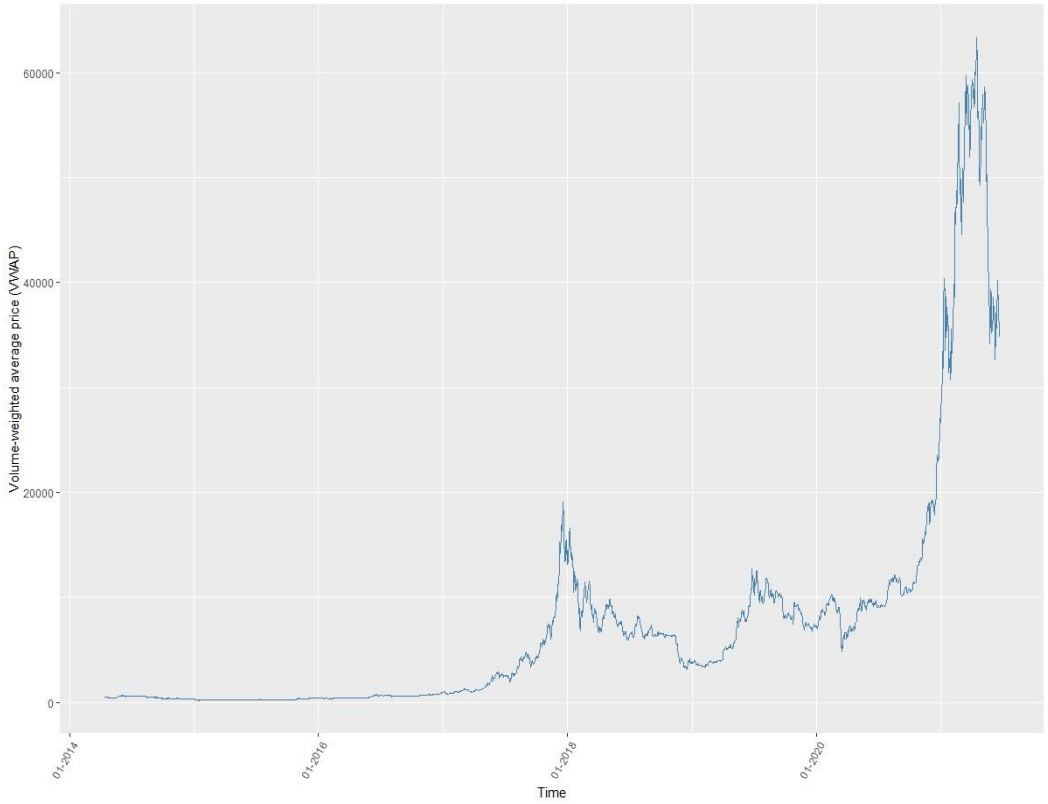


Figure 4. Bitcoin Prices (original time ordering)

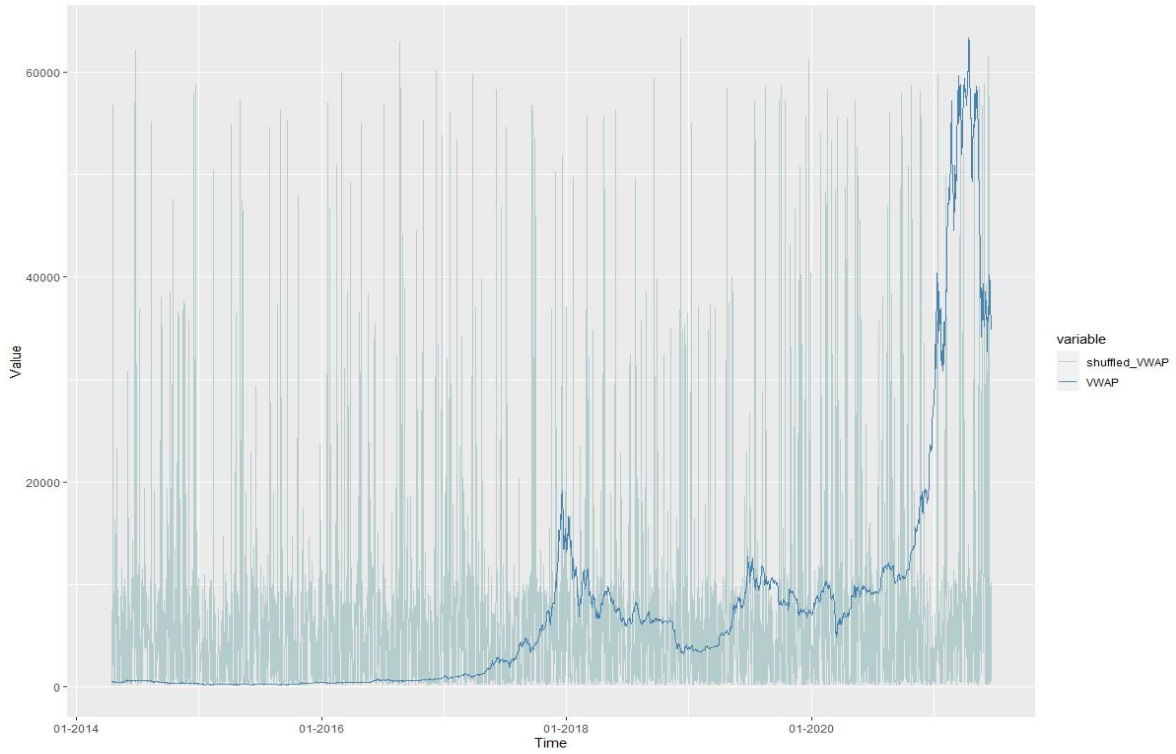


Figure 5. Bitcoin Prices (original time ordering and shuffled values)

Simple X-Function: Identity and LPD

- We propose a family of Explainability (X-)functions xf(.) for assigning meaning to the net's response or output o_t over time t = 1, ..., T, where $o_t = (o_{1t}, \ldots, o_{npt})$ is a n_p dimensional vector of possibly multiple output neurons.
- By selecting the identity $xf(o_t) = o_t$ we can mark preference for the sensitivities or partial derivatives $w_{ijt} = \frac{\partial o_{jt}}{\partial x_{it}}, i = 1, \ldots, n, j = 1, \ldots, n_p$, for each explanatory variable x_{it} of the net.
- In order to complete the 'explanation' derived from the identity one can add a synthetic intercept to each output neuron o_{jt} defined according to:

$$b_{jt} := o_{jt} - \sum_{i=1}^{n} w_{ijt} x_{it}$$

- For each output neuron o_{jt}, the resulting derivatives or 'explanations' $b_t, w_{1t}, w_{2t} \ldots w_{pt}$ generate a new data-flow which is referred to as **Linear Parameter Data**
- The LPD is a matrix of dimension $T * (n + 1)$, irrespective of the complexity of the neural net, with t−th row denoted by $LPD_t := (b_t, w_{1t}, w_{2t} \ldots w_{pt})$
- The LPD can be interpreted in terms of **exact replication of the net by a linear model at each time point t** and the natural time-ordering of LPDjt subsequently allows to examine changes of the linear replication as a function of time.
- We are then in a position to assign a meaning to the neural net, at each time point t = 1, ..., T, and to monitor non-linearities of the net or, by extension, possible non-stationarities of the data

LPDs & **RISK MANAGEMENT**

- Our initial insights suggest that **the dependency of the LPDs can be related with certain movements in terms of the bitcoin price**.

- We here suggest that these episodes are possibly **indicative of unusual activity** that might call for accrued attention and care of investors or regulators

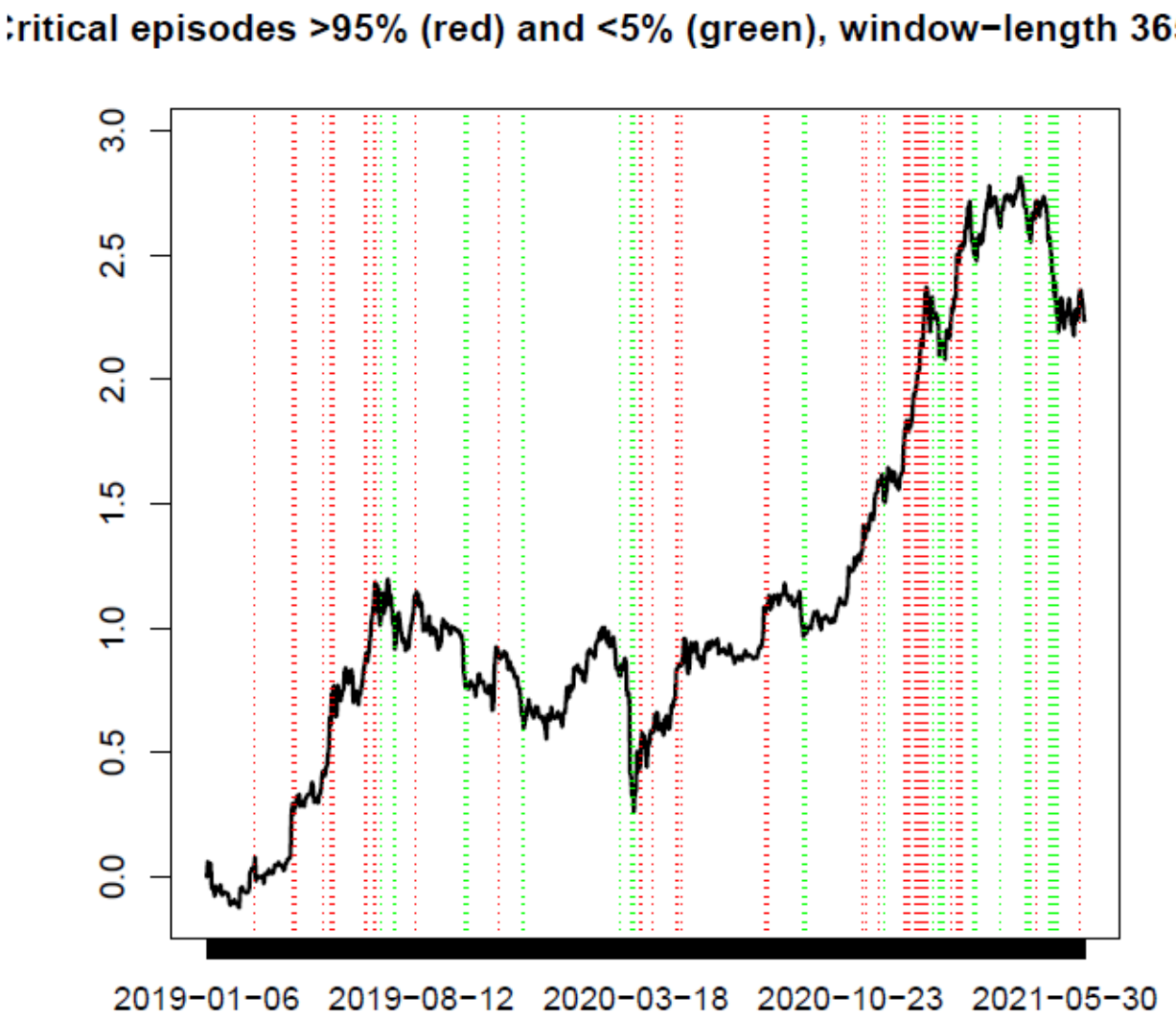


Figure 6. Critical time points: two-sided exceedances of the 5-percent (green) and 95-percent (red) quantiles of the out-of-sample mean-LPD (mean over 100 nets) based on a rolling-window of length one year of its own history. The LPD corresponding to the lag-6 BTC-value is used.

Outlook

- The **lack of algorithmic transparency is one of the main barriers** for the wider adoption of AI-based solutions in credit risk management

- **Three-fold objective** of the work:
  - Identify the mathematical & implementation issues related with classic XAI methods applied in finance
  - explore the stability and robustness of explanations provided
  - Explore the applicability of classical methods for time series data

- **Technical issues** → many challenges

- **Stability** → state-of-art methods offer certain level of stability

- **Applicability for time series data** → we need new methods which preserve the natural ordering of time.