Machine Learning in Finance Workshop 2021

eXplainable AI in Credit Risk Management

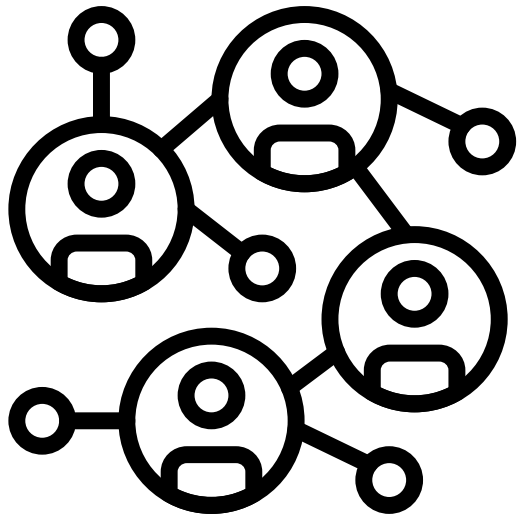
Branka Hadji Misheva

ZHAW Zurich University of Applied Sciences

NOTE: Contributors

Dr. Branka Hadji Misheva &
Prof. Dr. Jörg Osterrieder
ZHAW, School of Engineering

Prof. Dr. Ali Hirsa,
Columbia University



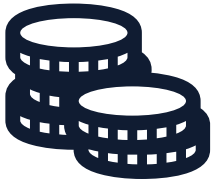
Alex Raita & Phillip Kim,
Columbia University

Onkar Kulkarni & Stephen Fung Lin
Columbia University

AGENDA

The Need for XAI

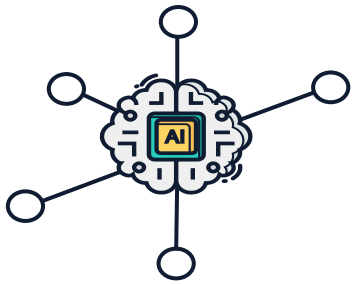
Deploying Explainability



XAI in Credit Risk Management

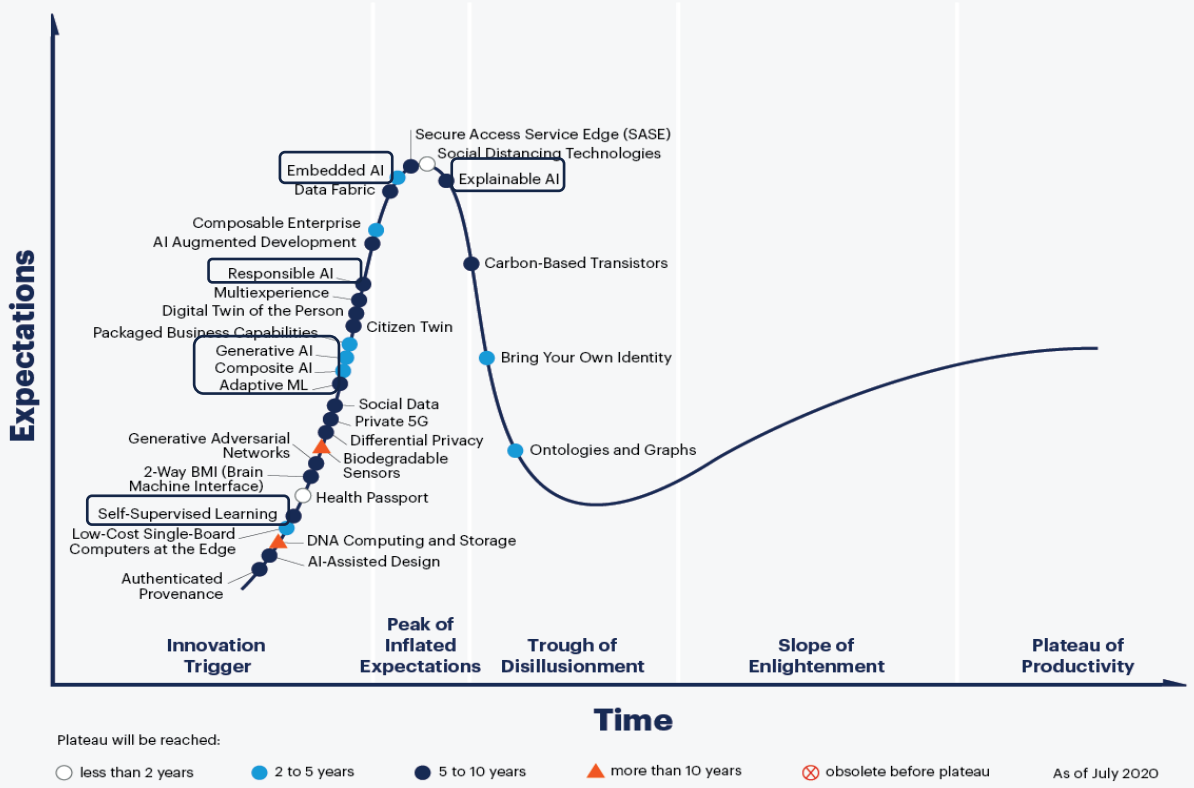

The Need for XAI

Hype vs Real?



AI in finance?

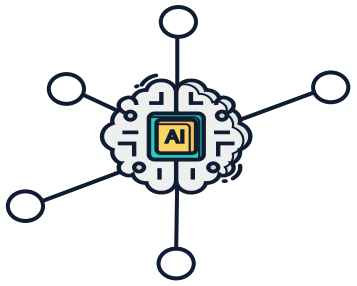
Hype Cycle for Emerging Technologies, 2020



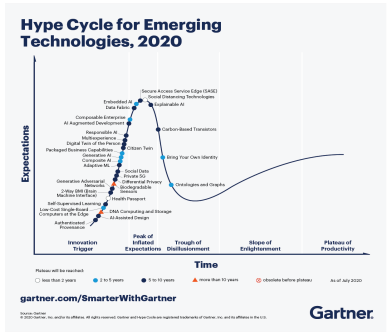
gartner.com/SmarterWithGartner

Gartner

Hype vs Real?



AI in finance?



VOLUMES OF DATA

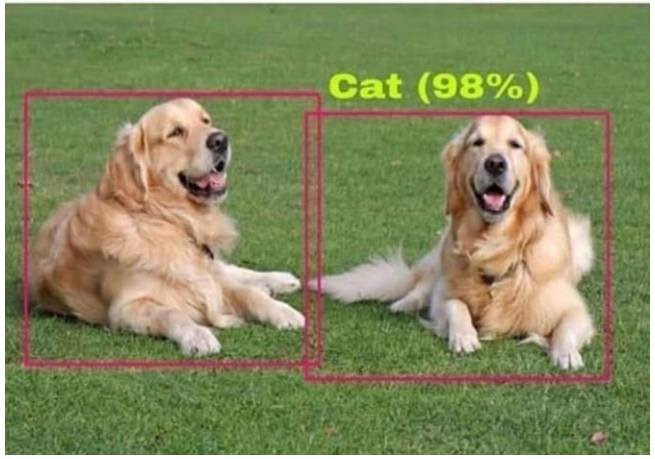
QUANTITATIVE PROBLEMS

Where is the progress?

Well …

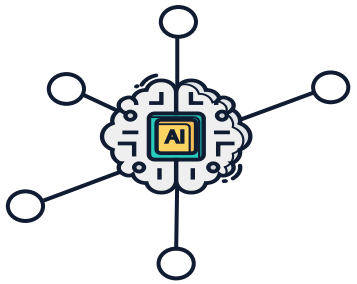
People: *fearing*AI takeover

AI:

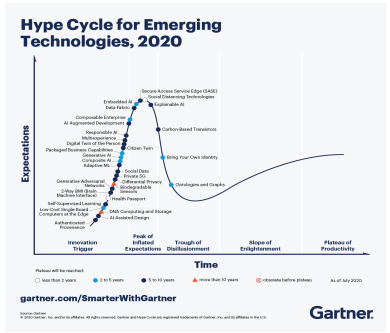


AI in practice is difficult

Hype vs Real?



AI in finance?

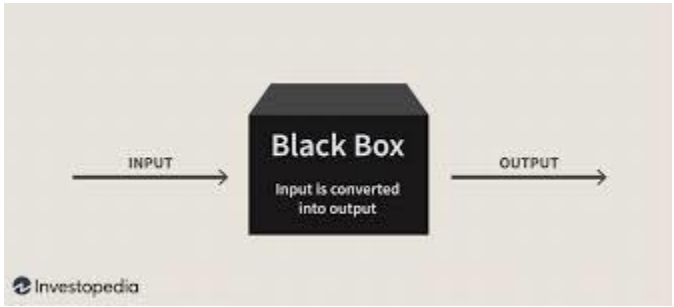


VOLUMES OF DATA

QUANTITATIVE PROBLEMS

Where is the progress?

Well …



Hype vs Real?

AI in finance?

Well …

Explainability is the name of the game!

The Need for **eXplainable AI**

It is not clear how variables are being combined to make predictions!

Why do we **NEED** this?

- Trust in models is **key**!

One Mistake!

Image source: medium.com

Why do we **NEED** this?

It has found some snow!

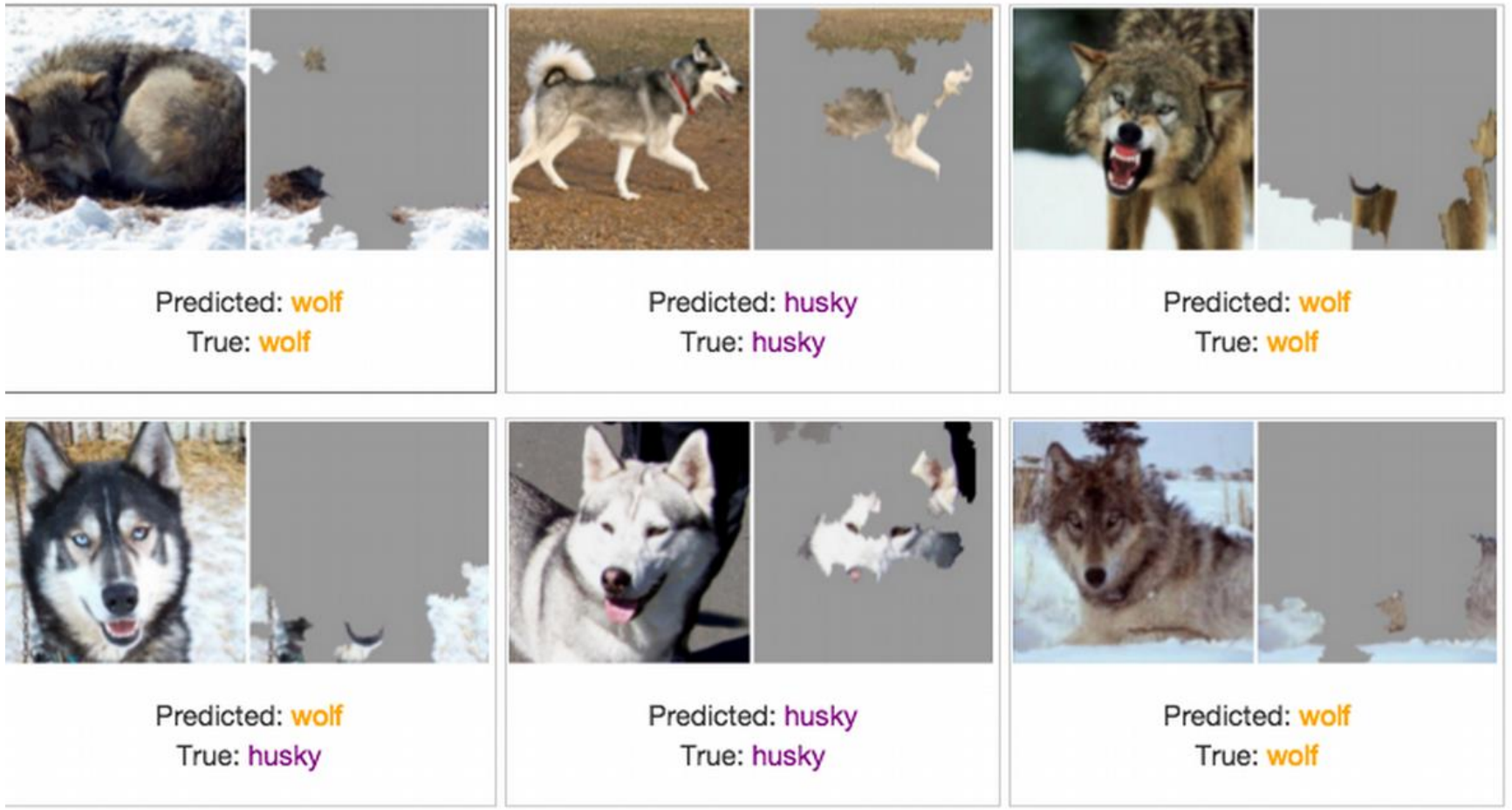
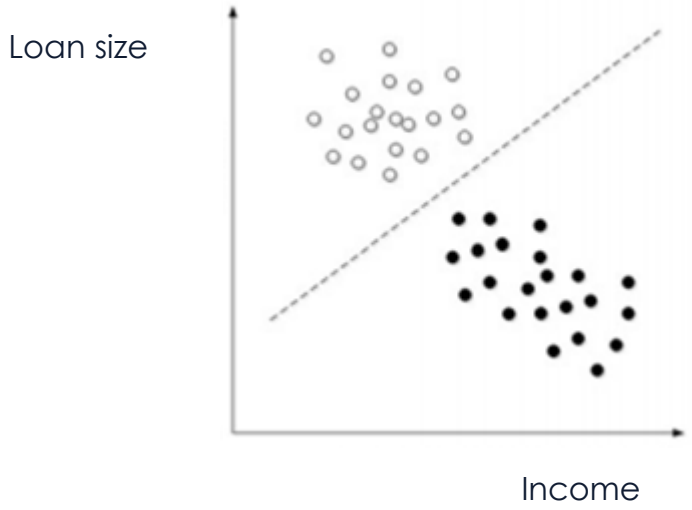


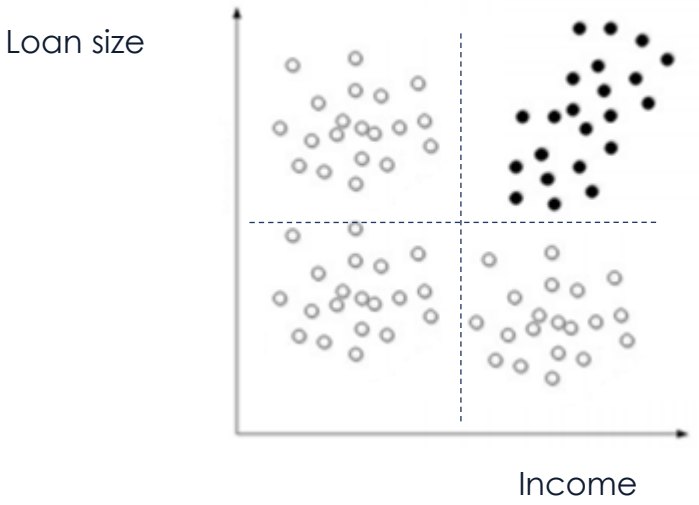
Image source: medium.com


Deploying
Explainability

CREDIT RISK Management

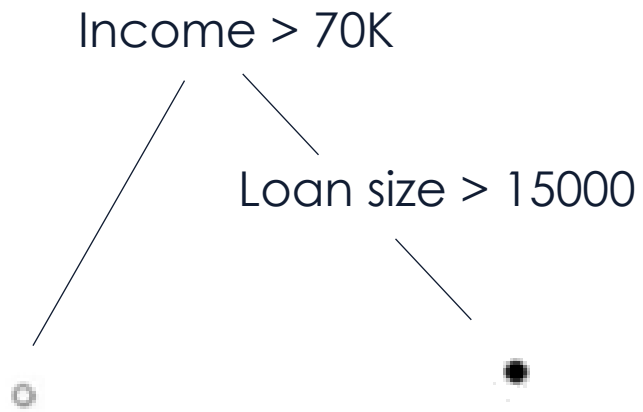


CREDIT RISK Management



What about **non-linear relationships**?

Still interpretable!



N-dimensions and **HIGH COMPLEXITY**

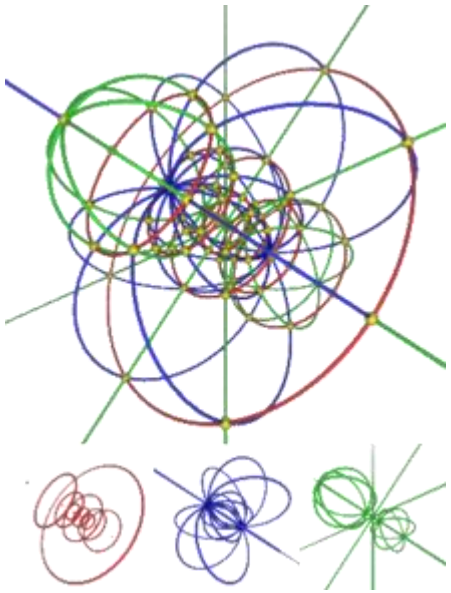


Image source: wikipedia

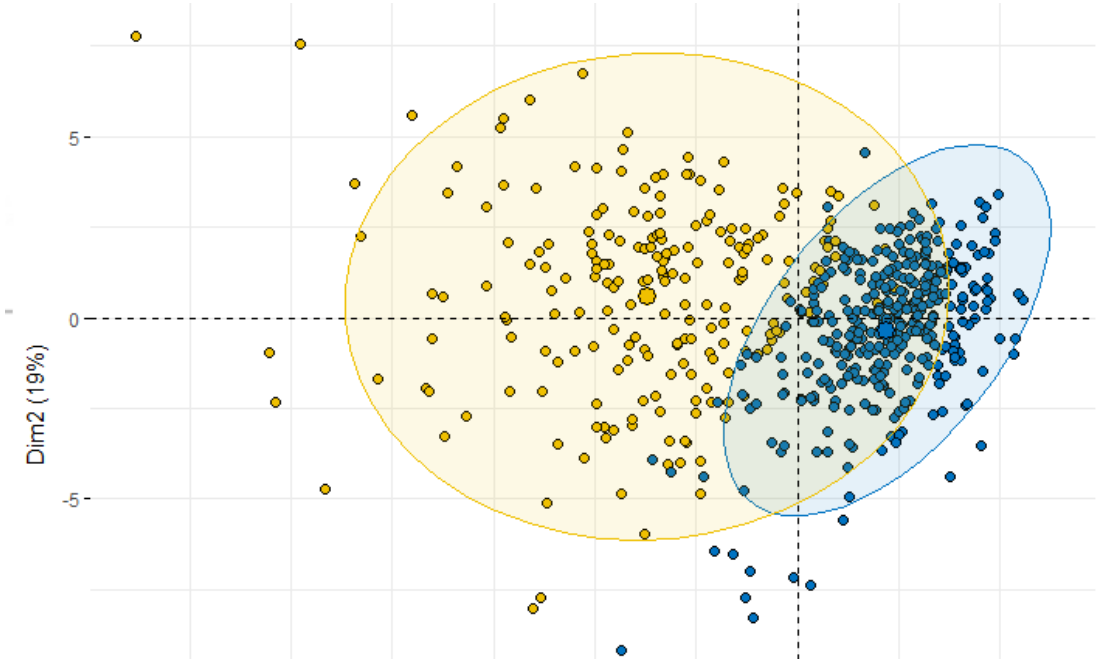
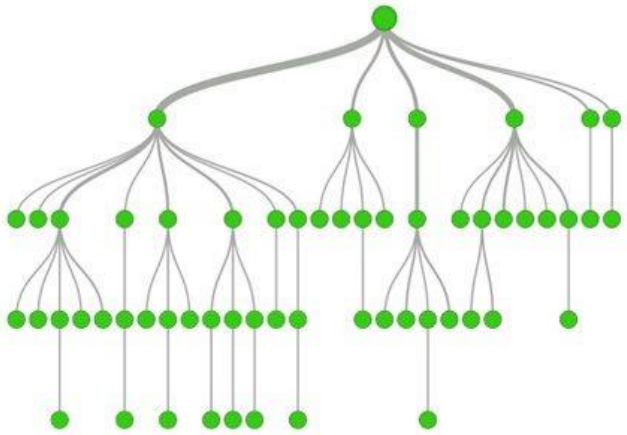
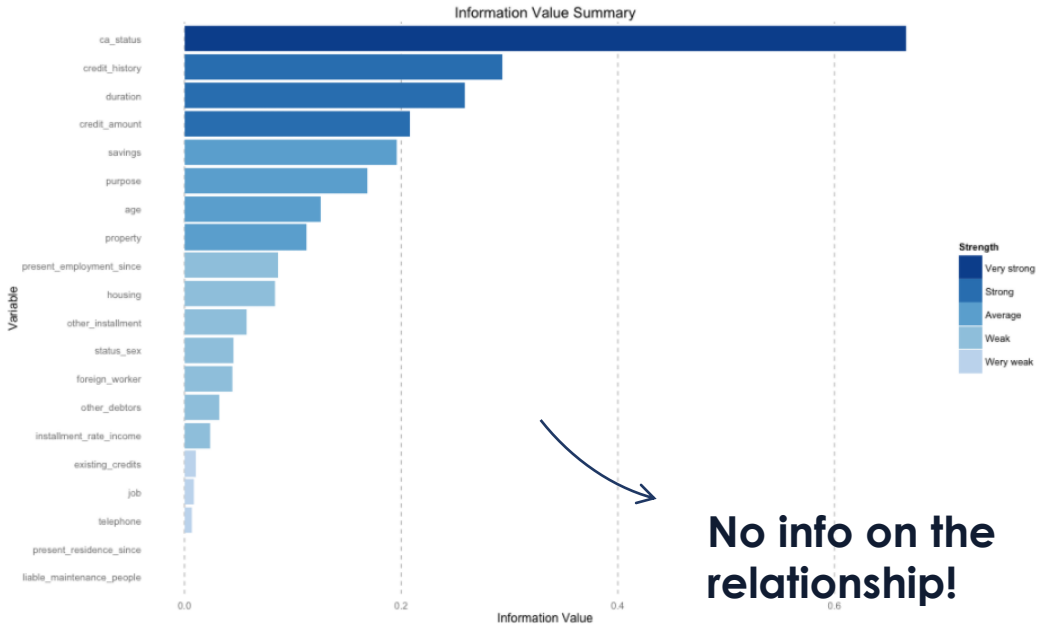


Image source: towardsdatascience.com

FEATURE IMPORTANCE





No info on the relationship!

Image source: opendatascience.com

Image source: stackoverflow.com

POST-HOC Explainability

- 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!

- We distinguish between two approaches:
  - those that are designed for their application to **any ML models;** and
  - those that are designed for **a specific ML model** and thus, can not be directly extrapolated to any other learner.

POST-HOC Explainability

- 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!
- We distinguish between two approaches:
  - **those that are designed for their application to any ML models; and**
  - those that are designed for a specific ML model and thus, can not be directly extrapolated to any other learner.

Local Interpretable Model-agnostic Explanations

LIME → explains the prediction of **any machine learning classifier** by learning an interpretable model **locally** around the prediction.

"Why Should I Trust You?"
Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

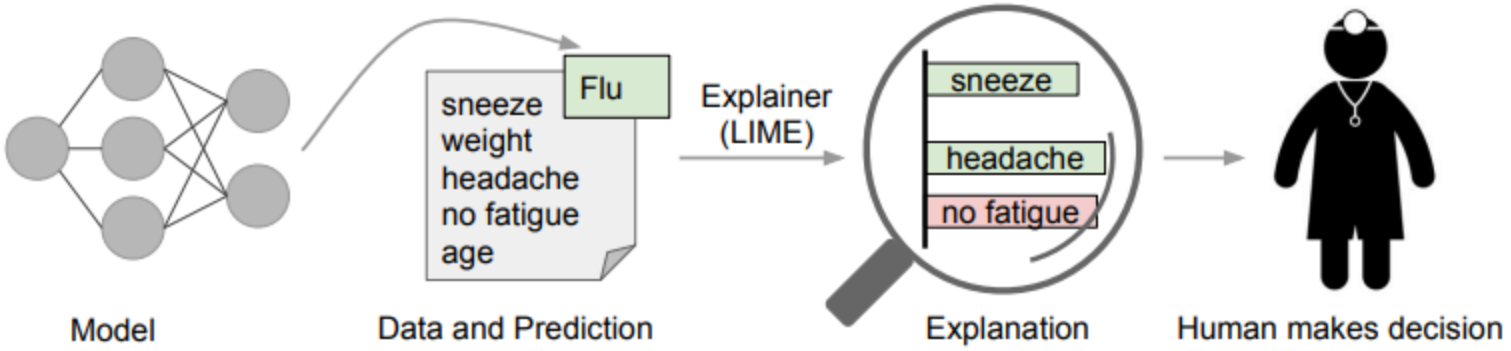


Image source: Ribeiro et al. (2016)

LIME: **Details**

- 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

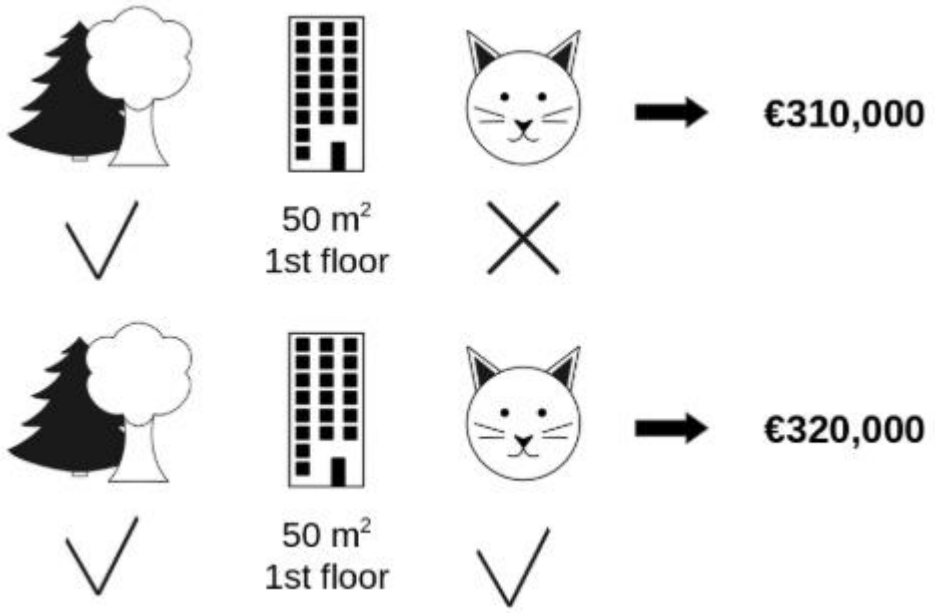
$\Omega(g)$ - Complexity parameter (e.g. number of features)

- The goal is to **minimize the locality aware loss 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 Values

- The Shapley value is the average marginal contribution of a feature value across all possible coalitions.

The contribution of the cat-banned is -10K!



This greatly depends on our random pick!

We repeat the sampling step and average the contribution!

Image source: christophm.github.io

Shapley Values: **DETAILS**

—

- Given a model

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

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

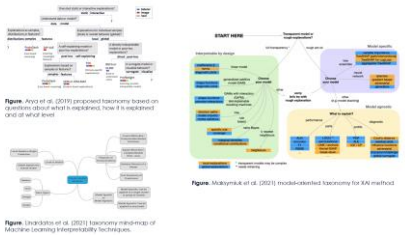
- Marginal contribution $\Delta_v(i, S)$ of a feature i:

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

- 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 Credit Risk Management



Match explainability needs of stakeholders with the XAI methods

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

What is the best way to bring those explanations to different stakeholders in the financial world?

Wider adoption of AI-based use cases in finance

XAI Research

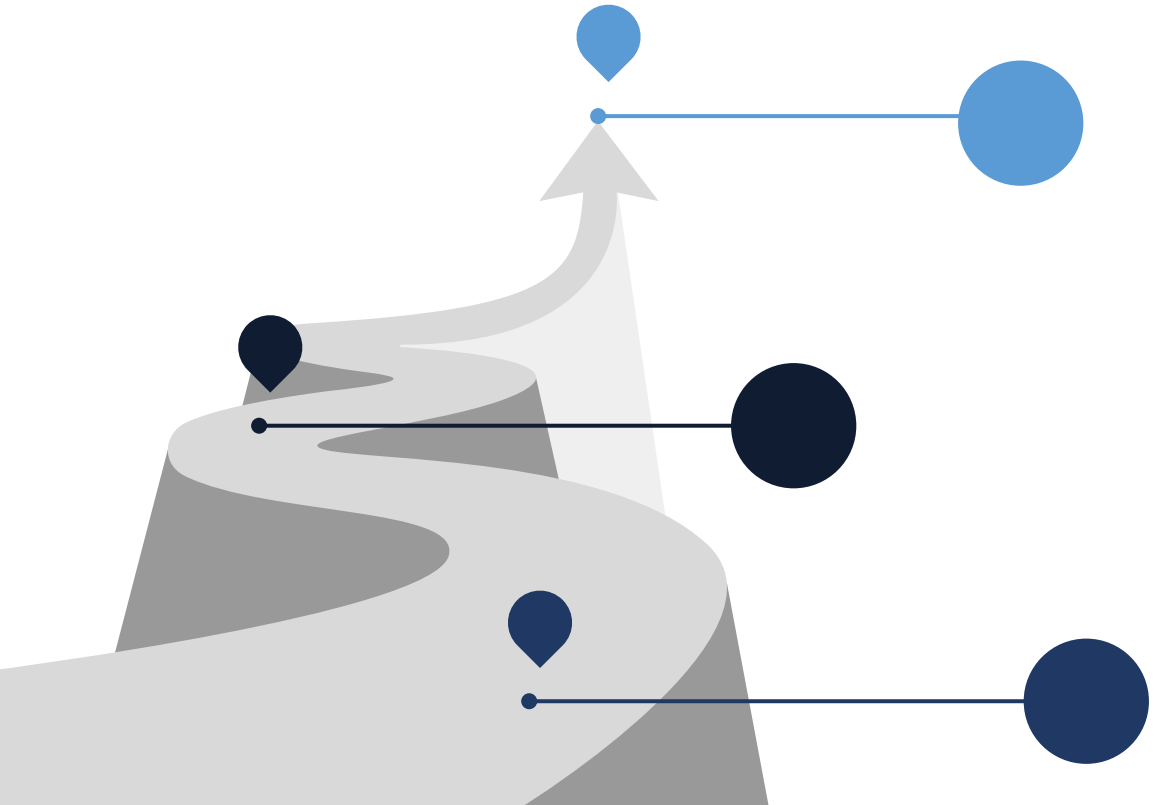
XAI research in **FINANCE**

Deployment

Productivity

Use Case: **OBJECTIVES**

Context: Credit Risk Management



To explore the **utility of both SHAP and LIME frameworks in the context of credit risk management**

Stability and robustness of explanations

Human-centric and mathematical issues

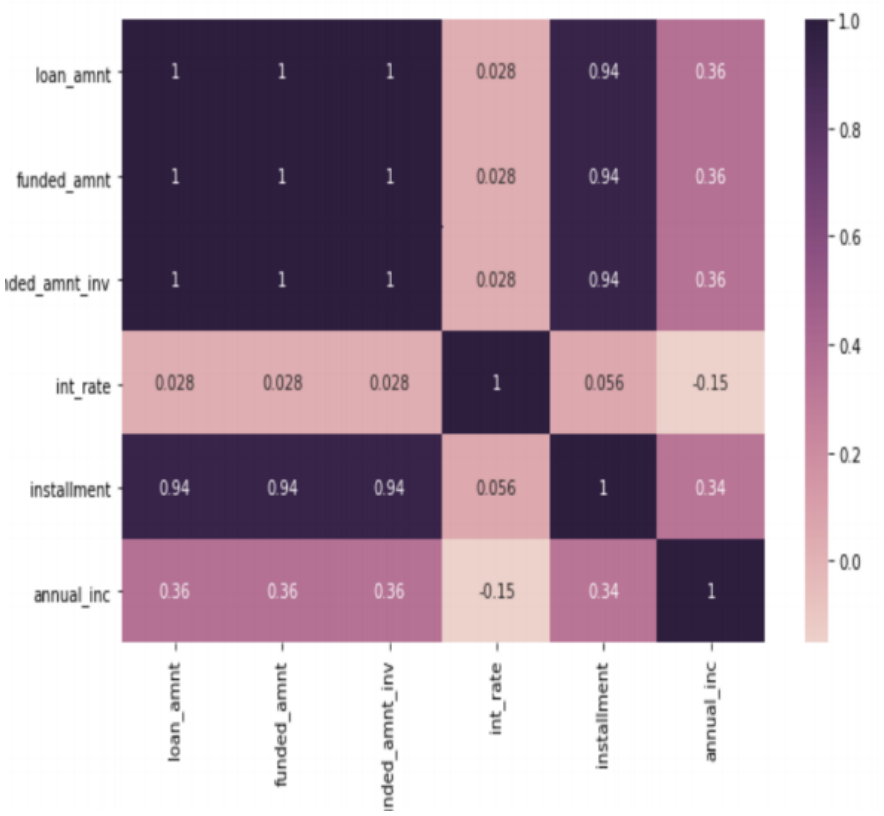
Use Case: **DATA**

—

LendingClub

- 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



Use Case: **FEATURE SELECTION**

Package 'Boruta'

Original features

F1	F2	F3	F4
1	4	7	10
2	5	8	11
3	6	9	12

S1	S2	S3	S4
2	5	9	12
1	4	7	10
3	6	8	11

Shadow features

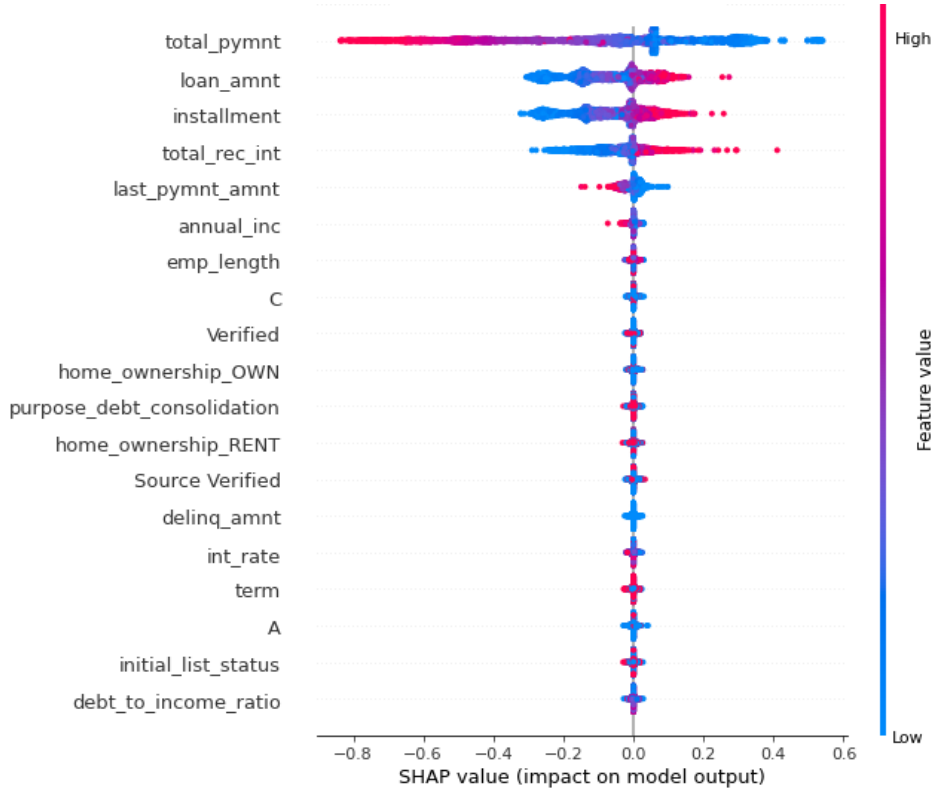
```
> fit = randomForest(y~x, data = trainingset, maxnodes = 10, ntree = 500)
```

	F1	F2	F3	F4	S1	S2	S3	S4
Importance score	1.2	1.01	0.85	0.92	0.001	0.02	0.41	0.91
Hit	✓	✓		✓				

Use Case: **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]

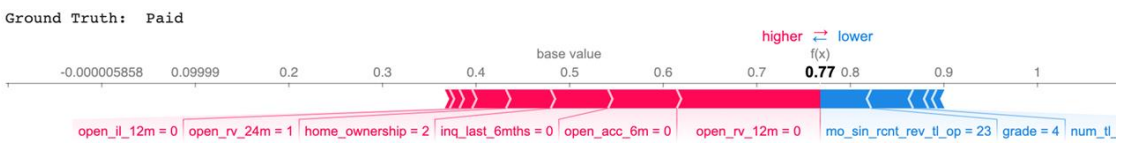
HUMAN-CENTRIC Issues

Interviews carried out with various stakeholders.

The main barriers for wider adoption of ML-based solution in finance;

The need for explainable and interpretable ML;

Specific explainability needs and XAI methods



- Explanations for **model developers**
- Could provide value for end users as well – however, **counterfactual explanations preferred**
- Visualization **not suited for end users**

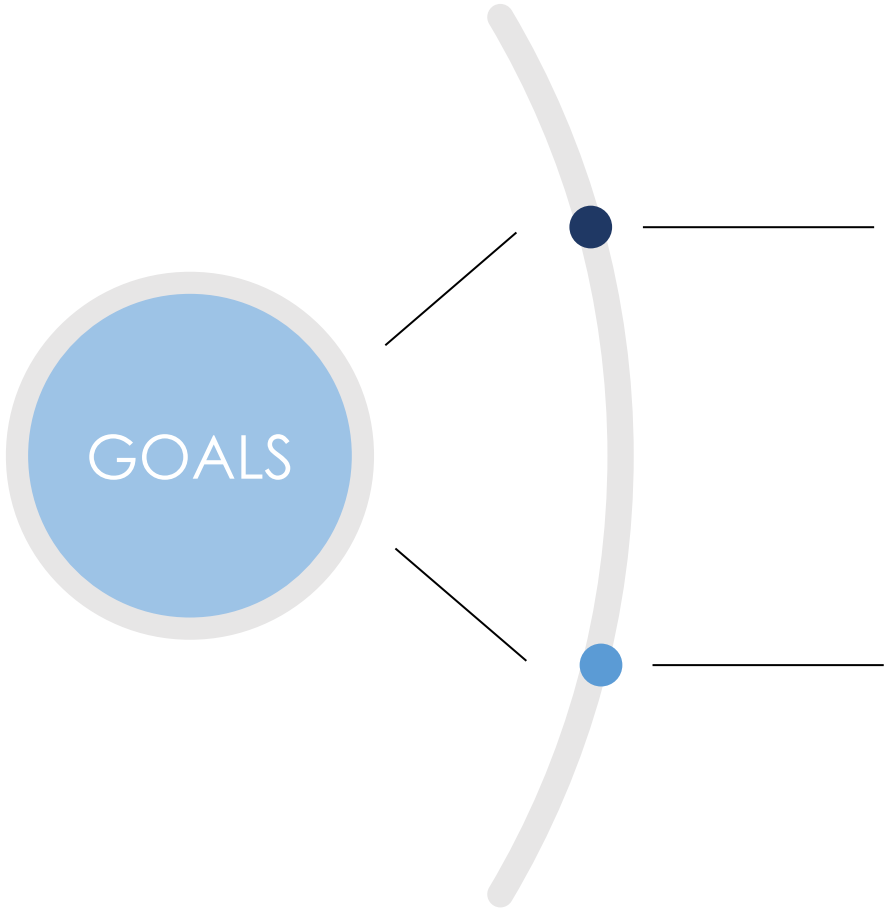
TECHNICAL Issues

- One-point access to data

- Issue with the **different estimation procedures**
  - the exact computation of the Shapley value is computationally intensive
  - Feature selection can be crucial
    - The choice of features that count as players can affect the resulting explanations

- Only few model-specific solutions for the computational complexity

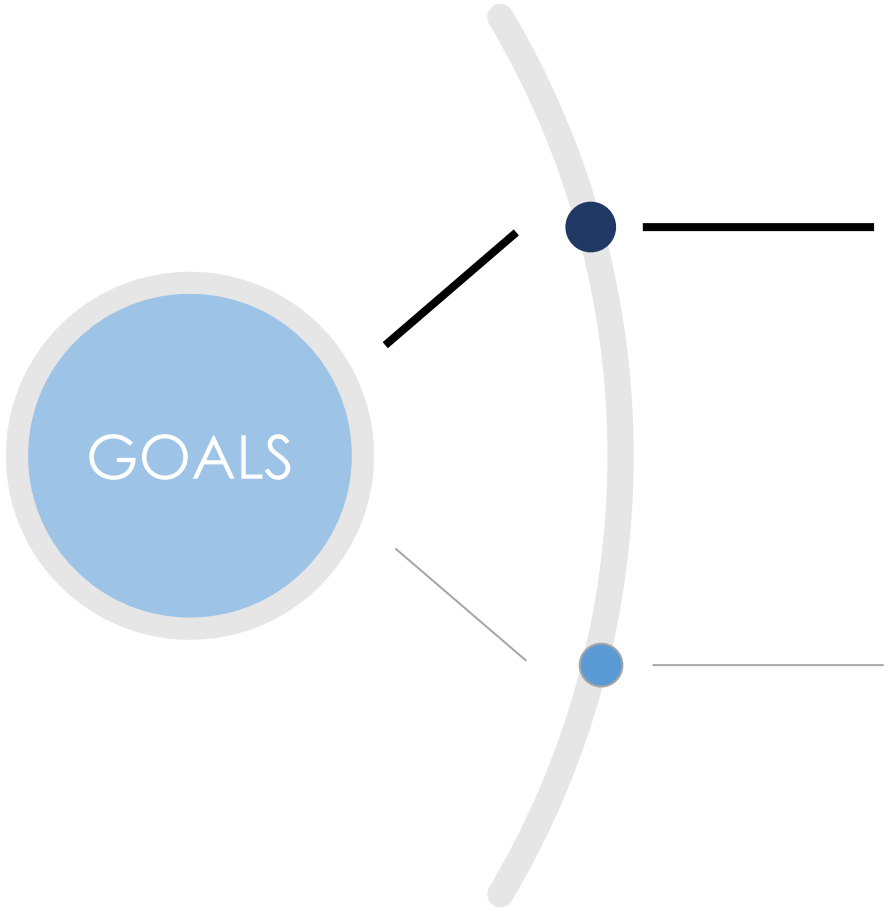
ROBUSTNESS & STABILITY of Explanations



Similar data points/loan contracts should have similar outputs and **similar explanations**

Explanations across different XAI methods should be similar for similar data points

ROBUSTNESS & STABILITY of Explanations



Similar data points/loan contracts should have similar outputs and similar explanations

Explanations across different XAI methods should be similar for similar data points

Stability of Explanations though **GRAPH THEORY**

—

- Use concepts from **graph theory** to investigate whether similar loan contracts have obtained similar explanations

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

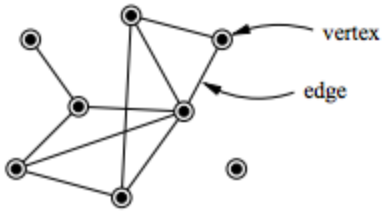


Image source: wikipedia

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

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

The Minimal Spanning Tree

- We derive the **Minimal Spanning Tree (MST)** representation of the loan contracts

- 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.

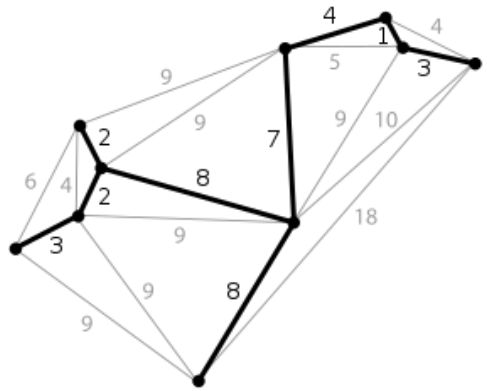


Image source: wikipedia

Stability of Explanations though **GRAPH THEORY**

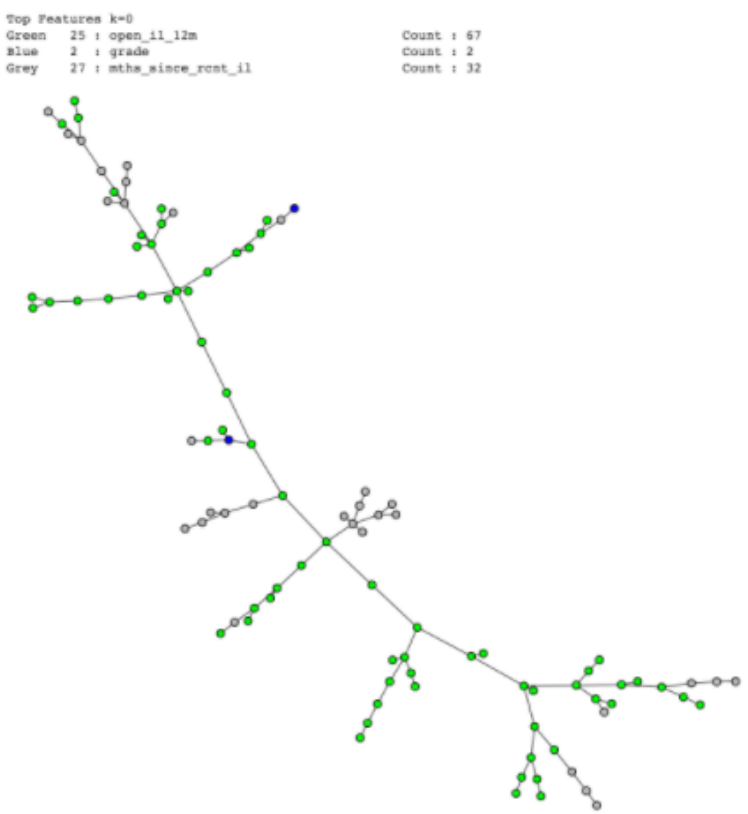


Figure. MST tree representation of 100 random data points. Coloring based on the top explanatory feature [green = "Number of instalment accounts opened in past 12 months"; grey = "Months since most recent instalment accounts opened" ; blue = "Grade"]

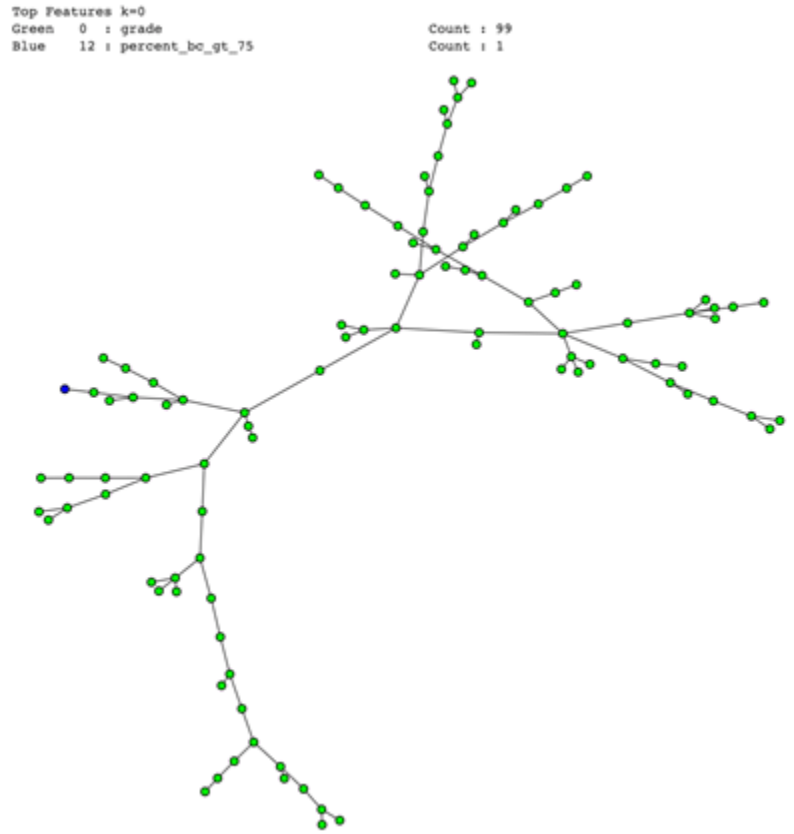


Figure. MST tree representation of 100 random data points. Coloring based on the top explanatory feature [green = "Grade", blue = "Percent of trades never delinquent"]

Stability of Explanations though **GRAPH THEORY**

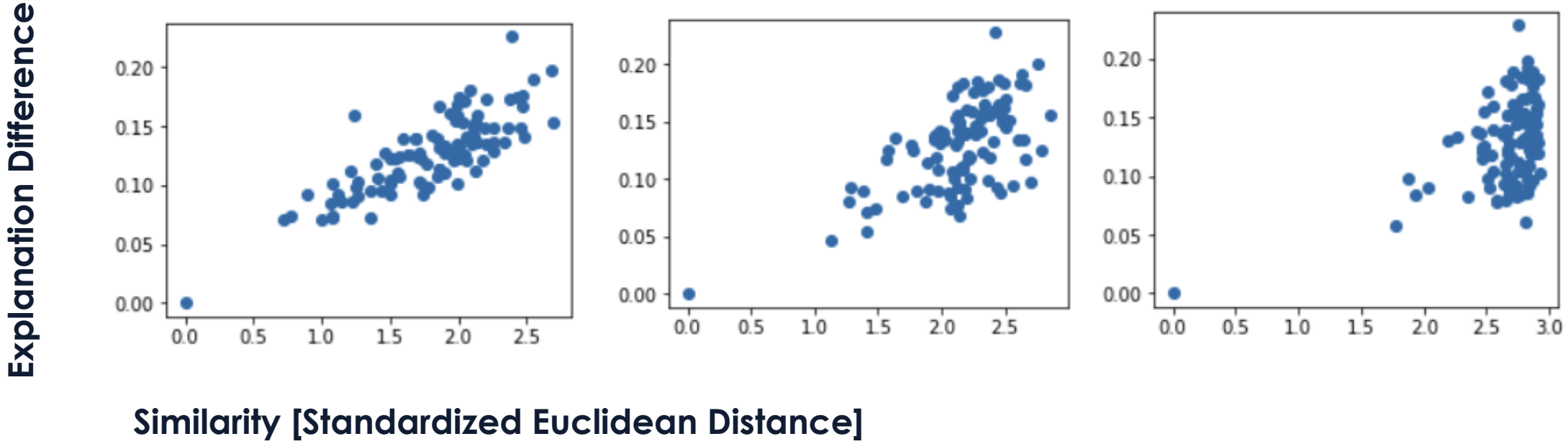
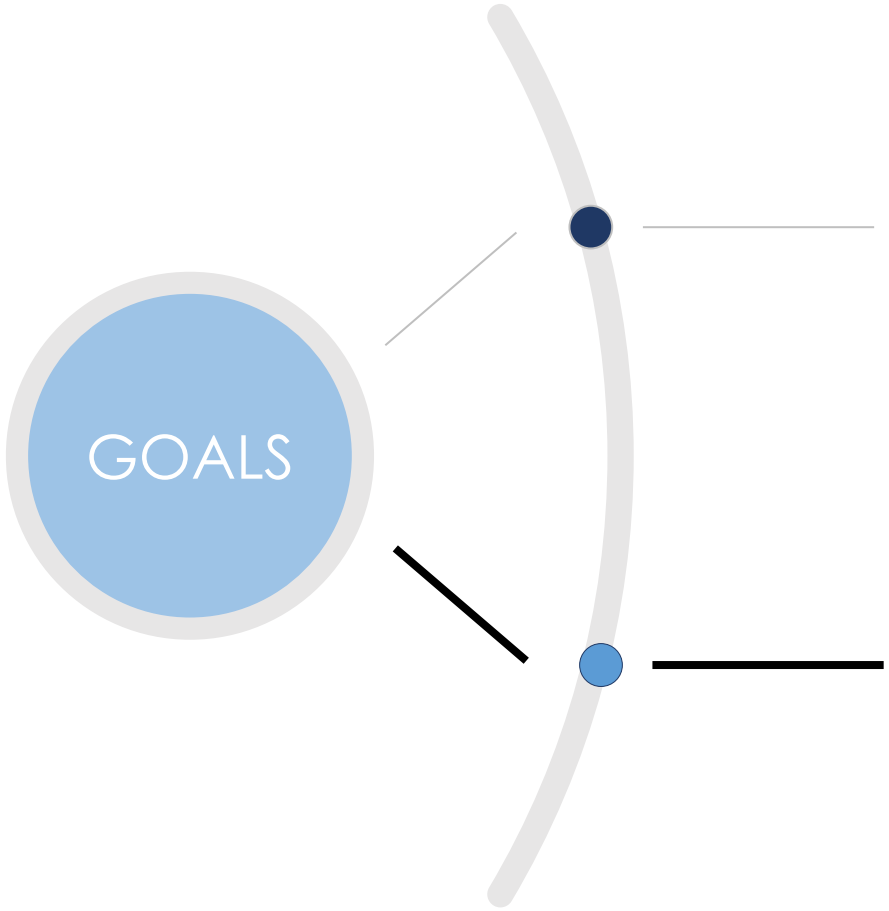


Figure. Explanation Difference vs Spatial Distance for ref_i = 1000, n = 100 for 5, 10, and 20 Features.

*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.

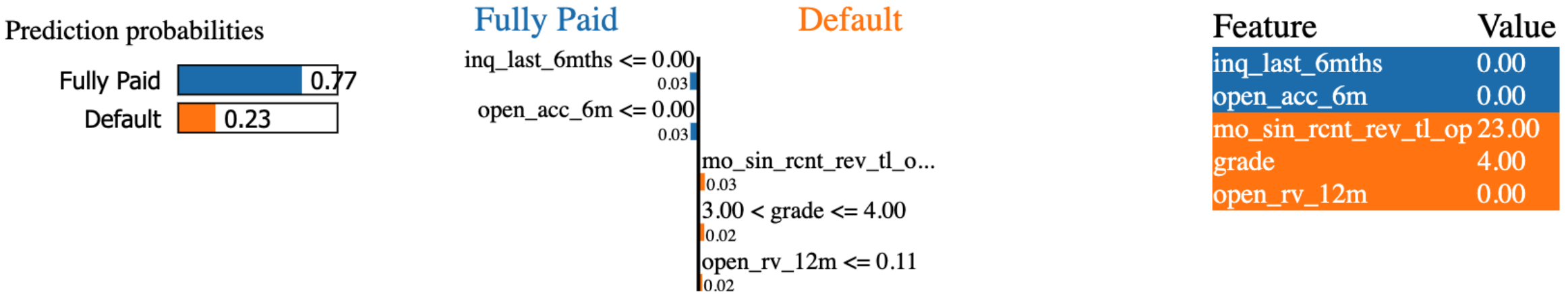
ROBUSTNESS & STABILITY of Explanations

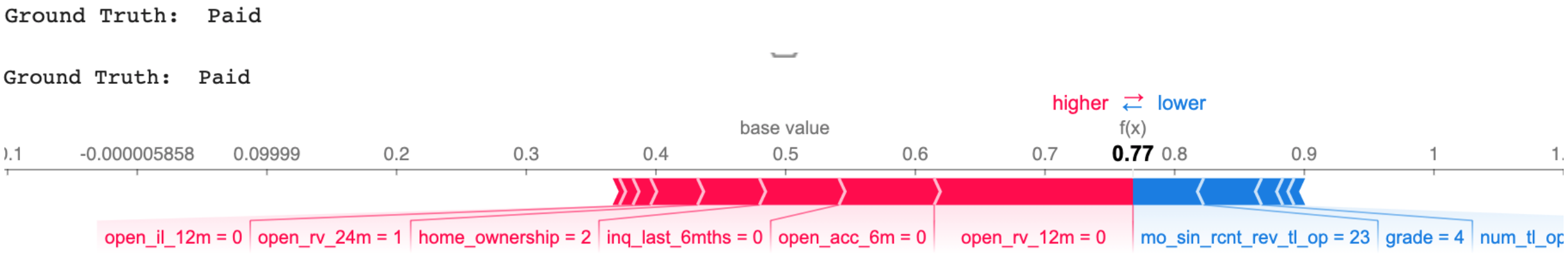


Similar data points/loan contracts should have similar outputs and **similar explanations**

Explanations across different XAI methods should be similar for similar data points

Stability across **XAI METHODS**





CONCLUSION Remarks - I

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

- **Research on XAI applications in finance remain limited**

- **Two-fold objective** of the work:
  - human-centric and mathematical issues related with the implementation of XAI methods in finance, and
  - explore the stability and robustness of explanations provided

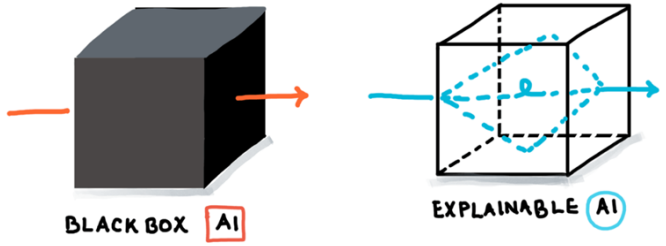
- **Human-centric issues** → we find that that XAI methods are suited to the needs of ML engineers

CONCLUSION Remarks - II

- **Deployment** → various problems arise from the estimation procedures that are in use for some of the post-hoc explainability techniques
  - This in turn affect their practical utility

- **Stability and robustness:**
  - State-of-art methods offer certain level of stability
  - Similar loan contracts obtain similar explanations
  - Explanations across XAI methods for similar loans are consistent

- Future work: **brining XAI literature closer to industry**



WORKING Papers

- Hadji Misheva, B., Osterrieder, J., Hirsa. A., Kulkarni, O., Lin, S. (2021) Explainable AI in Credit Risk Management. DOI:10.2139/SSRN.3795322
- Hadji Misheva, B., Osterrieder, J., Hirsa. A., Kim, P. and Raita, A. (2021). XAI in Finance: Focus on Stability and Robustness of Explanations. Working Paper
- Hadji Misheva, B., Osterrieder, J., Posth.A., Gramespacher, T. and Jaggi, D. (2021). Audience-dependent Explanations for AI-based Risk Management Tools: A Survey. Working Paper