Derisking the Black Box

How Explainable AI Validation help building (and actually using) Machine Learning systems we can trust

Incontro Annuale Comitato Regionale della Lombardia
17 dicembre 2021

Presenting today

Elena Pizzocaro

Partner, McKinsey & Company

elenapizzocaro

Machine learning related risks arise over various dimensions and create new challenges for risk management functions

Legal and regulatory risks	Using certain customer characteristics is illegal in some use cases/geographies (e.g. gender discrimination in motor insurance) – bias in model outcomes is the new focus for ML models Legal consequences and regulatory fines can have a significant negative impact
Reputational risks	Machine learning model outputs and actions that are publicly available (e.g. quoted prices, accidents of self-driving cars, …) can lead to reputational risks Damaged reputation can have impact in various ways (e.g., revenue loss, loss of talent, …)
Model performance risks	Higher risks of overfitting ML models, leading to poor performance in production Self-learning algorithms can suffer performances drops in the course of deployment depending on intake of new training data
Operational risks	Self learning algorithms require frequent data feeds – data pipelines need to be constructed and quality of data monitored continuously, e.g. to detect anomalies like changes in data definition in sub-systems to avoid underperformance or breakage Overly complex model landscape can lead to inefficiencies and loss of control

Source: McKinsey, Risk Dynamics

Derisking the use of AI and ML with a twofold approach

Extended approach to Model Validation

Extended approach to validation and monitoring of models including use of new tools and techniques where required

Explainable AI (XAI)

New methods able to shed light on model outputs both at the individual and global level

Example of extended Model validation framework

Similarity to traditional validation Identical Some modifications New element

Dimensions	A	B	C	D	E	F	G
1 **Model environment**	Intended use(s)	Intended domain of applicability	Model requirement(s)	Model specification(s)			
2 **Input**	Development data set	Quality	Treatment(s) & assumption(s)	Input model(s)	Feature engineering		
3 **Model development process**	Theory	Modeling techniques	Modeling assumption(s)	Hyper- parameters			
4 **Output**	Accuracy	Precision	Robustness	Business operational Indicators	Interpretability	Bias	
5 **Implementation**	System documentation	Production environment	Data import process	Processing code	Report generation	Implementation controls	Scalability
6 **Ongoing monitoring**	Ongoing monitoring plan coverage	Program execution	Escalation process	Metrics and acceptance criteria			
7 **Reporting & use**	Report(s) contents	Model effective use(s)	Output(s) adjustment				
8 **Model governance**	Review Plans & Controls	Model Risk Scoring					

Source: McKinsey, Risk Dynamics

Derisking the use of AI and ML with a twofold approach

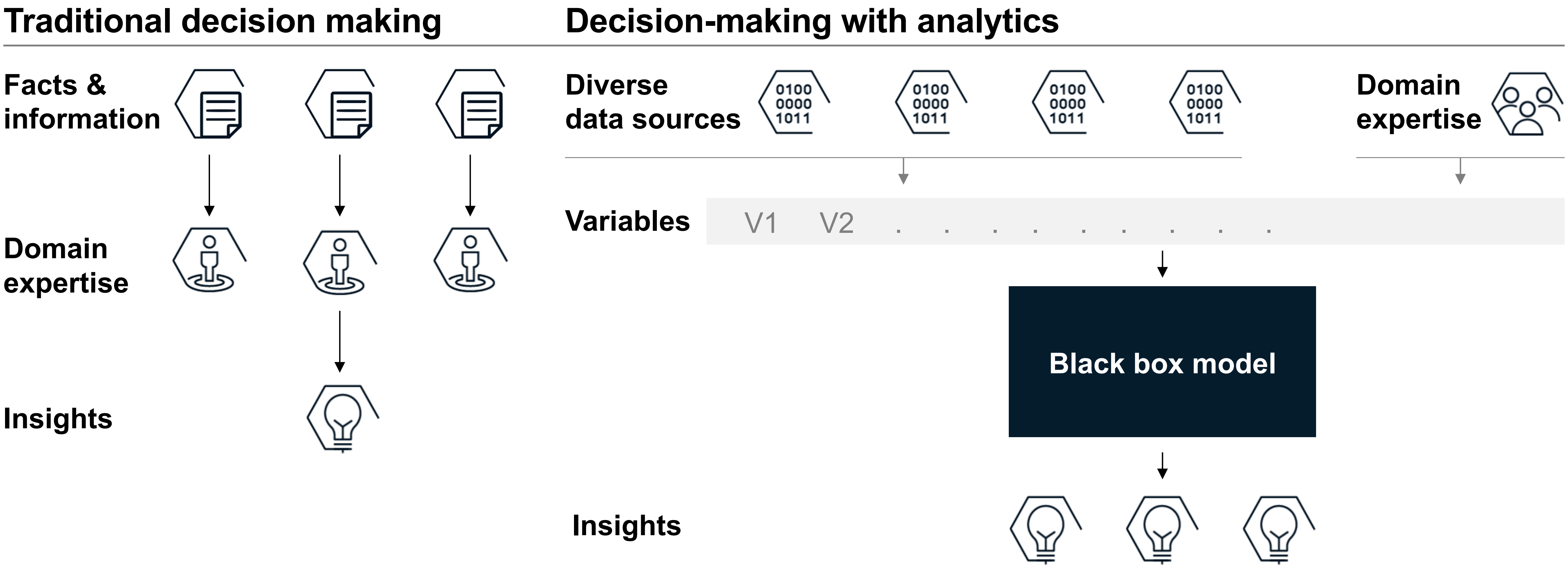
Extended approach to Model Validation

Extended approach to validation and monitoring of models including use of new tools and techniques where required

Explainable AI (XAI)

New methods able to shed light on model outputs both at the individual and global level

Machine Learning models have been increasingly embedded in business decision making



Do we need interpretable or high performing models?

Advocates of interpretability

Regulators

Users

Brokers

//

Need to fully understand how the model works to trust it

Advocates of performance

Corporate decision makers

Large scale institutions

Analytics experts

//

Predictive performance in real-life evaluation trumps interpretability

From The Mythos of Model Interpretability [Zachary C. Lipton]

Source: McKinsey, QuantumBlack

Do we need interpretable or high performing models?

Advocates of interpretability

Regulators

Users

Brokers

//

What is their argument?

There is a "right to explanation"[1]

Sometimes a single error can incur enormous costs

Sensitive information (race, gender) may be misused or inferred by models

Advocates of performance

Corporate decision makers

Large scale institutions

Analytics experts

//

What is their argument?[2]

A powerful model is more profitable to an understandable one

Human decision-makers can be biased too

Machine Learning can be more accurate at predicting than human experts

1. The Mythos of Model Interpretability Zachary C. Lipton
2. A.I. vs M.D, Siddhartha Mukherjee

How do you achieve model explainability?

#1: (Traditionally)

Create easy-to-explain features

Domain knowledge, low dimensional datasets

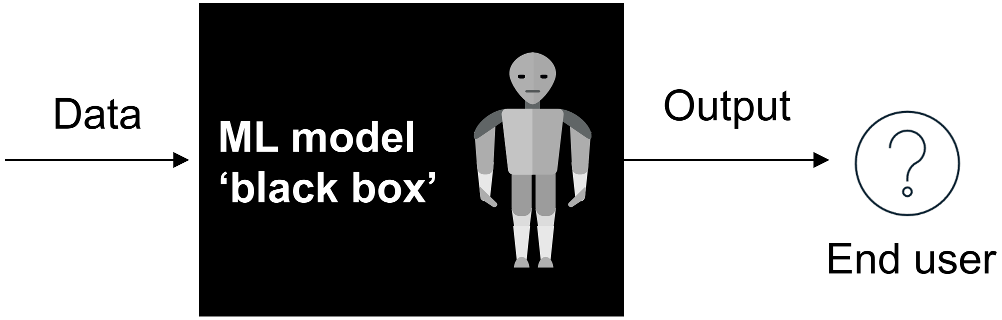
#2: (State of the art methods)

Explain each sample post-hoc

Integrated explainability algorithms

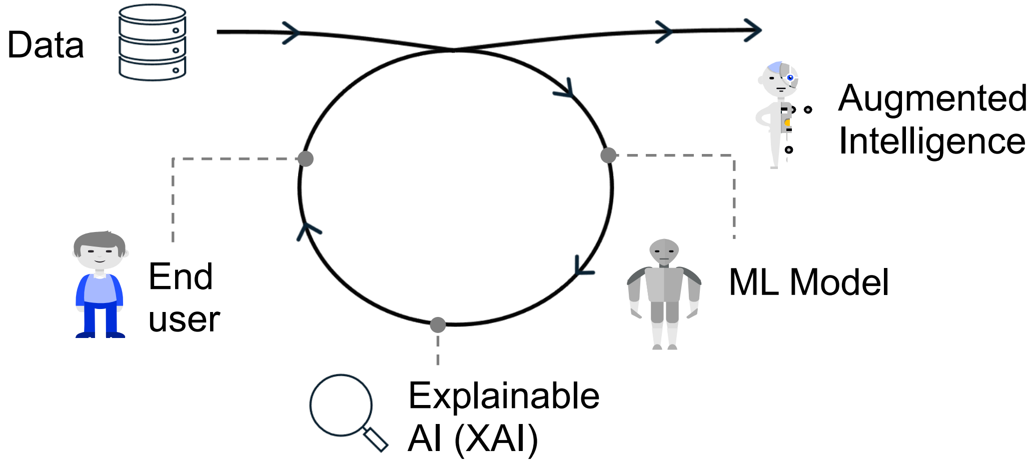
'Explainable AI' (XAI) bridges the gap between 'black-box' Machine Learning models and the users

'Black-box' Machine Learning



- Very high **predictive power**
- Limited input from human expertise
- **Lack** of **transparency** hurts adoption
- Increased ethical / regulatory risks

'Explainable AI'



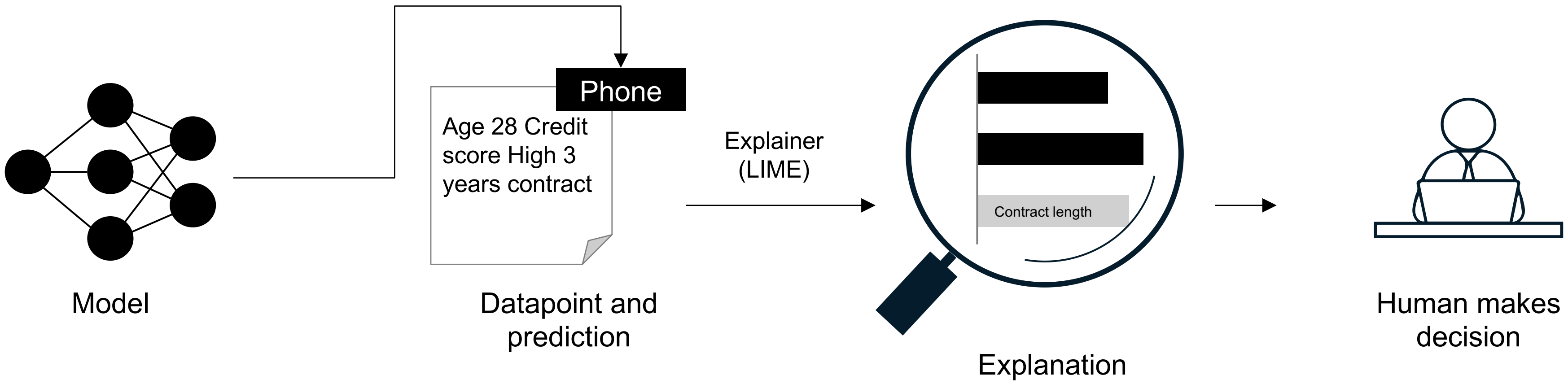
- Very high **predictive power**
- **Trust** in model output enables adoption
- **Intelligence augmentation**, combining human and machine insight
- **Addressing regulatory / ethical requirements**

Source: McKinsey, QuantumBlack

XAI methods work to shed light on model outputs both at the individual and global level

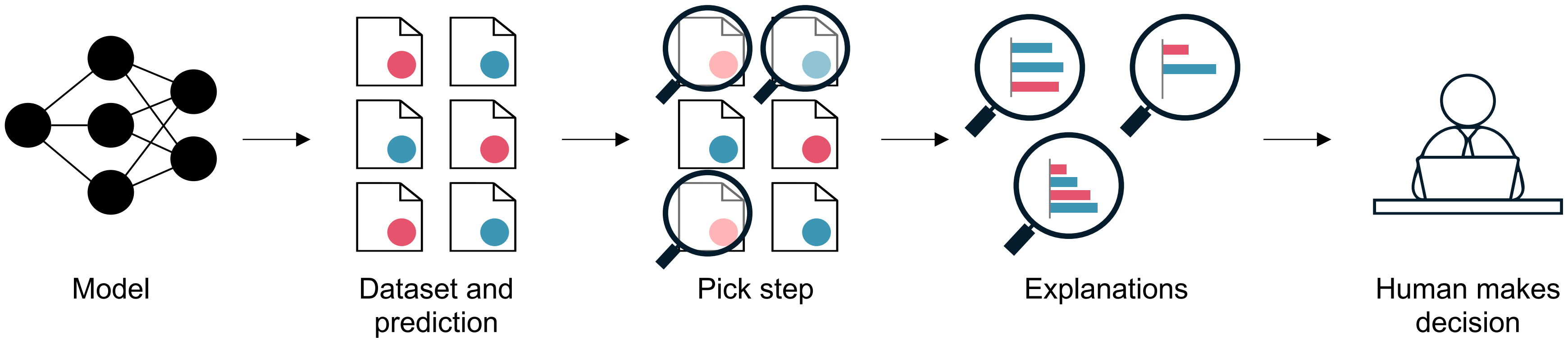
Individual explanations

Explain why the model generates this output for one particular instance



Global explanations

Pick representative examples from a dataset or illustrate global-level relationships/ patterns learnt by the model

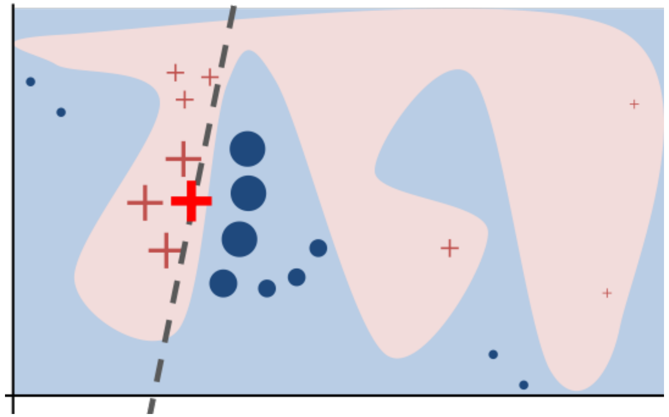


Images adapted from: https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

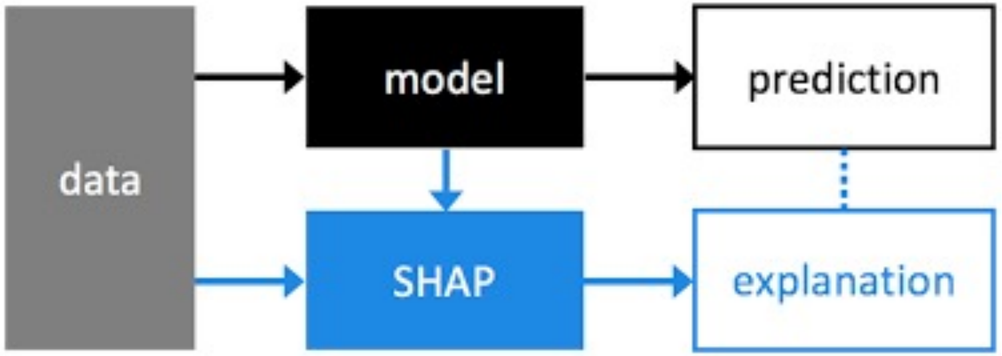
Source: McKinsey, QuantumBlack

Different examples of integrated explainability

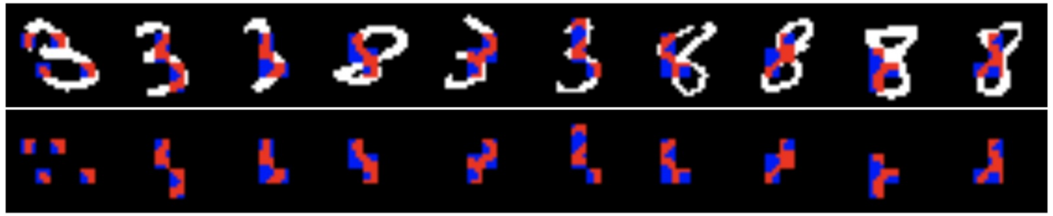
LIME (Locally Interpretable Model-agnostic Explanations)[1]



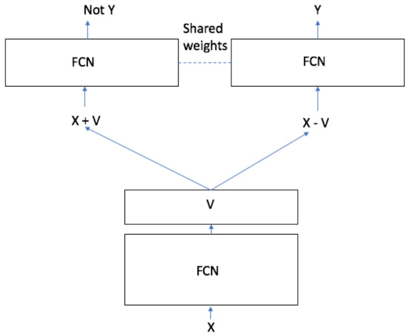
SHAP (Shapley Additive exPlanations)[2]



L2X (Learn to Explain)[3]



EMAP (Explanations by Minimum Adversarial Perturbations)[4]



1. Ribeiro et al., "Why Should I Trust You?": Explaining the Predictions of Any Classifier, https://arxiv.org/abs/1602.04938
2. Lei et al., Rationalizing Neural Predictions, https://arxiv.org/abs/1602.04938
3. Letham et al., Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model, https://arxiv.org/abs/1511.01644
4. QuantumBlack

Source: McKinsey, QuantumBlack

XAI is relevant to several types of users in insurance

Agents	**Identifies leads** with greater confidence and the **preferred channel** (email, phone, etc.) **Better conversations with customers**
Commercial strategist	Generates additional **business insights** for **strategy**, **product design**, **marketing**, etc.
Risk manager	Uses XAI to ensure **regulatory compliance** Reviews population cohorts to **identify sources of bias** in the model
Actuaries	**Improves model performance** by**:** • Collecting **input from business experts** • **Analysing misclassified examples**

Source: McKinsey

How explainability is key in adopting AI in actuarial problems (e.g. pricing, reserving)

Identify drivers of deviances between ML models and traditional actuarial methods and understand structural/exceptional perturbations

Validate business rational underlying estimates, and correct potential bias

Overcome internal resistances in adopting the advanced models to assist the business-as-usual (e.g., open/closed file reviews)

Source: McKinsey