Delta Capita

# Challenges in Machine Learning Adoption

Edward Adcock
Thuy Nguyen

Version v1.0
February 2019

DC MINT

# Background

## Machine Learning

Machine Learning is the study of algorithms and models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of computer science and a branch of artificial intelligence (AI).

Machine Learning is used in applications such as image recognition, speech recognition or email filtering.

Machine Learning algorithms can be classified in groups based on the tasks they are designed to solve. Common Machine Learning tasks are:

- **Classification** – Output variable is a category, for example, predicting whether the colour of the box is 'Red' or 'Green'
- **Regression** – Output variable is a real value, for example, predicting the price of oil
- **Clustering** – Data points are assigned to clusters, for example, organizing documents in clusters based on their content.

## Machine Learning Interpretability

Machine Learning Interpretability refers to the ability to make the behaviour and predictions of Machine Learning systems understandable and helps explain crucial aspects of our models:

- What drives model predictions?
- Why did the model take a certain decision?
- How can we trust model prediction?

Interpretability is not only relevant to system engineers and data scientists, but also end-users, model validation teams, risk and compliance for whom it is important to understand the results generated from Machine Learning models.

There are two important types of interpretability:

- **Global** – Is the model as whole sufficiently understood to the extent required to trust it (or to convince someone else that the model can be trusted)?
- **Local** – Is it possible to explain the reason for a particular decision with respect to model inputs?

## Financial Models

In the past, decision-making was the domain of individuals trusted by the firm to competently assess the available information and make a risk weighted decision on the outcome. If this decision proved to be incorrect, then the individual could be questioned for the reasoning behind it.

Early adoption of machine learning methods in the finance sector was limited to well-studied statistical methods, like:

- Linear regression
- Time series analysis
- Binary decision trees

More recently, machine learning models are becoming increasingly complex reflecting the advances and increasing sophistication of the underlying training algorithms. However, the accuracy of these models is in many cases superior, hence their ubiquity throughout finance and many other sectors, with a clear growth trend over time.

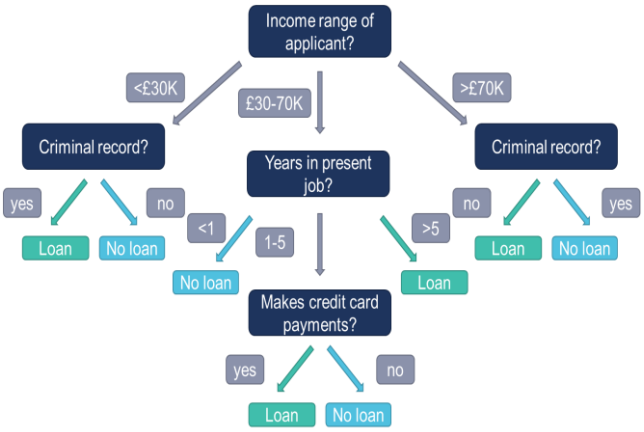



*Figure 1: Example of binary decision tree on credit card application*

These algorithms are now helping key decision-making within corporations

- Risk analysis
- Portfolio management
- Loan applications

These state of the art deep learning models are trained on very large datasets and have millions of internal parameters.

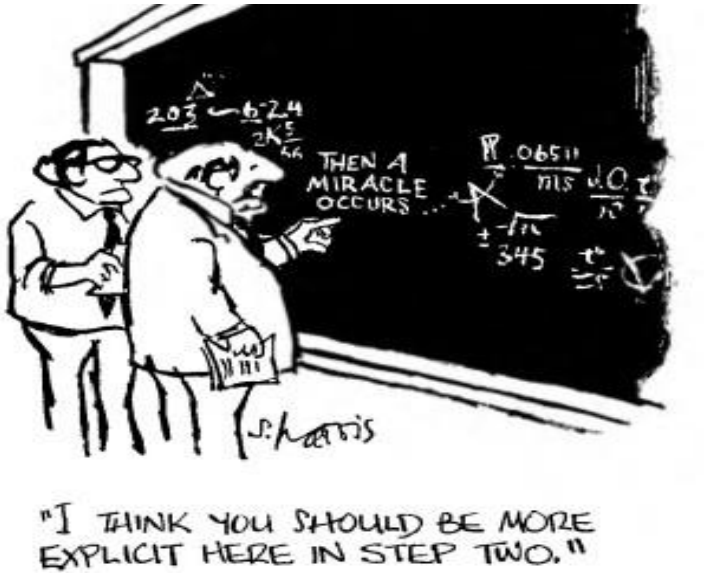
# Challenges

## Model Interpretation

Interpretability should be seen as a range, not a binary state, which means it is invalid to say that a model is 100% explainable or 100% unexplainable. Interpretation refers to the degree to which a human can understand the cause of a decision made by machine learning models.

Explanations may not fully explain the situation, but they should address the root cause.

For example, in the case of product recommendations, we always think about why certain products have been advertised to us. While we may not fully understand how recommendation systems derive the decision, it is obvious that an advertisement about oven gloves follows me on the Internet because I recently bought a new oven online, and oven and oven gloves are frequently purchased product combinations




Machine leaning model challenges can be summarized as follows:

1 **Model Bias** - The underlying data, feature selection and training process may introduce bias into models

2. **Model Validation** - Understanding model behaviour is a pre-requisite for independent model validation

3. **Regulatory Framework -** Regulators require firms to be able to explain how their models work (e.g. GDPR, SR 11-7)

4. **Model Confidence -** Explanations of model predictions creates confidence in model users

> *Model interpretability is about making "Black Box" models a "White Box".*

In the context of big data, the model itself becomes the source of knowledge instead of data. The state of art models can be very complex Neural Networks or large ensemble models using a large set of data. Implementing model interpretation enables us to extract additional knowledge captured by the model

> *A correct prediction only partially solves our problem, the model must also explain how it derives the prediction.*

Interpretability implementation is also based on the nature of the problem. Examples of models not requiring explanations include:

- A movie recommender system where making mistakes will not have serious consequences
- An optical character recognition where the method has already been extensively studied and evaluated

The need for interpretability comes from an incompleteness in problem formalization. To be specific, a correct prediction only partially solves our problem, the model must also explain how it derives the prediction.

Examples of models requiring explanations include:

- GDPR and SR11-7 regulation – People have the right to question why a machine learning mortgage default prediction model rejects their loan application.
- Trading desk will want to know how an FX currency pair price prediction machine learning models arrived at a decision as this may have a significant market impact.
- Credit Card fraud detection – False positives (incorrect classification of non fraudulent transactions as fraud) can prove to be costly.

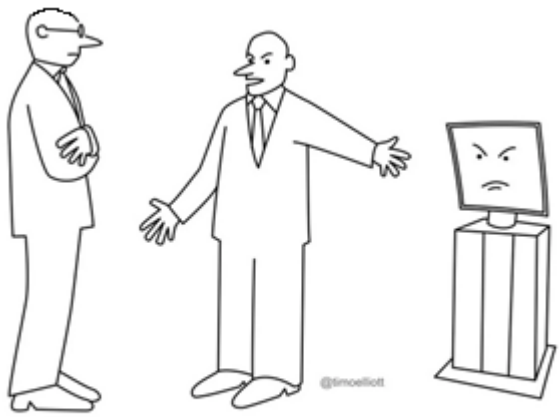
# Challenges

## Global and Local Explanations

The main approach to tackling challenges in Machine Learning Interpretability is to provide global or local explanations.

Global model explanations give an overview about how the entire model makes decisions and which key factors contribute to the prediction. A separate, less complex interpretable model is used to mimic the performance of our original model, from which important information of the original model is captured. By inspecting the interpretable model, we can offer explanations.

Local model explanation zooms in on a single instance and examine each feature's contribution to the prediction of that particular observation. Feature perturbation is one of the methods that can be used for this purpose.



*His decisions aren't any better than yours — but they're WAY faster...*

Being able to evaluate the quality and robustness of explanations is another important factor requiring careful consideration. There are a number of criteria to asses this:

- **Consistency** – How much does an explanation differ between models that produce similar predictions?
- **Stability** – How similar are the explanations for similar data examples?
- **Complexity** – How well do humans understand the explanations?
- **Representativeness** – How many instances does an explanation cover?

## Human Interpretation

The role of humans in a machine learning context should not be overlooked. Humans are involved in many steps during the process of data preparation, development, training, testing and deployment of machine learning models.

- Data collection and augmentation
- Model and algorithm selection
- Hyperparameter tuning
- Model evaluation
- In field use

All these aspects may have a significant impact on the final model performance and suitability. It is crucial that domain experts/practitioners are involved in this process.

> *Humans build trust in a model once they understand it and consider the actions it takes given certain information reasonable*

Whether the model is used as a tool by a practitioner or deployed to run fully autonomously, if the users do not trust the model or a prediction, they will not use it.

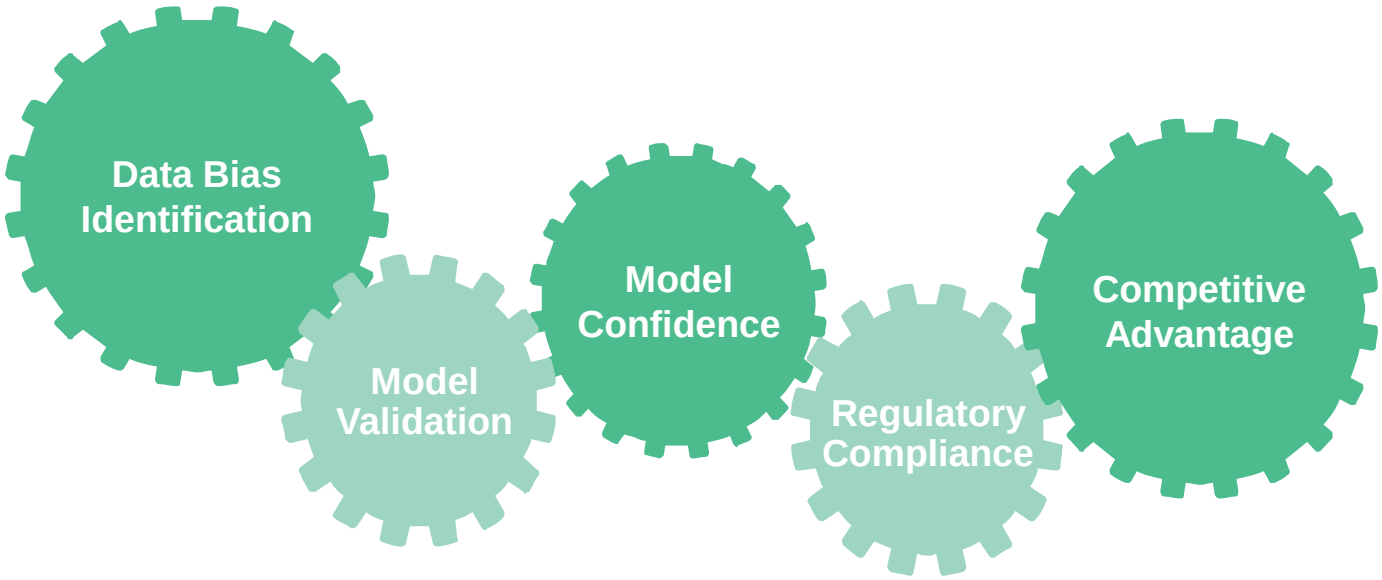
> *Businesses that can build trust and public understanding of their models will have a competitive advantage.*

Ultimately, the human interpretation of machine learning models is as an evaluation tool. The answer that the model gives should be merely one of many information sources that a domain expert will use to interpret the outcome.

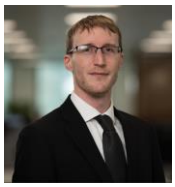
Finally, regardless of which method of model interpretation has been applied, it is always good practice to incorporate domain expert knowledge to assess the quality of the explanations .i.e. whether the factors explained by the model are actually meaningful and have an effect on the outcome in real life and production environments.

# DC MINT

The Delta Capita DC MINT platform helps our clients to address the machine learning interpretability challenges. It is a product for data scientists, validation teams, risk, compliance and other end-users.




**Edward Adcock**
**Data Science Consultant**



Data scientist and Machine Learning specialist with 14+ years experience in financial services, applying state of the art algorithms and neural network architectures on complex data sets.

**Thuy Nguyen**
**Data Science Consultant**

Data scientist consultant specializing in Machine Learning interpretability with experience in financial services, market research and CRM.

**Please contact Delta Capita at *capitalmarketsdelivery@deltacapita.com* if you are interested in any of the following areas:**

- **DC MINT - Model Interpretability**
- **DC COMPRESSION - Model Compression and Model Optimisation**
- **DC DOCS - Automated Model Documentation**
- **DC VOICE - Voice/NLP on large data sets and Wealth applications**
- **DC RISK - Contagion Risk Models**

DC MINT