Delta Capita

AI Adoption and LIME

Alexander Klemm
Huhan Yang

Version v1.0
March 2019

DC MINT

AI Adoption and LIME

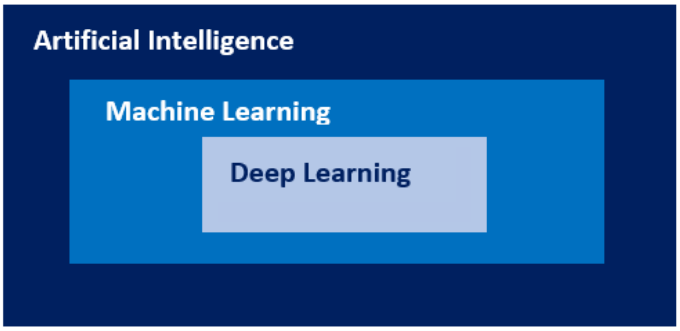
Growing AI Adoption

2018 saw a big push for artificial intelligence (AI) making inroads to become more mainstream and expanding successfully into more areas than before. Firms are increasingly exploring applications for AI and how best to commercialise this. Also, the general public has become more aware of the growing interaction with the technology on an almost daily basis.

The stage is set for AI to continue transforming businesses and individuals. In 2019 and well beyond, the technology will continue to grow in global prevalence and our daily working and personal lives. This prominence will fuel innovative business models but also further regulation and social responsibilities.

The main challenge in AI adoption is explainability i.e. understand what the model does. There are a number of explainability techniques and this paper discussed one such techniques called LIME (Local Interpretable Model-Agnostic Explanations) and why it is an effective technique.

Where does AI sit



AI and Machine Learning have been used interchangeably. This has caused much frustration with people in the know and worth revisiting:

- **Artificial Intelligence:** Any technique that enables machines to mimic human intelligence. This can be through explainable techniques such as decision trees or less explainable techniques such as Machine Learning and Deep Learning
- **Machine Learning:** A subset of AI that includes abstruse statistical techniques that enables machines to improve at tasks as it learns.
- **Deep Learning:** The subset of Machine Learning with algorithms that permit software to train itself to perform tasks by exposing large amounts of data to multilayered neural networks.

AI Explainable Techniques

As supposed to explainable techniques like decision tree, the rising use of AI has seen a surge of black-box techniques like deep learning or algorithms whose inner workings cannot easily be explained - how can you make a black-box a white-box?

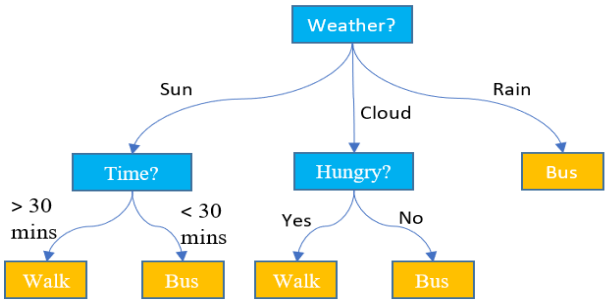


Figure 1: High Interpretability: Decision Tree

These techniques offer breakthrough capabilities and are so effective as they allow for systems to do completely new tasks with much better accuracy, but are unable to easily explain how they do those things.

The interpretability question has been a major factor in adopting AI, especially internally to a firm or end-user or outside the firm to customers, auditors or regulators who typically do not have a data science background.

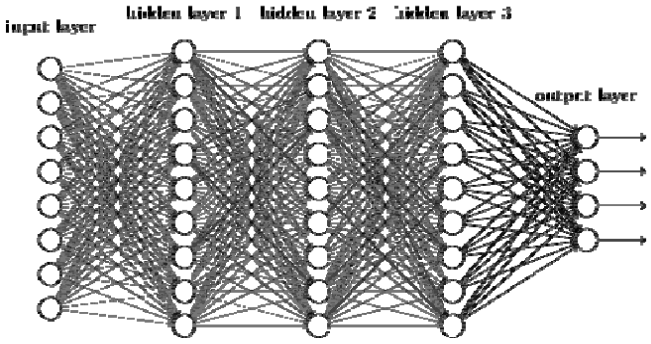


Figure 2: Low Interpretability: Neural Net

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

AI Adoption and LIME

Global and Local Techniques

From the point of view of a business deploying AI in a business process or product, there are two main important types of interpretability:

- **Global:** Do you understand the model as a whole sufficiently to deploy it in the real world?
- **Local:** — Do you understand the reason for particular decisions the model made?

A local technique called LIME (Local Interpretable Model-Agnostic Explanations) is discussed further in this paper. Other techniques are discussed in subsequent papers.

Local Technique: LIME

LIME takes as input a trained model and the training data one wishes to have explained. It then randomly perturbs the features of
the example, and runs these perturbed examples through the classifier. This allows it to probe the surrounding feature area to build up a picture of the classification surface nearby.

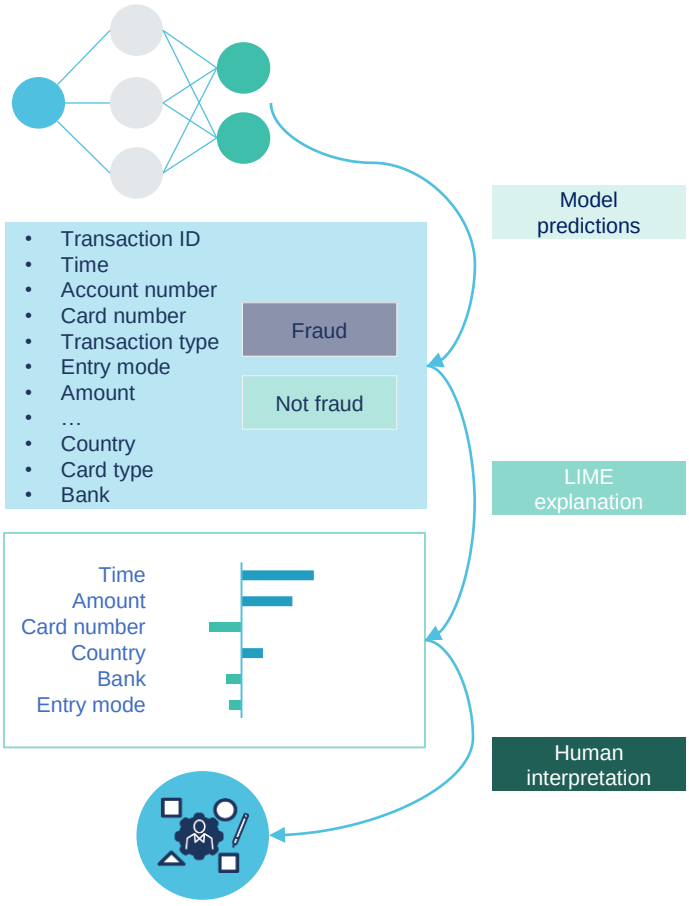


Figure 3: LIME human-friendly interpretations

It probes the classifier thousands of times, and then uses this information to ft a linear model.
The training examples are weighted by distance to the original example. The linear model can then be interpreted as usual to extract explanations.

LIME is a good technique to explain why the model made a particular decision.

LIME's approach is similar to the shadow model approach: the true original model is used to generate training data for a simpler, interpretable model. It should be noted through that whilst the shadow model offers a supposedly global explanation that may be wrong in unknown ways, LIME offers a local explanation that is correct.

LIME Steps

The workflow of LIME with a linear surrogate kernel can be summarized as following:

- Step1: Select the observation we want to investigate
- Step 2: Permute the features of the chosen observation several times, thus creating a new dataset from the local domain
- Step 3: Use the underlying machine learning model to make predictions based on the permuted data
- Step 4: Compute the distances between the permuted data points and the original chosen observation
- Step 5: Convert the distances to similarity scores
- Step 6: Fit an interpretable surrogate model with the results from steps 3 to 5
- Step 7: Extract the weights from the fitted surrogate model. The weights represent the contributions of the corresponding features to the prediction for the chosen observation.

AI Adoption and LIME

Model Agnostic Interpretability

The model-agnostic nature of LIME makes it flexible with respect to the model choice, explanation style, and representation type.

- **Model flexibility:** As there is no single machine learning model that can solve all sorts of different tasks in the real world, the interpretation method that can work with any machine learning model is preferred.
- **Explanation flexibility:** This refers to the various styles of the explanation that the method offers. Some explanations may come in the form of mathematical function of the simplified model while in other cases a graphic representation showing feature importance might be more suitable.
- **Representation flexibility:** The explanation should be representable in a way that is intuitively recognizable by humans even if this representation differs from the representation of the original features.

The user can choose the surrogate model to replace the black-box for interpretation.

The user can choose the type of model to replace the black-box model for interpretation, such as a decision tree, a k-nearest neighbour classifier, a LASSO or Ridge regression model, etc.

Depending on the nature of the machine learning task and the choice of black-box model, one specific type of surrogate model may provide a better approximation than others.

Explanations derived by LIME can generally be represented in the form of the original input data. For instance, in addition to numerical data with numerical and graphical explanations, LIME can also output interpretations in the form of text snippets (e.g., for text classification) or image-based representations (e.g., for face recognition). Hence, despite using a numerical abstraction of the data, like pixel values for images, the LIME output is human-friendly.

The image below illustrates this on an image of the handwritten digit “0” and the same digit overlaid with the LIME interpretation. The green areas contributes positively to the prediction that the digit is indeed “0” and the red area contributes to an alternative prediction where the digit would be classified as ”not 0”. The green areas frame the general shape of the digit “0” in the majority of handwritten samples whereas the red area frames a shape that is also common to “not 0” digits, such as “9”.

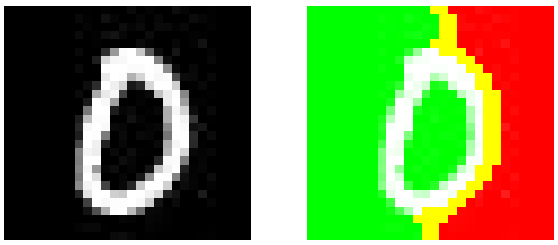


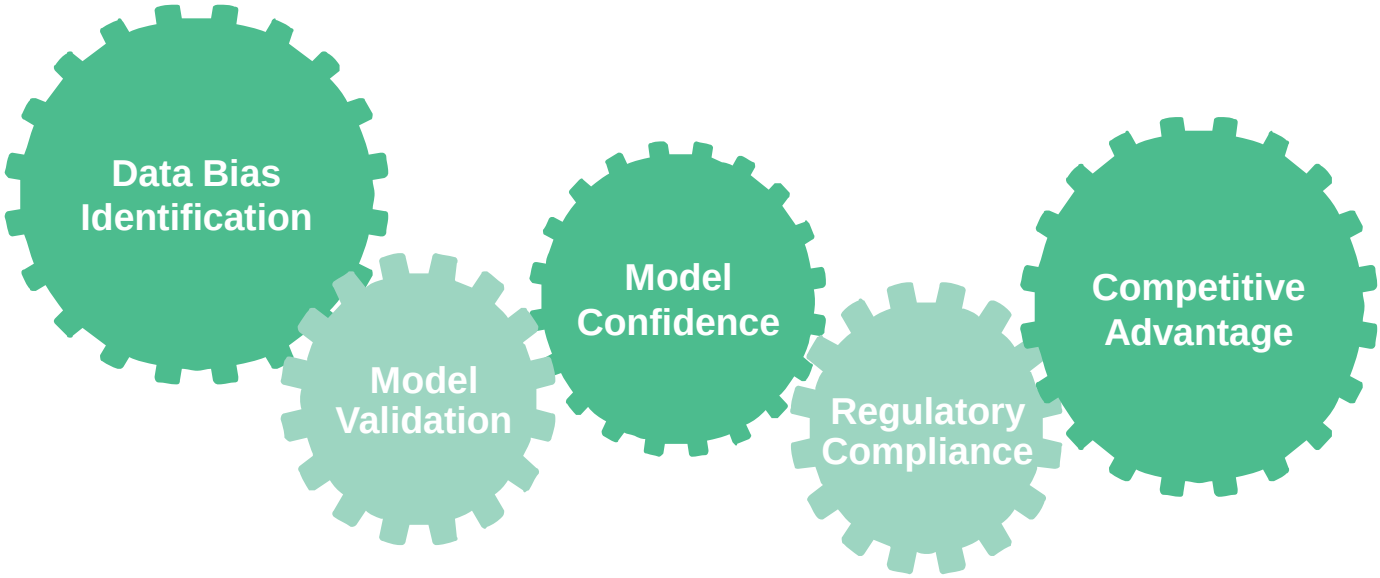
Figure 4: LIME human-friendly interpretations

For interpretations to be understandable, they should be as short as possible. LIME facilitates short explanations because it allows the user to focus on a small number of mechanisms through which the output is generated. For instance, through the weights in a linear surrogate model, we can directly observe the expected change in output when increasing or decreasing individual feature values.

LIME explanations also allow the user to contrast a given case with a reference case. This is important because humans tend like explanations that allow them to understand why A occurred instead of B and how the inputs would need to be different for B to occur rather than A.

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.



Alexander Klemm
Data Science Consultant

Data Science Consultant with several years of experience in various areas across finance, including consulting, investment banking, sales, business development, fintech, and research.

Huhan Yang
Data Science Consultant

Data science consultant specializing in machine learning algorithms with experience in financial service, developing machine learning model-driven products and research.

Other contributors:
Sylvia Smit, Managing Partner
Ricardo Cruz, Senior Consultant
Khrystyna Andronova, Senior Consultant

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