Machine Learning and Official Statistics

Is Explainability an issue?

Maurizio Naldi

LUMSA University

6 December 2023

Machine Learning

- Machine Learning is gaining wide adoption in many fields (you name it, you got it)
- This is also happening in Official Statistics
- Machine Learning also has a (not always deserved) reputation for not being transparent
- Is this going to hamper its application in Official Statistics?
- I hope to show that does not need to be the case

A roadmap in brief

- What can Machine Learning be useful for in Official Statistics?
- Why is Explainability relevant in Official Statistics?
- How can the Explainability requirement be met?

What is Statistics?

The science (or art) of counting?

- We need to decide:
  - What is to be counted
  - Counting them all or not
- Both have links with ML
- What is to be counted is related to classification (a major task for ML)
- Counting a sample instead of the shown population is related to sampling is what is done when training an ML algorithm

Machine Learning in Official (and near-official) Statistics (1/3)

- **Data Imputation**:
  - Filling in missing or incomplete data in official statistics
- **Data Quality Assurance**:
  - Identifying and correcting errors, inconsistencies, or outliers in datasets
- **Survey Sampling**:
  - Optimizing survey sampling methods to select representative samples more efficiently
  - Reducing the cost and time associated with data collection
- **Data Classification**:
  - Automatically classifying data (e.g., products into different industries or services into various sectors)

Machine Learning in Official (and near-official) Statistics (2/3)

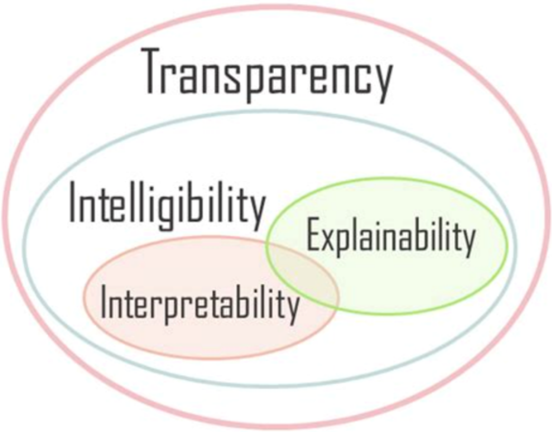
- **Anomaly Detection**:
  - Detecting unusual patterns or anomalies (outliers or irregularities in data)
  - Anomalies may indicate errors or significant changes in trends
- **Time Series Forecasting**:
  - Predicting future trends based on historical data (e.g., for economic forecasting and population projections)
- **Sentiment Analysis**:
  - Analyzing social media data and other unstructured sources using natural language processing
  - Providing insights into public sentiment and understanding public opinion and potential biases.
- **Geospatial Analysis**:
  - Aiding in mapping and monitoring trends and spatial patterns, such as population distribution, land use, and environmental changes.

Machine Learning in Official (and near-official) Statistics (3/3)

- **Data Linkage**:
  - Linking datasets from different sources (e.g., integrating data from multiple domains)
- **Fraud Detection**:
  - Examining financial and economic data to detect fraudulent activities and transactions
- **Natural Language Processing**:
  - Extracting structured information from unstructured text data (reports, news articles, and survey responses)
  - Generating summaries
  - Offering a text-based input/output interface that helps familiarity with non-expert users of ML software
- **Data Visualization**:
  - Creating advanced data visualization tools to help present statistics

Interpretability vs Explainability

- **Interpretability** means understanding the cause and effect with an AI system:
  - Which features are most relevant for the AI algorithm output?
  - What if some input changes?
  - And why did the algorithm get it wrong?
- **Explainability** is related to communicating the why and how of interpretability
  - Reaching a wide audience using a human-readable form

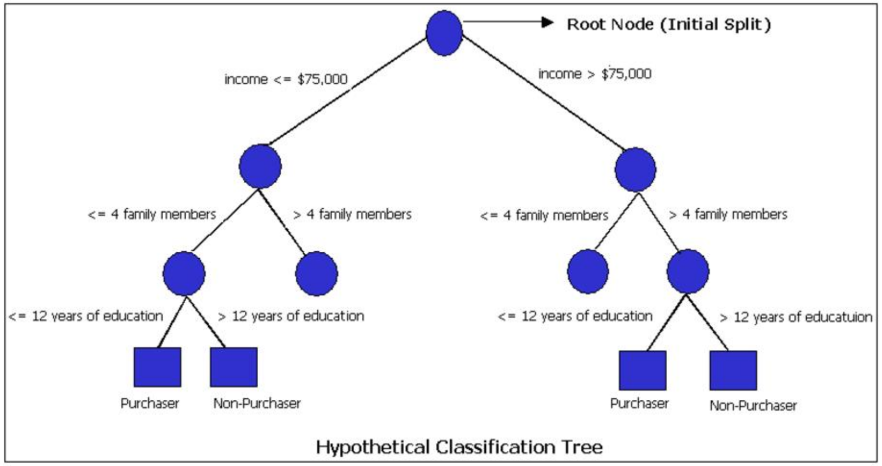


Is Explainability needed in Official statistics?

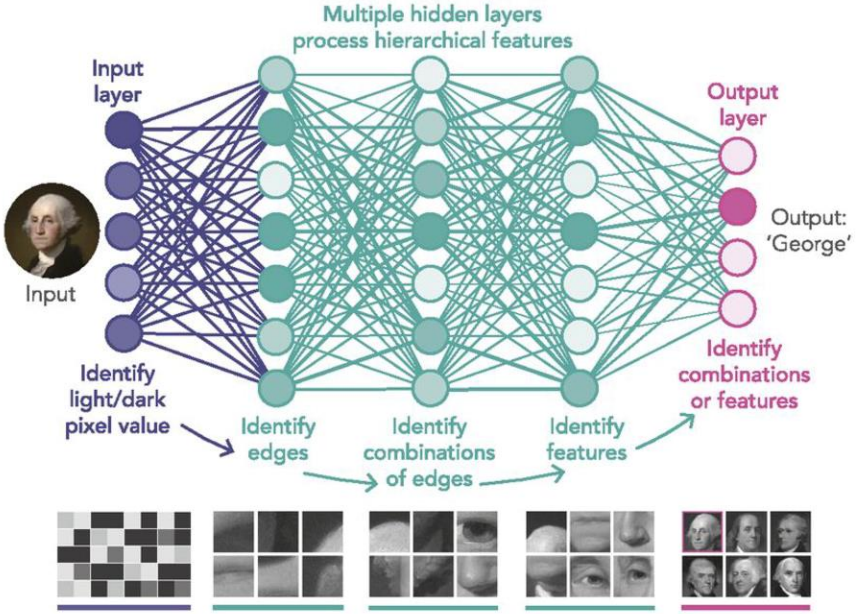
- Transparency (see the Quality Assurance Framework of the European Statistical System Eurostat or the Guidelines for Official Statistics by the US National Academies of Sciences, Engineering, and Medicine and others)
    - Statistical authorities have to document their production processes
- Accountability: Government agencies responsible for official statistics are also to be accountable for the accuracy and reliability of the data they produce (see the Fundamental Principles of Official Statistics stated by the United Nations)
- Tracing of bias or unfairness
- Quality control
- More informed use (e.g, in economic policies or public health strategies)
- Public trust

Explainability in ML

- Some algorithms are intrinsically white-box, i.e., interpretable/explainable
  - A classification/regression tree provides a clear path to the algorithm output
  - The amount of information to describe the path is of the order of the number of features (less if we prune)
- But others (ensemble methods, deep learning) are intrinsically black-box ones
  - A deep learning network may have billions of parameters to tune

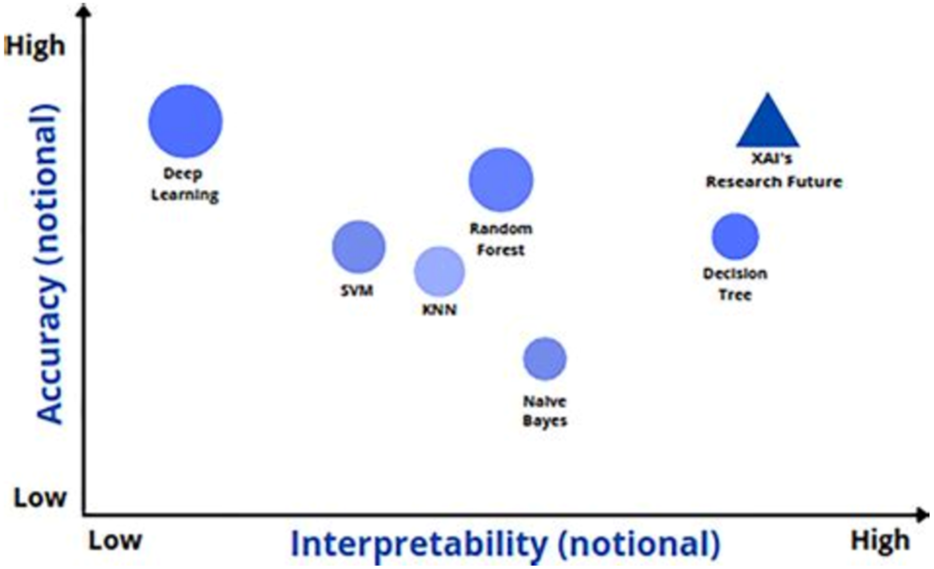


Hypothetical Classification Tree


Multiple hidden layers process hierarchical features
Input layer
Input
Output layer
Output: 'George'
Identify light/dark pixel value
Identify edges
Identify combinations of edges
Identify features
Identify combinations or features

Accuracy vs Interpretability

- We would like to achieve high accuracy and high interpretability at the same time
- Unfortunately, higher accuracy methods are also less interpretable



Which XAI methods to use

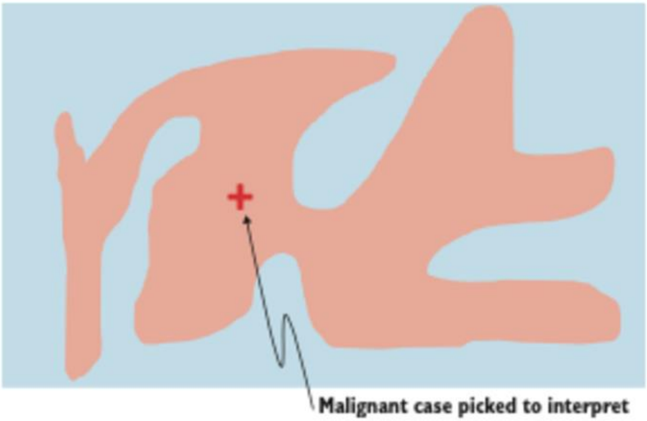
- The most widespread so far are
  - SHAP (Shapley Additive exPlanations)
  - LIME (Local Interpretable Model-agnostic Explanations)
- Both are
  - Post-hoc: first, build the model; then, explain it
  - Local: they explain prediction for a single instance
  - Model-agnostic: they are largely applicable

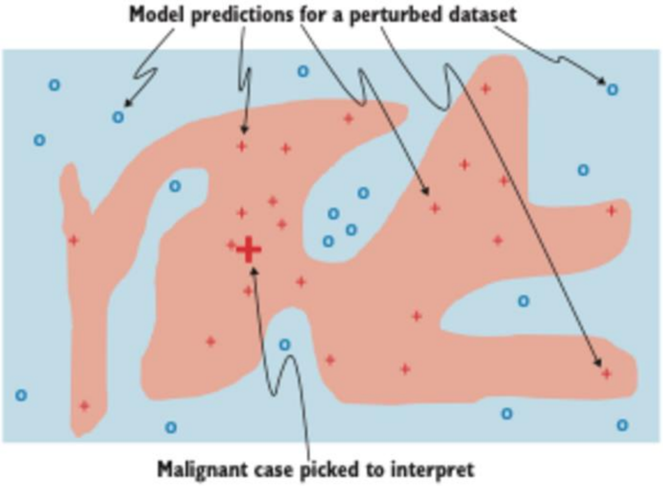
The LIME approach

- LIME relies on a linear surrogate model to identify the most relevant features in the neighbourhood of the instance of interest
- It relies on the assumption that global nonlinear boundaries are approximately linear in a small neighbourhood of the instance of interest

The LIME approach

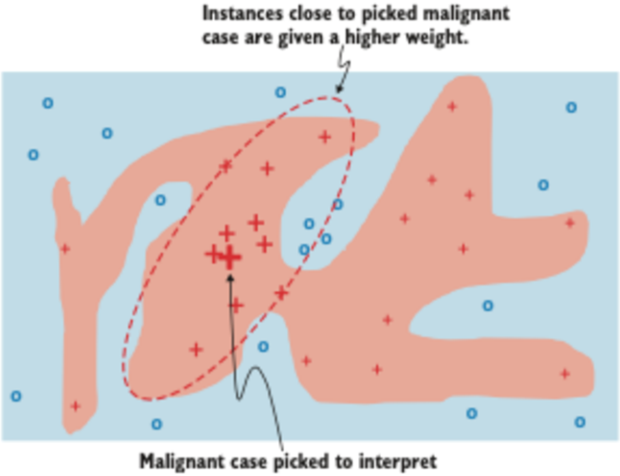
- LIME relies on a linear surrogate model to identify the most relevant features in the neighbourhood of the instance of interest
- It goes through the following steps:
    1. Create a perturbed dataset
    2. Use the model to make predictions for the perturbed dataset




Model predictions for a perturbed dataset
Malignant case picked to interpret

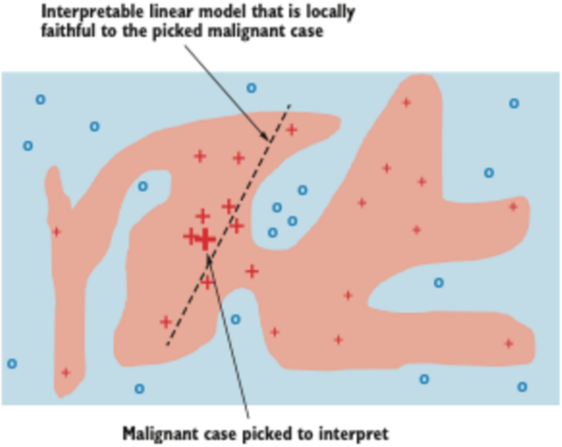
The LIME approach

- LIME relies on a linear surrogate model to identify the most relevant features in the neighbourhood of the instance of interest
- It goes through the following steps:
  1. Create a perturbed dataset
  2. Use the model to make predictions for the perturbed dataset
  3. Weigh the instances in the perturbed dataset according to their distance from the instance


Instances close to picked malignant case are given a higher weight.
Malignant case picked to interpret

The LIME approach

- LIME relies on a linear surrogate model to identify the most relevant features in the neighbourhood of the instance of interest
- It goes through the following steps:
  1. Create a perturbed dataset
  2. Use the model to make predictions for the perturbed dataset
  3. Weigh the instances in the perturbed dataset according to their distance from the instance
  4. Apply a linear model to identify the most relevant features


Interpretable linear model that is locally faithful to the picked malignant case
Malignant case picked to interpret

LIME drawbacks

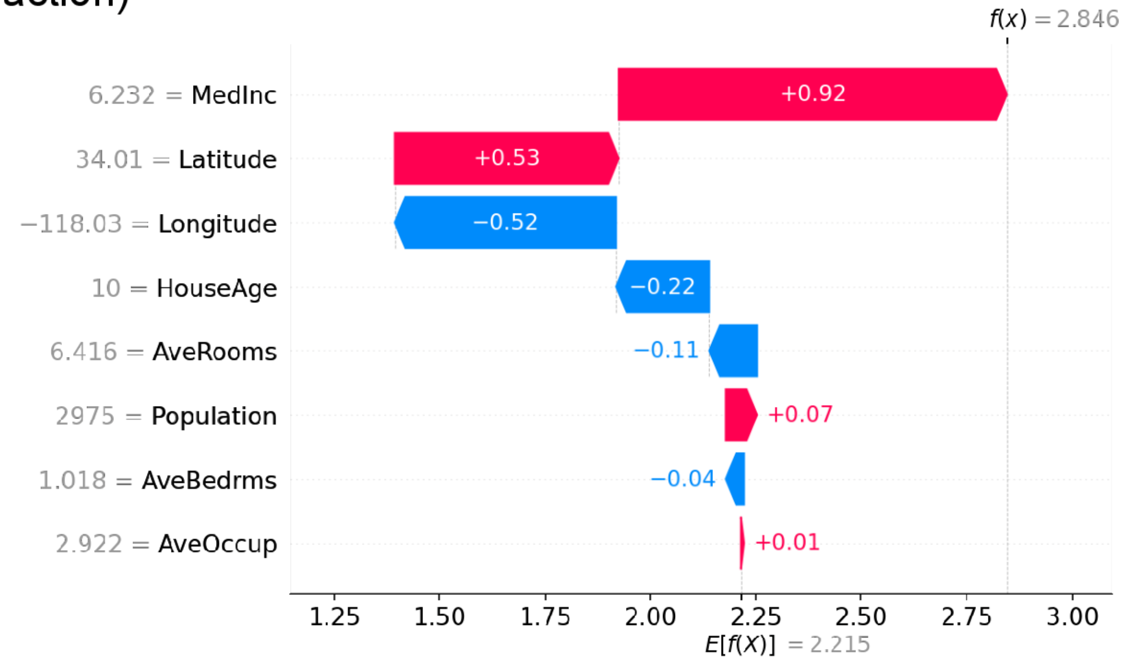
- A linear model may not be adequate
- The perturbed dataset does not account for correlations between features
- The degree of locality may lead to unstable explanations

The SHAP approach

- It is based on a game model, where a set of players form a coalition to make a decision
- We wish to assess the contribution of each coalition member to the overall decision
- The contribution is assessed by considering all the sub-coalitions where that member does not belong and see what that member adds
- The features play the role of players
- The model plays the role of the rules of the game

The SHAP approach

- It starts with the average prediction over all the instances
- Then, it considers the contribution of each feature to the actual prediction for the specific instance (SHAP values)
- However, the contribution is conditional to the features we have examined earlier (due to feature interaction)

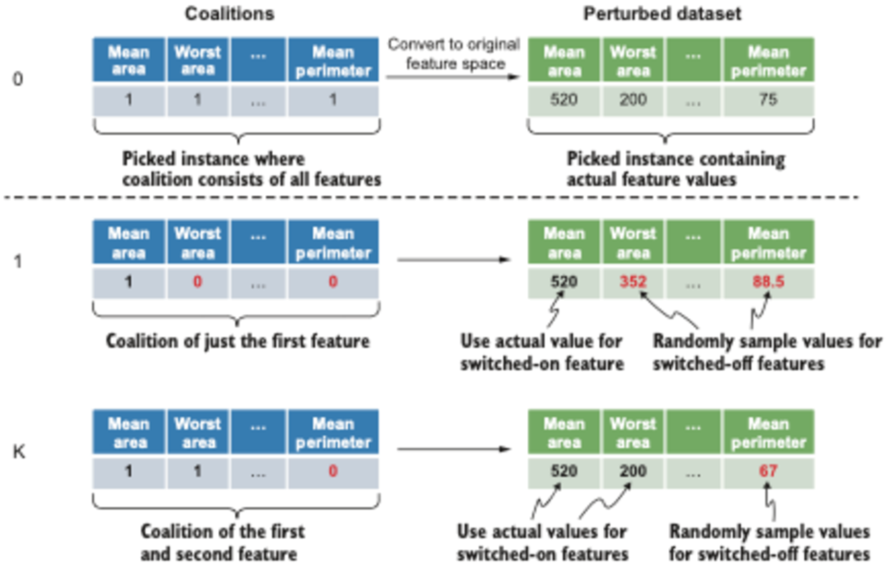


How to bypass the influence of order

- We could average over all possible orders…
- …but that means n! different sequences for a set of n features (the permutations over that set)
- For 32 features we would need to consider over 17 billion coalitions

How to reduce the computational load

- We can pick a set of randomly generated coalitions
- Each coalition is weighted according to its distance from the distance of interest
- The weight is given by the SHAP kernel function (based on the number of features, the number of those included and excluded, and the number of coalitions of the same size)
- A linear regression can then be applied to the weighted data



Coalitions

Perturbed dataset

Convert to original feature space

0

Mean area	Worst area	...	Mean perimeter
1	1	...	1

Picked instance where coalition consists of all features

Mean area	Worst area	...	Mean perimeter
520	200	...	75

Picked instance containing actual feature values

1

Mean area	Worst area	...	Mean perimeter
1	0	...	0

Coalition of just the first feature

Mean area	Worst area	...	Mean perimeter
520	352	...	88.5

Use actual value for switched-on feature

Randomly sample values for switched-off features

K

Mean area	Worst area	...	Mean perimeter
1	1	...	0

Coalition of the first and second feature

Mean area	Worst area	...	Mean perimeter
520	200	...	67

Use actual values for switched-on features

Randomly sample values for switched-off features

SHAP limitations

- Though reduced, the computational load may still be significant
- It again assesses the importance of features one at a time, while combinations of features may be more relevant

Conclusions

- We need more explainable methods
- Global explanations may be difficult to achieve or too vague
- Local explanations must be handled with care, but improvements are being sought
  - Looking for relevant combinations of features
  - Working across application domains
  - Establishing performance metrics
  - Addressing counterfactual and contrastive questions
- Happily enough, it is not a problem for the Official Statistics community only