**Transparency – Intermediate**

**Lesson 1**

* **Presentation:**

[**https://trilateralcloud.sharepoint.com/:p:/s/sociotechteam/ESvLr\_n6J8ZPiHeo2tDkSmgBzLZb10TsTHzX1Po-KVjkhw**](https://trilateralcloud.sharepoint.com/:p:/s/sociotechteam/ESvLr_n6J8ZPiHeo2tDkSmgBzLZb10TsTHzX1Po-KVjkhw)

* **Self-Reflection Exercise:** [**https://trilateralcloud.sharepoint.com/sites/sociotechteam/\_layouts/15/doc.aspx?sourcedoc={a0f6017a-136e-48da-8fb1-2634a74a3e02}&action=edit**](https://trilateralcloud.sharepoint.com/sites/sociotechteam/_layouts/15/doc.aspx?sourcedoc=%7ba0f6017a-136e-48da-8fb1-2634a74a3e02%7d&action=edit)
* **Pre-Reading (1- hour)**
  + Principles and practices of explainable machine learning:
    - <https://arxiv.org/pdf/2009.11698.pdf>
  + **For indicator types- you can review this summary table:**
    - [**https://trilateralcloud.sharepoint.com/:w:/r/sites/sociotechteam/\_layouts/15/Doc.aspx?sourcedoc=%7BDE76D6D3-37DB-463D-A775-3D48B05E43D0%7D&file=Pre-Reading.docx&action=default&mobileredirect=true**](https://trilateralcloud.sharepoint.com/:w:/r/sites/sociotechteam/_layouts/15/Doc.aspx?sourcedoc=%7BDE76D6D3-37DB-463D-A775-3D48B05E43D0%7D&file=Pre-Reading.docx&action=default&mobileredirect=true)
* **Mandatory Reading (1- hour)**
  + **Exploratory data analysis review this blog:**
    - [**https://towardsdatascience.com/an-extensive-guide-to-exploratory-data-analysis-ddd99a03199e**](https://towardsdatascience.com/an-extensive-guide-to-exploratory-data-analysis-ddd99a03199e)
  + **Model performance evaluation**
    - Choosing the right metric https://www.altexsoft.com/blog/machine-learning-metrics/
    - **Review Sklearn model evaluation metrics here:** [**https://scikit-learn.org/stable/modules/model\_evaluation.html**](https://scikit-learn.org/stable/modules/model_evaluation.html)
  + Reading: <https://christophm.github.io/interpretable-ml-book/simple.html> (chapter 5, interpretable models). This will give an overview of interpretable models (white-box) and their desirable interpretable features.
  + For some good insight on desirable features of model-agnostic explanation, read here <https://christophm.github.io/interpretable-ml-book/agnostic.html> (chapter 6).
  + How to read the xAI pyramid <https://medium.com/@ModelOriented/dalex-v-1-0-and-the-explanatory-model-analysis-419585a4ba91>
  + Review this unified implementation of explainable AI methods here: <https://github.com/salesforce/OmniXAI>
  + **Review evidently AI for machine learning model monitoring**
    - [**https://www.evidentlyai.com/**](https://www.evidentlyai.com/)
    - [**https://analyticsindiamag.com/concept-drift-vs-data-drift-in-machine-learning/**](https://analyticsindiamag.com/concept-drift-vs-data-drift-in-machine-learning/)
  + Accountability of AI Under the Law: The Role of Explanation: <https://dash.harvard.edu/bitstream/handle/1/34372584/2017-11_aiexplainability-1.pdf>

**Module 3: Applying technological concepts (2 lessons)**

Discuss how to approach transparency and explainability alongside technical development:

a. Assessing transparency and explainability requirements of the problem at hand (considering end-user needs, application, data and models used)

In order to understand our explainability requirements we need to remind ourselves of what we are trying to achieve.

Why is transparency and explainability important within technical development? When we build insights or algorithms from data, we are trying to build tools to augment the end users understanding of the situation and assist them in their decision-making process. We are very explicitly not trying to automate or replace any decision-making processes. This means that any tools we build must be transparent, understandable and explainable to the end user, if they are to effectively incorporate the outputs to help their understanding. Understanding how a tool produced a result allows the end user to effectively assess how that result fits into their understanding of the situation.

The key question to answer to gather your requirements is:

**What does the end user need to know about the insight to be able to trust, understand, and make decisions?**

The answer to this question will depend on three things:

* **Who** the end user is – what level of technical understanding or interest do they have?
* **What** insight is being provided – is it a probability from a classification model? Is it a network visualisation? Is it the binary results of a text classifier? Has there been a lot of data manipulation?
* **How** that insight is operationalised – how are they going to use the insight to help them make a decision?

Who?

It is important to understand your audience and what level of technical understanding they have, so that the explainability can be tailored to the end user needs. There is a balance to be obtained between ensuring the user has access to the information required to understand, trust and use the insights, without overwhelming the user with too much information such that they do not engage with it at all. If the end user is not using/engaging with the explainability tools, it is equivalent to not providing any explainability in the first place.

* Are they familiar with certain types of data presentation in their current work? For example, do they use percentiles, bar graphs, scatter plot? If they are not familiar with any, perhaps text is the most effective.
* Do they need to know how the model works or just the relative importance of the different inputs?

What?

What data is being used? Is the end user familiar with the data? How much context needs to be provided?

What model is being used? Is there a requirement for a simpler model?

What are the limitations of the results? What do they not take into account? How accurate is the model?

How?

How much time is the end user going to have to look at the results?

Are they trying to make comparisons? Or isolated decisions? Or just gain some context?

Is it as part of an application? Need to think about the design

b. Discuss way to perform a detailed technical breakdown of model performance and algorithmic development that assesses the performance, transparency, justifiability and accountability of your algorithms.

**Performance**

The goal of a ML model is to predict the values of a variable of interest based on the values of other variables. To assess how good is a model at making these predictions we need a criterion to evaluate its performances. In other words, we need a way to know how ‘good’ the model is. Several performance metrics have been developed over the years and standard questions that a data scientist should answer before choosing the right performance metric(s) to compute are:

* Which performance metrics are available for the specific ML task my model ought to solve?
* Given the options available and the specificities of problem at hand, which metrics are more important to assess how good is the model at solving the problem? For instance, in a binary classification scenario, are we interested in maximising the precision of a model or its recall? Which trade-off we are willing to accept?
* What is a ‘good enough’ model for the problem at hand? Which is the performance threshold below which we can consider a model performance ‘poor’ or not useful at all?

Usually, while we might be interested in choosing a model that maximises a specific performance metric, computing other relevant performance metrics is useful to better understand the model behaviour.

Reading: https://www.altexsoft.com/blog/machine-learning-metrics/

**Transparency and explainability**

If we want to develop an explainable model, looking only at maximising performance is not enough; we should be able to understand (and explain) why the model achieved that performance, which means investigating why the model made certain predictions and inferences.

A wide range of techniques and metrics that describe the model behaviour exists since different models solve tasks in different ways and different stakeholders have different explainability needs. Some key questions we need to ask to define our approach to explainability are:

* How much explanation my model requires (how comprehensible is the model in the first place)?
* Which granularity of explanation is required?
* What do I need the explanation for?

*Explanation by Model Comprehensibility*

Before deciding on the approach to evaluate a model and understand its performance we should consider how comprehensible is the model in the first place. Having a too few explanations might result in a poor understanding of the model, on the other hand, computing all the possible performance and explainability metrics available for our problem might overwhelm a stokehold with information and be computationally and time inefficient. Is it common to distinguish models in ‘white box’ and ‘black box’ models. White box models are simpler models (examples are linear regression, decision trees) with an observable and understandable input-output relation. Black box models are more complex and lack clarity on the inner process that leads from inputs to outputs. White box models are easier to explain and usually their explanation relies on the model inner parameters. Black box models usually require more explanations and therefore computation of a number of metrics that can capture different aspects of their inner working. Several explainability metrics are model-agnostic and can therefore be applied to any (white and black) box model while other are specific one of the two models’ categories or even to specific models (model-specific metrics).

Reading: <https://christophm.github.io/interpretable-ml-book/simple.html> (chapter 5, interpretable models). This will give an overview of interpretable models (white-box) and their desirable interpretable features. For some good insight on desirable features of model-agnostic explanation, read here <https://christophm.github.io/interpretable-ml-book/agnostic.html> (chapter 6).

*Granularity of the explanation*

In terms of granularity of explanation, the main categorisation is into:

* Global explanations: explanations that describe model behaviour on the whole data set. This allows us to understand how the model behaves on average. The overall model performance falls under this category and it provides a first summary of the model behaviour. Going deeper in our global investigation some questions we can ask are: which variables are important to the model? How a single variable affects the average prediction? How the variables interact and how their interactions affect the prediction? How the model generalise, on average, on unseen data? How the model performance varies among different sub-groups of the dataset?
* Local explanations: refer to the model prediction of a single instance/data point. Local explanations try to explain how variables affect specific predictions. Some of the questions we should ask at global level, are questions we can address at local level to: which variables contribute more to the prediction of a specific instance? How do they and their interactions affect the instance prediction? How are the performance of the model on instances similar to the specific instance?

Diagram

Description automatically generated

The “explainability pyramid” above [ref <https://ema.drwhy.ai/>] provides a general

framework for investigating a model global and local behaviour. Questions we should address to assess the model behaviour on single instances are provided on the left, on the right the questions refer to the global model behaviour. Questions of the top are more general and can be usually answered with one or few metrics, the lower we go in the pyramid, the more specific the questions become, and we might require several metrics and wider context to address them in a satisfactory way. In evaluating our model, it is common to obtain general (top of the pyramid) global and local metrics, however, depending on the requirements of our problem, we might be more interested in exploring more deeply only one of the two types of behaviour.

Reading: <https://medium.com/@ModelOriented/dalex-v-1-0-and-the-explanatory-model-analysis-419585a4ba91> (a more thorough explanation of the xAI pyramid)

*Reasons for explanation*

The choice of the explainability metrics to compute should also be guided by questions around what I want or need to explain and for which reason. Common use cases are:

* Biases in the prediction
* Importance of different features
* Correlations among features
* Insights on causality relationships
* Metrics resulting from monitoring the performance of the model with respect to changes in input data or parameters.

These explanations are useful not only to ensure stakeholders trust in the model but also to allow modeller to more easily identify problems with the model predictions and behaviours so that mitigations and corrective actions can be taken.

**Relation with justifiability and accountability**

The goal of justification is to convince that a decision is “good” or appropriate. Though related concepts, justifiability and explainability are different: we can understand how a model produced certain outputs, but we might not think these outputs are good (and vice versa). As with explainability, justifiability of a model output can be provided at global or local level. The assessment of whether a model outcome is good lies outside the algorithmic development, however, explainability metrics provide useful means to experts and decision makers to make this assessment.

When a model makes a ‘wrong’ (e.g., harmful) decision, it is important to know who should be accountable for this; this is a non-technical decision, however, explainable metrics can help identifying what caused the failure and to correct the errors.

Technical considerations to make when investigating model performances to support the justifiability and accountability assessment include:

* Document each step of model development, from input data collection to output delivery
* Understanding and communicating biases and limitations of the model
* Make the model outputs reproducible to external users
* Assure compliance with best coding practices in the developers' team
* Perform extensive technical testing of the model
* Consider cyber-security requirements

Readings: This article provides a list the technical considerations that must be considered if we desired AI systems that could provide kinds of explanations that are currently required of humans under the law. https://dash.harvard.edu/bitstream/handle/1/34372584/2017-11\_aiexplainability-1.pdf

c. Provide an overview of quantitative metrics and indicators available (e.g., feature importance analysis, data provenance, SHAP values) in a supervised machine learning context.

In this section, we aim to provide an overview of what indicators are available in the following categories:

* **Exploratory data analysis**
* **Bias**
* **Feature Importance**
  + Through intrinsic model parameters
  + Permutation Feature Importance (PFI)
  + Individual Conditional Expectation (ICE)
  + Partial dependence plots (PDP)
  + Accumulated Local Effects (ALE)
* **Model performance monitoring**
  + Prediction metrics
  + Bias
  + Production Model Monitoring via Evidently AI.
    - Data quality
    - Data drift
    - Target drift
* **Model Explainability**
  + LIME
  + SHAP
  + Anchors
  + Counterfactuals
  + Integrated Gradients

**Exploratory data analysis:**

This covers indicators that help you to exhaustively analyse your data. For instance, we might be interested in understanding the distribution of our data; this can include questions like is a specific feature in our data normally distributed? Or in a classification scenario, are the target labels imbalanced? Etc. We might also be interested in assessing how different features are correlated with one another in our dataset. For a machine learning model, we would ideally like to remove redundant features, and this is where we can look at correlation matrixes to remove features that are redundant or not too helpful for our model. Exploratory data analysis can also entail assessing data quality of your dataset to determine suitability for machine learning. This phase is critical for successful machine learning model development. Data quality checks can include looking for missing values, duplicates, detecting outliers in dataset and various profiling statistics that can help you to explore the distribution of your data.

**Feature importance:**

This entails answering questions like *what matters most to the model and how much does it matter for it to make a decision?* Feature importance's include an array of methods that can shed light on the overall importance of features and how likely- individually or through combinations do they impact a model's outcome. For transparent models like decision trees and linear models, we can use the models' intrinsic parameters to determine which features are useful for the model to make predictions. In other cases, methods exist such as permutation feature importance (PFI) that can rank features by observing the increase in prediction errors through shuffling features in an iterative fashion. In addition to this, some other methods like the partial dependence plots (PDP) and Individual Conditional Expectation (ICE) help to visually investigate the relationship of a feature with a target variable by computing the marginal contribution (average for the former and individual for the latter) across all features. An improvement to these methods is feature importance via Accumulated Local Effects (ALE) which does not assume feature independence and helps to better visualise the importance of features with the target variable.

**Model performance monitoring:**

This includes a selection of indicators that can help to assess the model performance. Here we are interested in answering questions like

* *how well does the model generalize to unseen data instances?*
* *Is the model biased in its predictions towards certain cohorts, for example cohorts which are underrepresented etc?*
* *How much does the distribution of new unseen data vary with data distributions on which the model was previously trained on?*
* *How much do predictions vary between current/re-trained machine learning model and an existing (benchmarked) model that has previously worked well in production.*
* *How much do predictions vary with respect to different features and target class distributions?*

For evaluating model performance in a supervised machine learning context, there are standard classification and regression metrics available, that can quickly identify the model performance to your test data. For classification, the main metrics include accuracy, precision, recall, F1 score and confusion matrices- which are highly used to diagnose model performance. Two concepts implemented in the sklearn library include the macro and micro averaging for each of these metrics to monitor the performance for your classification model. The macro-averaging technique with these metrics is insensitive to the imbalance of the classes and treats them all as equal- giving more accurate reflection of the model performance on imbalanced datasets. For regression, the mean squared error, mean absolute error and R-squared statistic are used to determine the goodness of fit of the model to unseen data.

For bias, we can initially look at the classification and regression metrics for each sub-group/cohorts of our population. This entails identifying groups with certain characteristics and then computing the classification/regression metrics for each of these sub-groups depending upon the problem statement. Additional fairness metrics include demographic parity, selection rate, equivalised odds for classification tasks and bounded group loss for regression. Bias in model can also be observed in a post-hoc manner by visually inspecting predictions for your cohorts.

In addition to the above, when a model has been launched in production, it is important to monitor performance on unseen and new data instances. The indicators available here help to quantify data drift and target/concept drift. Data drift seeks to answer questions on how distribution of the features in the data vary by quantifying the similarity of the existing feature distributions in the data with feature distributions when new data instances that get added to the existing data. The dataset used as a reference to compare new data instances against is called the benchmark/reference dataset.

The Target/Concept Drift helps detect and explore changes in the target variable and predictions made by the machine learning model between the current dataset (with new data instances added) and a benchmark/reference dataset. Target drift helps to investigate change in predictive power of machine learning models. During training, the model learns a function that maps the target variable, but over time, it can unlearn them or is unable to use these patterns in a new and dynamic environment as more data samples are added- target drift helps to capture this. It can also be used to investigate the distributions of correlations between individual features and target variable. This can indicate how feature correlations change with the target variable as new data is available. Some tests that help to quantify the similarity between two data distributions include Kullback-Leibler divergence, Jensen-Shannon distance, Population Stability Index etc. Evidently AI is a great platform for model performance management, incorporating all these metrics to in a dashboard format.

**Model Explainability/Interpretability:**

Indicators here answers questions on

* *How do I explain the decisions made by the model to different stakeholders?*
* *How can we be confident in the predictions made by the model?*

The indicators here come from different explainability approaches that aim to understand model behaviour locally (for each individual prediction) and globally. There are a lot of methods available, and some rely on looking at model parameters intrinsically to draw conclusions, whereas the other techniques follow a post-hoc approach. The post-hoc approaches in explainability can be model specific and model agnostic. Some model specific approaches include integrated gradients, tree SHAP, Linear SHAP, Deep SHAP. Some model agnostic techniques include Shapely values, LIME, Anchors, Kernel SHAP, Counterfactuals. Some of these approaches like Shapely, SHAP, have solid mathematical foundations in cooperative game theory that make them a reliable model interpretability method.

d. Analysing issues of transparency and explainability in the police and security domain through concrete examples and exercises drawn from CESIUM product, which includes a custom user-facing xAI component

Here we will present examples and exercises drawn from the CESIUM product. All the exercises are based on lengthy discussions and feedback that took place with the end user. This emphasises the value and importance of co-design within explainability.

SCRIPT FOR CESIUM DEMO (already exists, many have been done)

**Example 1: Classification algorithm from CESIUM**

Aim: present the results of a classification model designed to support the identification of children who may be at risk to child exploitation

Assessing requirements

Who?

Safeguarding professionals at Lincolnshire police and children's social care. Intelligent and engaged. No technical background. Familiar with percentiles, quantifying risk via numbers or colours. Not familiar with machine learning algorithms, histograms

What?

End users are very familiar with the dataset

Technical details of the model not necessary

Need to know what information was used as inputs by the model and what information was most important in contributing to the result, e.g., why is an individual's score especially high

Probabilities from a classification algorithm

The model is learning patterns based on past decisions. The information that has been considered most relevant of that available has been selected as inputs to the model. As with any model, it can get things wrong. There will be cases where the factors that suggest vulnerability are not included in the model, as such the model would not be able to identify this. There will also be cases that do not follow the standard pattern of risk, which will be also difficult for a model to identify.

How?

Prioritise and review cases of potential child exploitation risk

Help guide which children might be important cases to review

Support a case review of an individual by providing insights and reasoning

Firstly, we need to state what question the model trying to answer and what is the result

How likely is John Smith to be involved in a MACE meeting at this time?

**Exercise 1:** Suppose John Smith has a model output of 0.78 (which is the 95th percentile). Discuss the relative merits of the following presentation of this result

* John Smith has a model output of 0.78
* The likelihood for John Smith to be involved in a MACE meeting is 0.78
* John Smith is in the top 5% of children for referral to MACE
* John Smith is in the 95th percentile of children for referral to MACE

Possible discussion points

* Using the phrase ‘model output’ does not contextualise the result for the end user
* Using the output of the model does not provide a comparison of the result compared to other people that the police need to prioritise against.
* The end user is interested in which children they should focus their time and resources on, as such as score relative to other children would be more beneficial to their operations. Does the model output have any operational value in isolation?
* It is useful for big numbers to mean higher risk.

Visualise the result

* Gives a multi-model approach to the explainability and allows end users to understand the result in different ways
* Providing a visualisation helps to make it obvious that we are comparing an individual's score to all other children and the result corresponds to their position
* Different people might understand the sentence better or the visualisation better, so both will help to provide understanding to people who think in different ways
* You can see this distribution of scores, so whether there are peaks in certain areas and whether the individual you are looking at fall in a peak

**Exercise 2:** Discuss the relative merits of the two different designs of visualisation

Chart, histogram

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Possible discussion points:

* End users not comfortable with histograms
* Pictograms are more common in everyday life
* Pictogram connects the fact that the area represents people
* Colour helps clarify the point
* Statement on the visualisation helps make the point

Interactive feature importance and data provenance

We need to ensure that it is clear what data went into the model (transparency of data provenance) and what data was important in determining the model result (explainability). The feature importance

* Without selecting anything it is clear what data went into the model and how important it was in contributing to the score
* Details about the data provenance and contextualisation of the data can be explored when selecting any one feature

Chart

Description automatically generated with medium confidence

**Example 2: Classification algorithm for violent and sexual content in text**

Aim: present the results of two NLP classification models designed to make it easier and quicker for the reader to identify important text relating to vulnerability of the child

Assessing requirements

Who?

Safeguarding professionals at Lincolnshire police and children's social care. Intelligent and engaged. No technical background. Familiar with percentiles, quantifying risk via numbers or colours. Not familiar with machine learning algorithms, histograms

What?

Binary results from a text classifier

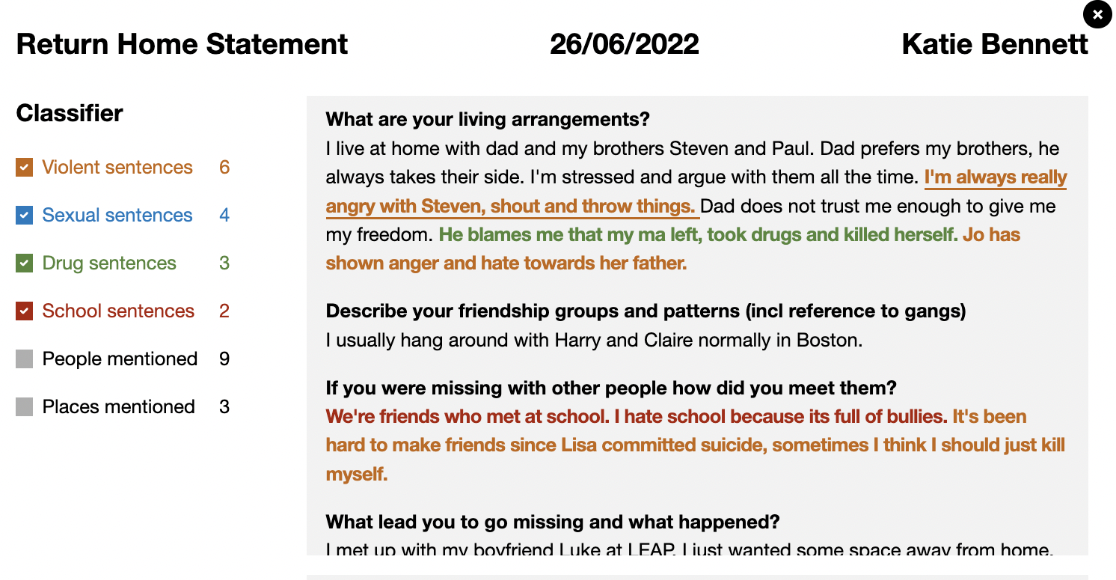
Need to know what information was most important in contributing to the result, e.g., why a sentence was flagged as sexual

As with any model, it can get things wrong - it will flag sentences that are not relevant, and it will miss sentences that are relevant.

How?

Speed up the review of an interview with a missing child to support the identification of any significant vulnerability factors of importance

**Exercise 3:** Discuss the relative merits of explaining the results of a text classifier using visual explainability (a bar chart to give relative importance of each word), compared to textual explainability (a sentence listing the 3 most important classifying words in the sentence).



Possible discussion points:

* Bar chart has more detail – it is a more complete explanation
* As we discussed before, it is not about giving the most amount of explanation, but giving it at the right level
* Textual explanation is much more intuitive and easier to understand
* A bar chart for each sentence can be confusing and overwhelming
* Textual explanation provides the most important information succinctly
* Textual explanation does not give the whole picture

**Reflection sheet**

**References**

[ ] Henin, C. and Le Métayer D.. Beyond explainability: justifiability and contestability of Algo-  
rithmic Decision Systems. AI & Society: Knowledge, Culture and Communication, Springer Verlag, 2021, 10.1007/s00146-021-01251-8. hal-03165232

**[ ]** Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). christophm.github.io/interpretable-ml-book/

[ ] Biecek, P. and Burzykowski, T., 2021. *Explanatory model analysis: Explore, explain and examine predictive models*. Chapman and Hall/CRC.

Molnar, C. (2019). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). christophm.github.io/interpretable-ml-book/

Jacovi, Alon, and Yoav Goldberg. "Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?." arXiv preprint arXiv:2004.03685 (2020).

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations." Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1. 2018.

Lundberg, S., Allen, P. and Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. [online] Available at: <https://arxiv.org/pdf/1705.07874.pdf>.

Shapley, L.S. (1952). *A Value for N-Person Games*. [online] [www.rand.org](https://ukc-word-edit.officeapps.live.com/we/www.rand.org). Available at: <https://www.rand.org/pubs/papers/P295.html> [Accessed 22 Sep. 2022].

‌

Štrumbelj, E. and Kononenko, I. (2013). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3), pp.647–665. doi:10.1007/s10115-013-0679-x.

Wu, T., Ribeiro, M., Heer, J. and Weld, D. (2021). *Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models*. [online] pp.6707–6723. Available at: <https://www.cs.cmu.edu/~sherryw/assets/pubs/2021-polyjuice.pdf> [Accessed 22 Sep. 2022].

Masís, S. (2021). Interpretable Machine Learning with Python: Learn to build interpretable high-performance models with hands-on real-world examples. Packt Publishing Ltd.

Han, X., Wallace, B. C., & Tsvetkov, Y. (2020). Explaining black box predictions and unveiling data artifacts through influence functions. arXiv preprint arXiv:2005.06676.

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K. R. (2019). Layer-wise relevance propagation: an overview. Explainable AI: interpreting, explaining and visualizing deep learning, 193-209.

Sundararajan, M., Taly, A. and Yan, Q. (2017). Axiomatic Attribution for Deep Networks. *arXiv:1703.01365 [cs]*. [online] Available at: <https://arxiv.org/abs/1703.01365>.

‌

Mucha, M. (2020). Open the Black Box: Understand What Drives Predictions in Deep NLP Models. [online] Medium. Available at: <https://towardsdatascience.com/open-the-black-box-understand-what-drives-predictions-in-deep-nlp-models-833f3dc923d0> [Accessed 22 Sep. 2022].

Shrikumar, A., Greenside, P. and Kundaje, A. (2019). *Learning Important Features Through Propagating Activation Differences*. [online] Available at: <https://arxiv.org/pdf/1704.02685.pdf>.

‌

‌ Ancona, M., Ceolini, E., Öztireli, C. and Gross, M. (2022). *Towards better understanding of gradient-based attribution methods for Deep Neural Networks*. [online] openreview.net. Available at: <https://openreview.net/forum?id=Sy21R9JAW> [Accessed 22 Sep. 2022].

Lal, G.R. (2021). *Explainable Neural Networks: Recent Advancements, Part 3*. [online] Medium. Available at: <https://towardsdatascience.com/explainable-neural-networks-recent-advancements-part-3-6a838d15f2fb> [Accessed 22 Sep. 2022].

Iwanaga, T., Usher, W., & Herman, J. (2022). Toward SALib 2.0: Advancing the accessibility and interpretability of global sensitivity analyses. Socio-Environmental Systems Modelling, 4, 18155. doi:10.18174/sesmo.18155

Herman, J. and Usher, W. (2017) SALib: An open-source Python library for sensitivity analysis. Journal of Open Source Software, 2(9). doi:10.21105/joss.00097

‌