TRUSTWORTHY AI CASE STUDY

Firstname1 Lastname1 & Firstname2 Lastname2
Amsterdam Business School
University of Amsterdam
1018TV Amsterdam, the Netherlands
email@student.uva.nl and email2@student.uva.nl

ABSTRACT

Here goes a short abstract.

1 Introduction

This section of the report should consist of a short introduction to trustworthy AI and the context of your dataset. Also provide an overview of what will be discussed in the following. This report should be written in the style of a conference paper, meaning formal language and correct citations. Here are examples on how to cite: Barredo Arrieta et al. [2020] provide an overview of XAI methods. Many more review reports on XAI have been written in recent years [*e.g.,* Ali et al., 2023, Guidotti et al., 2018]

2 Methods for explainable and interpretable AI

This section is reserved for an overview and explanation of the three methods that you chose. Each method will be in a separate sub-section where you explain the core essentials and the main steps of each method.

You should choose three different methods. You should choose at least one glass-box model and at least one post-hoc explanation method. Two methods should be taken from the course material, *i.e.,* you should choose at least two methods that we discussed in class. The third method you may choose freely and try out other methods not discussed in class, however this is not a requirement.

Throughout the entire report you're free to add subsections or paragraphs as you see fit.

2.1 Method 1

Discuss your chosen method 1. You may rename this section according to the method you chose. Make sure to also cite the relevant material for each method.

2.2 Method 2

Discuss your chosen method 2. You may rename this section according to the method you chose. Make sure to also cite the relevant material for each method.

2.3 Method 3

Discuss your chosen method 3. You may rename this section according to the method you chose. Make sure to also cite the relevant material for each method.

3 Modeling

This section should contain everything related to the modeling stage.

3.1 Data description and processing

Provide a description of the dataset. Also include summary statistics and processing steps you take.

3.2 Black-box model

Report on your choice of black-box model you fit to the data. Also report any considerations and steps in the modeling process. Remember to also include an assessment and evaluation of the performance of your chosen model (you may do that here or in Section 4..

3.3 Modeling choices for explainability methods

If the methods you chose require any parameters you should specify your choice for these.

4 Results

While the previous section was reserved for an overview and general explanation of each method, this section will report the results of each method.

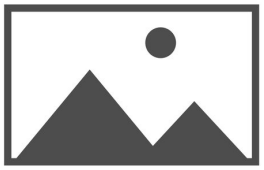

Figure 1: Caption

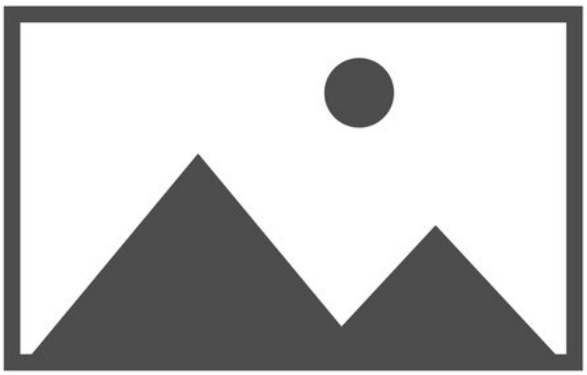

Figure 2: Caption

4.1 Results method 1

For example, you could provide a visualisation of some sort. In Figure 1 I am showing a placeholder image as example.

4.2 Results method 2

In the results section, results are reported and described. This means that you provide a description of the results and what certain plots show. You can also interpret the results of each method *in isolation*. That means that there is no comparison (yet) of the methods. That will come in Section 5.

4.3 Results method 3

To include an image that spans the entire page like Figure 2, you can use `\begin{figure*}` instead of `\begin{figure}`.

Similarly, for tables you can use `\begin{table}` for a table that will be fit into one column or `\begin{table*}` for a wide table spanning both columns.

5 Discussion

This section is reserved for a discussion of the different methods and their outputs. Evaluate and compare each method and its results you obtained for your dataset. Which method is useful for your dataset? Also provide a general discussion of the pro's and con's of each method and possible use cases they would fit to. Discuss general points about the methods you chose and make it concrete using your dataset (or vice versa – start with concrete results of your dataset and then move toward a general discussion).

Paragraph. You may want to use `\paragraph{}` to create in-line headings for paragraphs.

End your report with a discussion of the socio-ethical context of trustworthy AI and your specific dataset.

6 Conclusion

Provide a summarizing paragraph of this report.

References

S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805, 2023. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.101805. URL `https://www.sciencedirect.com/science/article/pii/S1566253523001148`.

A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2019.12.012. URL `https://www.sciencedirect.com/science/article/pii/S1566253519308103`.

R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51 (5), aug 2018. ISSN 0360-0300. doi: 10.1145/3236009. URL `https://doi.org/10.1145/3236009`.

A Additional information

The appendix comes after the references and contains additional material that may support discussions in your report but are not essential for the report to be complete. Anything that is essential should go in the main part of the report.