Text

Description automatically generated

**Master Degree in Data-Driven Marketing,**

**with a specialization in Data Science for Marketing**

Explainable AI

Ludovico Toscano 20220044

**Supervisor:** Professor/Prof. Fernando Bação

**Context:**

One of the problems of AI is that there is a trade-off between the explainability and performance of the model, in recent years some solutions have come out to try to solve this issue. Explainability is essential for many reasons, like gain of knowledge, improving the model itself, understanding if the model is making a choice in the right way or not, and so on.

**Research gap and objectives:**

As written in this paper [1] most of the study in the field of explainable ai are related to models that are tested on text or image in static datasets while reinforcement learning work on datasets that changes at each state. Another difference in reinforcement learning is the performance metric used which is usually an ad-hoc metric. The objective is to produce a review of existing interpretability methods and build an example based on reinforcement learning deep network.

**Methodological approach:**

To achieve the objective of building an example based on reinforcement learning, we will try to explain the result of the model using some of the best and most used methods such as SHAPE, LIME and Integrated Gradients according to [1,2,3,4]. We will use the Python implementation of the three algorithms [5,6,7].

**Expected results and contributions:**

The main contribution will be to have an updated review of the explanation methods related to deep networks, which will also include an example based on reinforcement learning which is not the focus of the studies already available.

**Bibliographical references:**

[1] Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine

Learning Interpretability Methods. Entropy 2021, 23, 18. <https://dx.doi.org/10.3390/e23010018>.

[2] Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why should I trust you?” Explaining the predictions of any classifier. In Proceedings of the

22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August

2016; pp. 1135–1144.

[3] Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the Advances in Neural

Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.

[4] Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference

on Machine Learning, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 3319–3328.

[5] <https://shap.readthedocs.io/en/latest/index.html>

[6] <https://github.com/marcotcr/lime>

[7] https://www.tensorflow.org/api\_docs/python/tf/gradients