Artificial Intelligence Ethics: 5th lesson – Model Cards

A model card is a short document that provides key information about a machine learning model. Model cards increase transparency by communicating information about trained models to broad audiences.

In this tutorial, you will learn about which audiences to write a model card for and which sections a model card should contain. Then, in the following exercise, you will apply what you have learned to a couple of real-world scenarios.

Model cards:

Though AI systems are playing increasingly important roles in every industry, few people understand how these systems work. AI researchers are exploring many ways to communicate key information about models to inform people who use AI systems, people who are affected by AI systems and others.

Model cards are one way for teams to communicate key information about their AI system to a broad audience. This information generally includes intended uses for the model, how the model works, and how the model performs in different situations. You can think of model cards as similar to the nutritional labels that you find on packaged foods.

***Who is the audience of your model card?***

A model card should strike a balance between being easy-to-understand and communicating important technical information. When writing a model card, you should consider your audience: the groups of people who are most likely to read your model card. These groups will vary according to the AI system’s purpose.

For example, a model card for an AI system that helps medical professionals interpret x-rays to better diagnose musculoskeletal injuries is likely to be read by medical professionals, scientists, patients, researchers, policymakers and developers of similar AI systems. The model card may therefore assume some knowledge of health care and of AI systems.

***What sections should a model card contain?***

Per the original paper, a model card should have the following nine sections. Note that different organizations may add, subtract or rearrange model card sections according to their needs:

* Model details

Include background information, such as developer and model version.

* Intended use
* What use cases are in scope?
* Who are your intended users?
* What use cases are out of scope?
* Factors

What factors affect the impact of the model? For example, the smiling detection model's results vary by demographic factors like age, gender or ethnicity, environmental factors like lighting or rain and instrumentation like camera type.

* Metrics

What metrics are you using to measure the performance of the model, and why did you pick those metrics?

* Classification systems

In which the output is a class label, potential error types include false positive rate, false negative rate, false discovery rate, and false omission rate. The relative importance of each of these depends on the use case.

* Score-based analysis

In which the output is a score or price, consider reporting model performance across groups.

* Evaluation data
* Which datasets did you use to evaluate model performance? Provide the datasets if you can.
* Why did you choose these datasets for evaluation?
* Are the datasets representative of typical use cases, anticipated test cases and/or challenging cases?
* Training data

Which data was the model trained on?

* Quantitative analysis

How did the model perform on the metrics you chose? Break down performance by important factors and their intersections. For example, in the smiling detection example, performance is broken down by age (e.g., young, old), gender (e.g., female, male), and then both (e.g., old-female, old-male, young-female, young-male).

* Ethical considerations

Describe ethical considerations related to the model, such as sensitive data used to train the model, whether the model has implications for human life, health or safety, how risk was mitigated, and what harms may be present in model usage.

* Caveats and recommendations

Add anything important that you haven’t covered elsewhere in the model card.

***How can you use model cards in your organization?***

The use of detailed model cards can often be challenging because an organization may not want to reveal its processes, proprietary data or trade secrets. In such cases, the developer team should think about how model cards can be useful and empowering, without including sensitive information. Some teams use other formats, such as FactSheets, to collect and log ML model information.