Deloitte.

The new AI Fairness Paradigm

Deloitte's AI Ethics offering: Model Guardian

Deloitte's AI ethics solution "Model Guardian" helps clients test their machine learning models for fairness alongside accuracy.

The Need

Machine Learning [ML] models bring unprecedented accuracy and efficacy to challengingly complex optimization problems. In contrast to rules-based models, ML models derive the rules directly from the data... the more, the better (generally). While the foundation in data spares ML models from risks around assumptions and other pre-conceived notions, it does not make them invulnerable.

If built on non-representative data, ML models can perpetuate human prejudice just as the human-designed rules can skew classical models. Just as AI gains its strength from harvesting deep and wide datasets, a poor dataset will condemn an AI to inadequacy, risking to mislead rather than to inform. Skewed data injects subjectivity into an otherwise objective decision-making process.

This is doubly harmful, both to the organizations using AI as a decision-making aid, as well as to the customers or constituents of those organizations subject to the outcomes of those decisions. Bias may influence an organization to turn down profitable business. A potential customer may not obtain a much needed product or service on time.

As AI becomes more prominent across all industries, lawmakers and regulators are increasingly taking interest and an active role in shaping an AI-enabled future. They are keenly aware of the bias risk, issuing cautionary guidance or even passing into law new consumer protections, such as those enshrined in the General Data Privacy Regulation. The European Commission has, for example, set expectations for AI-enabled systems to satisfy certain trustworthiness requirements before being deployed at scale.

Our Solution: Model Guardian

Deloitte's AI ethics solution Model Guardian adds "fairness" alongside accuracy to judge the adequacy of machine learning models. It is versatile, able to handle models of any variety. It is intuitive, providing a step-by-step guide to the user on how to identify and quantify bias.

Model Guardian is built on the understanding that bias can take many forms and derive from several sources. Forms include intentional bias (prejudice set in policy) and unintentional (unwitting under-representation of a target market due to lack of historical experience). Sources span from raw data to the subset selected for training a ML model to the design of the model itself (algorithm, hyper-parameters). As such, bias can be tricky to isolate… and to properly measure, with different metrics applicable by situation.

Model Guardian starts with universal bias detection, examining source data, training set and model predictions for over- and under-representation of protected classes within the variables, or features, of the model. It then progresses into use-case specific metrics, for example credit or hiring decisions, leveraging metrics familiar to practitioners in those areas to explain fairness risk.

Beyond risk management, Model Guardian serves as a useful companion to the model developer. It provides the developer with a means to evaluate successive iterations of the model under development in the context of both accuracy and fairness.

AI Ethics is both an important topic as well as a difficult one to manage. Several prominent organizations have struggled to establish enduring control mechanisms. Model Guardian provides a means to facilitate this crucial task at both practitioner and management levels.

Advantages/Benefits

- Companies may deploy sophisticated ML models with confidence, aware of the degree of bias within their decisioning tools.
- They can demonstrate to regulators that models have been carefully designed to minimize unwanted discrimination against projected classes.
- Practitioners can build better models, focusing discriminatory power on permissible features that are good predictors of desired outcomes.
- Companies may find new growth opportunities into areas where certain classes had been historically under-represented or otherwise disadvantaged, simply due to lack of data history on which to build decisioning tools.

Example Use Cases

- Preventing prejudice in granting of loans/ credit decisions.
- Ensuring objectivity around hiring or promotion decisions.
- Demonstrating inclusiveness for public entities, such as university admissions.
- Optimizing discriminatory power for permissible model attributes.

Contacts

David Thogmartin
Leader aiStudio
dthogmartin@deloitte.de

Issue 05/2020