OPERATIONALIZING RISK MANAGEMENT FOR MACHINE LEARNING: BUILDING A PROTOCOL-DRIVEN SYSTEM FOR PERFORMANCE, EXPLAINABILITY, & FAIRNESS

Imran Ahmed, Giles L. Colclough, Daniel First*, & QuantumBlack contributors¹

*Corresponding author: daniel.first@quantumblack.com | QuantumBlack, 1 Pall Mall E, London, UK

Teams of data scientists are often required to build machine learning models at a rapid pace. This speed of execution, together with the complexity of machine learning techniques and the novelty and progression of the field as a whole, introduces a large degree of risk to these projects. At QuantumBlack, we have found three areas are especially prone to risk for teams building models: *performance, explainability*, and *fairness*. Here, "risk" refers to potential mistakes that practitioners can make that result in adverse outcomes in any of these three areas. Teams can make mistakes that critically affect the ability of their model to: maintain strong performance upon deployment, provide sufficient explainability to satisfy regulatory or business requirements, or avoid discriminatory biases against minority groups.

To help data scientists and other practitioners identify and mitigate these risks, we introduce a *comprehensive protocolbased risk management system for machine learning*. This system, built from our collective experience running hundreds of analytics projects across industries over the last ten years, enables data science teams to access best practices for identifying and overcoming risk, in a systematic and organised way. Teams can also share and read failure and success stories for each risk. A webapp interface (figure 1) increases the accessibility and usability of the risk management system. Each category of risk has a multimedia introductory page that outlines the topic at a high level (figure 2).

We organise knowledge about risk via a "protocol." This protocol breaks down the machine learning modelling process into over 30 high-level "activities" (for example, Define the Analytics Approach or Engineer Features), and splits these further into over 125 "tasks" (such as Define the Target Variable). For any given activity a practitioner is about to execute, the risk system provides a set of associated risks (figure 3) that can affect Performance, Explainability, or Fairness. For each of these \sim 75 risks, the system also provides users with "war stories" (successes or challenges from past experience), as well as "mitigations", which contain both technical and non-technical steps to identify and overcome a particular risk (figure 4). Within the webapp interface, users can read and contribute content, based on their experiences.

Previous approaches to risk in machine learning (Holland et al., 2018; Mitchell et al., 2019; Gebru et al., 2018; Arnold et al., 2018; Varshney et al., 2018) take the form of *checklists*: lists of questions that are typically considered or answered *after modelling is completed*. Our approach goes beyond these existing methodologies in four ways.

First and foremost, the risk mitigation system is organised around modelling activities, encouraging practitioners to manage risk *as they are building models*, rather than just auditing for risks after the models have been created. This structure also enables practitioners to quickly find the content that is most relevant to what they are doing.

Second, this is the first approach to managing risk in machine learning that uses a scalable system to *record mitigations* along with risks. Prior approaches (such as model cards, figure 5) typically prompt modellers to ask questions, without providing advice or processes to answer them. Our system allows users to capitalise on the experience of others over many projects, and ensures a consistent and reliable approach.

Third, in order to facilitate the scaling of this library, we propose a *unified conceptual structure for recording risks and mitigations*. Risks are defined in one sentence with a clause specifying what can be impacted if the risk is not controlled (figure 3). Each risk's *mitigation* includes three categories of content (figure 4): Assess (how to tell whether the risk is relevant), Mitigate (how to overcome the risk), and Communicate (what and with whom to discuss, if this risk applies). War Stories, attached to risks or to mitigations, catalogue specific examples of these risks, their impact, and mitigating steps taken. Our mitigations may also point the reader to relevant academic literature, or to software packages (internal or external) that may be useful in overcoming the risk.

Finally, teams can use this platform to provide transparency to team leaders and business stakeholders, by creating a customised risk worksheet for a specific project. This helps teams plan mitigations and record and audit their actions taken in response to risk.

Our approach ensures that the latest research can be deployed responsibly with senior-level business support, in the industries that need it most. For companies with many data science projects running concurrently, the experience of each project team becomes the collective experience of the whole, as mistakes and experiences overcoming challenges become fruitful codified knowledge available and accessible to all.

REFERENCES

- Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Darrell Reimer, Alexandra Olteanu, David Piorkowski, Jason Tsay, and Kush R. Varshney. FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. *IBM technical report*, arXiv:1808.07261, 2018.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for Datasets. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning, Stockholm, Sweden.* 2018.
- Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. arXiv:1805.03677, 2018.
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency—FAT* '19, Atlanta, GA, USA.* ACM Press, New York, NY, USA, 2019.
- Kush R. Varshney, Dennis Wei, Amit Dhurandhar, Karthikeyan Natesan Ramamurthy, and Jason Tsay. Automatic Generation of Factsheets for Trusted AI in a Runtime Environment. Presented at the 2018 NeurIPS Expo Demo, 2018.

¹The risk management system introduced here was created through a collaboration between over thirty colleagues, including data engineers, data scientists, machine learning engineers, product managers, advanced analytics consultants, and experts in legislation and information security. Contributors included but were not limited to: Shubham Agrawal, Roger Burkhardt, Rupam Das, Marco Diciolla, Mohammed ElNabawy, Konstantinos Georgatzis, Hege Larsen, Matej Macak, George Mathews, Ines Marusic, Helen Mayhew, James Mulligan, Alejandra Parra-Orlandoni, Erik Pazos, Antenor Rizo-Patron, Joel Schwartzmann, Vasiliki Stergiou, Andrew Saunders, Suraj Subrahmanyan, Toby Sykes, Stavros Tsalides, Ian Whalen, Chris Wigley, Didier Vila, Jiaju Yan, Jun Yoon, and Huilin Zeng. All authors are listed in alphabetical order.

Explore risk topics

Performance

Review factors and potential risks that impact model performance - ranging from mistakes in data collection, to a failure in addressing constraints and business requirements.

Learn More



Explore potential pitfalls a team may encounter if explaining model predictions and recommendations is a significant requirement on their engagement.

Learn More

Fairness

Models can sometimes be unfair to certain individuals or categories of people, for example women or non-whites, by having lower accuracy or biased results for these groups.

Learn More

Figure 1: THE RISK MITIGATION SYSTEM COVERS THREE CATEGORIES OF RISK. Users explore the risk mitigation system through a webapp interface.



Figure 2: EACH RISK CATEGORY HAS AN OVERVIEW PAGE. Each category of risk (performance, explainability, and fairness) has introductory explanations and questions to help practitioners new to the topic learn about risk at a high level. The overview page for fairness is shown here.



Figure 3: EACH ACTIVITY AND TASK IN THE MODEL-BUILDING PROCESS HAS RISKS LINKED TO IT. Sample fairness risks are shown associated with two activities, (a) "Assessing the data" and (b) "Developing the analytical solution." Each risk is related to a specific task within each activity. Risks are recorded in a consistent format, in one sentence, with a clause that articulates what can be impacted if this risk is ignored.

Sampling bias and population shift

Quality analysis fails to take into account data set shift, population shift, or a sampling bias in the data set, leading to performance loss and poor generalization performance



In a project for a heavy manufacturing business, we performed an analysis on the target variable, which revealed a strange target ratio pattern in different months. In some months, the percentage of broken manufacturing items was high; in others, it was low.

Mitigate

Correct the data or adjust the modeling approach in cooperation with the business.

- It may be necessary to throw out affected data.
- For co-variate shift, any features derived from this data column should also be flagged e.g. if the shift is around certain subgroups, consider only producing features that are unbiased (i.e. normalized values for each subgroup), if the shift is due to poor data quality, refer to pitfalls under 'Assessing data'.
- Flag data set shift as a consideration for future modeling.

Read Less

Communicate

Be transparent with the business that model generalization may be impacted due to co-variate shift*.



On the same project, the team suspected there was selection bias in the labeling process and came back to trace the source. They realized this was because the labeler was trying to catch all the target observations that are easy to label in the first run, and labeling all observations by time sequence in the second run. This inflated the target ratio in certain months because labeling was still a work in progress. Based on this, the team randomly divided all the observations into buckets and asked the labeler to annotate them one by one, and only put them into use when a full bucket of observations was labeled.

Figure 4: EACH RISK HAS STORIES FROM THE FIELD AND MITIGATION SUGGESTIONS ATTACHED TO IT. The stories either highlight the impact of the risk or help a team see how to overcome a challenging situation. Each risk has associated reactions to take in response, that are categorised into actions that Assess, Mitigate, or Communicate the risk.

Model Card
Model Details, Basic information about the model.
 Person or organization developing model
- Model date
 Model version
– Model type
- Information about training algorithms, parameters, fair-
ness constraints or other applied approaches, and features
 Paper or other resource for more information
 Citation details
– License
 Where to send questions or comments about the model
• Intended Use. Use cases that were envisioned during de-
velopment.
 Primary intended uses
 Primary intended users
 Out-of-scope use cases
 Factors. Factors could include demographic or phenotypic
groups, environmental conditions, technical attributes, or
others listed in Section 4.3.
 Relevant factors
 Evaluation factors
• Metrics. Metrics should be chosen to reflect potential real-
world impacts of the model.
 Model performance measures
- Decision thresholds
- Variation approaches
• Evaluation Data. Details on the dataset(s) used for the
quantitative analyses in the card.
- Datasets
- Motivation
- Freprocessing
• Training Data. May not be possible to provide in practice.
If such detail is not possible minimal allowable information
should be provided here such as details of the distribution
over various factors in the training datasets
Ouantitative Analyses
- Unitary results
- Intersectional results
Ethical Considerations

• Caveats and Recommendations

Figure 5: THE MODEL CARD SYSTEM SUMMARISES IMPORTANT CONSIDERATIONS WHEN DEPLOYING MODELS. The Model Card (Mitchell et al., 2019) helps model developers to document aspects of how the model was constructed, and implications for its usage, but does not provide advice on how to overcome risks.