Building Fair and Transparent Machine Learning via Operationalized Risk Management: Towards an Open-Access Standard Protocol

Background & Motivation

We introduce a risk management protocol and webapp platform for practitioners that highlight major risks As machine learning increasingly integrates into business decision processes with wide-ranging consequences, from hiring through to law enforcement, there is a need for models to be transparent, unbiased, and robust. There around fairness, bias, and explainability at each stage of development. Because risks are embedded in this protocol, practitioners can understand risks and follow mitigation advice associated with the tasks they are currently completing. are as yet no broadly-adopted standard approaches to ensure that models meet these requirements. It is critical that models ... Risks are embedded in an ML model creation protocol Are sufficiently Generalize well in Are fair to all The protocol for model-building is broken up into more than 20 activities. 000 production explainable groups Identifying ML Opportunities Defining Sucess Metrics Build Data Pipelines Monitor Performance Asessing Feasibility Identify Analytics Approach • Engineering Features **Literature Review** \diamond \diamond \diamond \rightarrow Specialized Technical **Worked Examples Post-Hoc** Ē Checklists Tools Sampling bias and population shi Quality analysis fails to take into account o bias in the data set, leading to performance Checklists, such as ML Test Qualitative assessments, such Technical packages, such as Task 6. Evaluate Performance as Racist in the Machine³, which Score⁴, covering potential risks What-If tool⁵, which are used to Using the agreed performance metri Performance metrics are agreed with the business before modeling starts. It is likely that a range of metrics might be used, to understand the goodness of fit between the model and the data, the performance of the model in relation to the business's value creation mechanism, and other attribu such as the filamess of the model. across a range of axes from model detail the impact of unaddressed debug various aspects of machine Assessing data Risks (5) - Mitigate Performance is assessed in relation to the <u>definition of success</u> agreed with the business, and to benchmarks of the existing way of working or other standard approaches. performance to legal concerns. risks on models, and their societal fodeling is an iterative process, and performance is re-availuated on the validation data as the moo mproved. Once a series of models are finalized, an estimate of the true generalization performance he test data is created and communicated to the business. (There may be no need to use separate test in **agtimization** projects.) learning projects Explicit sensitive attributes There are sensitive or protected attrib gender, or religion, which can lead to b implications. Task 5. Iteratively Develop Models When communicating performance to the business, it is crucial to employ clear language and pro description of the core concepts. The example below is an example performance description for a hypothetical generalized linear model (GLM) with a single feature: These are used to audit and Tools typically inspect for a Removed sensitive attributes Sensitive or protected attributes have been is, making it difficult to assess whether the Description of main components
The copy of the model is a time with an equal type of each of the model is a time with an equal type of the three of the model is a time with an equal type of the three of the the three of the th Simplified example of a GLMs with one feature Unrepresentative cross validation Train/test splitting does not equally reflect proportions of so data, leading to poor generalization of fairness assessment: document models after they're • Communicate the second seco particular or related set of issues. Task 1. Profile Data Task 2. Partition Data Set deployed. Implicit sensitive attributes Minority features removed ensitive attributes can be inferred from nonsensitive attributes in ncoding"), which heightens the possibility of an unfair model. Task 7. Model Diagnostics A Refs 3 eatures that are predictive for minority groups but not majority groups are discarded in sature selection, leading to lower model performance for minority groups Once the performance of the model has b iterations to the model design. Task 1. Profile Data Task 4. Refine & Select Features **User Research** Unequal accuracy Performance is lower for one subgroup relative to another Previous approaches to risk management in machine learning take the form of Each risk has detailed Extensive interviews revealed that our users wanted a pre-production checklists: lists of questions that are typically considered or answered mitigation advice and Each activity is broken **Risks are associated with** after modelling is completed (see, for example, Breck et al.'s rubric for ML production "one-stop shop" linked to the way they worked, that case studies readiness, or the Model Card framework (Mitchell et al., 2019) into tasks - over 125 total each tast would help them identify and overcome the most relevant Our user research indicated that this checklist approach was insufficient. risks. Sample Fairness Risks Selected from Library of 100+ Risks Requirements Activity Missing fairness metrics - Fairness metrics are not defined, when they Define success metrics may be useful for the use-case 03 Define Model Evaluation What metrics are Fairness/Performance imbalance - The trade-off between performance and fairness Metrics appropriate, and metrics is not defined, resulting in a model with poor performance or insufficient how should they be defined? emphasis on fairness PRACTICAL ACTIONABLE COMPREHENSIVE UNIFIED SCALABLE **Explicit sensitive attributes -** There are sensitive or protected attributes explicitly included in the data, such as race, gender, or religion, which can lead to bias **Risks considered** Each risk includes in a model against these groups **Risks are presented** Consistent format Assess the data across model Removed sensitive attributes - Sensitive attributes can be inferred from manage risk as they practical solutions to in stardardized enables scaling 05 Profile the Data nonsensitive attributes in the data ('redundant encoding'), which heightens the development Does the data exhibit any adress them form across all risk possibility of an unfair model. qualitiesthat s hould inform lifecycle and can know which categories the modelling Assess Data Quality Imbalanced data - If most of the data comes from one subgroup approach? then the model may be inaccurate for other subgroups, leading to lower performance as well as risk of discrimination **Inferior data quality -** Data for a subgroup is missing, inaccurate, - or otherwise biased, which can lead to unfairness and discrimination Personas Unrepresentative train/test split - Train/test splitting does not equally reflect proportions of sensitive characteristics in the data, leading to poor generalization of fairness assessments Partition Data Set PRACTITIONER **COMPANY TEAM MANAGER** Developing the analytical solution DS/DE LEADERSHIP \mathcal{D} **Minority features removed -** Features that are predictive for subgroups but 08 Partition Data Set What models should be built not majority groups are discarded in feature selection, leading to lower model performance for subgroups to solve the problem? Partition Data Set **Unequal performance -** Performance is lower for one subgroup relative to another "Where are our gaps?" "Which risks are relevant to the "How can I help my team tasks I am doing now?" prioritize and scope for risks?" "How can we make sure we **Structure of Risks and Mitigations:** "How have other teams handled don't make the same mistake "How I can be sure I've considered these challenges?' twice?" all risks, comprehensively?" Risks Mitigations Translating the problem Risks (3) 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 Each Risk Missing fairness requirements comes along Considerations relating to fairness are inadequately considered as part of the business project for a heavy manufacturing business, we performed an analysis **Operationalized Risk Management** Risks are embedded with a Mitigation problem and its constraints revealed a strange target ratio pattern in in tasks and have Mitigations include months. In some months, the percentage of broken manufacturir tems was high; in others, it was low war stories and a consistent Task 5. Identify Constraints ii. Mitigations That Capitalize on Expertise are broken up into conceptual structure Mitigate This is the first approach to managing risk in machine Assess, Mitigate, Unfair unit of analysis Correct the data or adjust the modeling approach in cooperation with the business. Risks are formulated Communicate The choice of unit of analysis* favors one group over another, which can lead to i. Risks embedded in an ML Model-Building Protocol learning that uses a scalable system to record in one sentence with discrimination mitigations along with risks, informed by historical an impact clause, Communicate enabling Task 2. Define Unit of Analysis experience and reviewed by experts Be transparent with the business that model generalization may be impacted due to co-variate "Risk War Stories" consistency and models have been created highlight challenges scalability as new Improper substitute target variable iii. Consistent Conceptual Structure risks are added teams have faced The choice for the target variable results in bias against a subgroup. This may be ame proiect, the team suspected there was selection bias in th in the past, to help Risks, Mitigations, and War Stories are captured in a because the target is a poor proxy for what the business is interested in predicting, and Enables practitioners to quickly find the risks and he source. They realized this wa **Risks are nested** bring risks to life and the proxy is biased. cause the labeler was trying to catch all the target observations that are consistent conceptual structure, to facilitate scaling by mitigation materials that are most relevant to the tasks illustrate impact within high-level easy to label in the first run, and labeling all observations by time sequence Task 3. Define Target Variable(s) adding risks and mitigations after each project "activities" and they are doing *in the second run. This inflated the target ratio in certain months because* "Mitigation War granular "tasks", so labeling was still a work in progress. Based on this, the team randomly Stories" help users can quickly divided all the observations into buckets and asked the labeler to annotate Defining success metrics Risks **(2)** them one by one, and only put them into use when a full bucket of identify the risk teams learn how to observations was labele Copyright © 2019 by QuantumBlack Visual Analytics Limited overcome challenges relevant to them



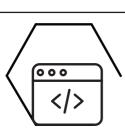














Practitioners can go through projects risks are relevant at any point in time



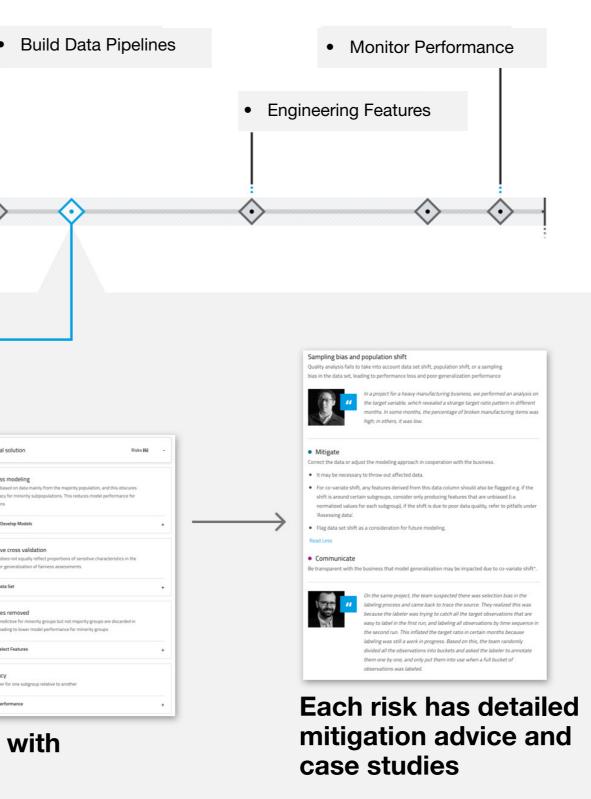




Key Contributions

Allows practitioners to manage risks as they are building models, rather than auditing for risks after the

Our Risk Management Protocol & Webapp



Future State: for Discussion

In this workshop, we invite discussion on how to make the protocol and platform open-access, communitysourced, and an industry-standard approach to building models that are fair, accountable, and transparent.

Future directions

Expand Scope of Risks/Mitigations



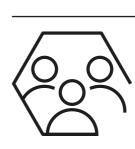
a. Technique-spe cific Risks i.e. Deep learning. casuality analysis or optimisation

Develop technical



Stress test risk protocol on applied ML studies across industries

Impact



ADOPTION

Businesses will be quicker to adopt ML, as it will be less risky

Potential Solution for Practitioners: a Risk Mitigation Worksheet to record and communicate project risks

Risk Library	Finding
Assess the Data	
Inconsistencies in data collection or recording	Dates were sometimes recorded in US format (MMDD), sometimes in UK format (DDMM)
Data for a subpopulation is missing, inaccurate, or otherwise biased, which can lead to unfairness and discrimination	"Star players" had more accurate injury data than players typically on the bench
Align on Modelling Requirements	
Difficulties forecasting of the amount of time it would take to implement XAI	LIME was not straightforward to apply due to missing values
Define the analytical problem	
Target variable is an imperfect proxy	Data for the target variable comes from actual injuries, which is a proxy for what the model aims to predict: 'likelihood for injury'
Evaluation Metric does not match business use case	ROC vs. APS: does the high-precision model matter most?

Questions for Discussion:

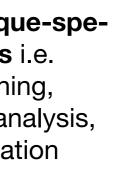
- performance, explainability, and fairness?

Corresponding author: daniel.first@guantumblack.com ? The risk management system was created through a collaboration between over thirty colleagues, including data engineers, data scientists, product managers, management consultants, lawyers, and information security experts. Contributors included: Shubham Agrawal, Roger Burkhardt, Jacomo Corbo, Rupam Das, Marco Diciolla, Mohammed ElNabawy, Konstantinos Georgatzis, Carlo Giovine, Stephanie Kaiser, Mate

Didier Vila, Jiaju Yan, Jun Yoon, and Huilin Zeng







b. Domain-specif ic Risks i.e. Healthcare, banking, or insurance

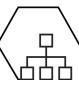


c. Risk Themes i.e. Information security or regulatory risks

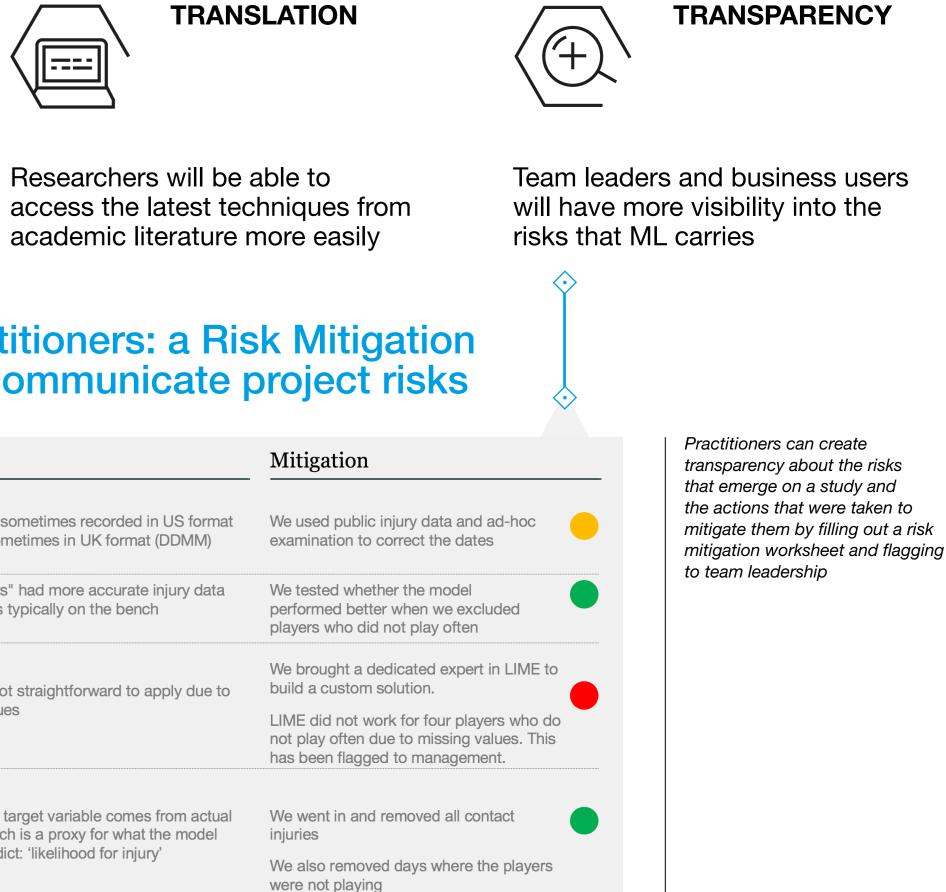


d. Open-Sourcing i.e. Make the risk and mitigation library public

A data linter that flags potential biases within data sources



An open-source model pipelining framework, that is able to assess risks at defined stage-gates



(01) Would you use a risk management approach in your work?

We used Average Precision Score instead

) What is your company's approach to ensuring

(03) Would you contribute to an open-source risk library?

(04) Do you have any suggestions for technical tooling?

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, Julian Waton, Ian Whalen, Chris Wigley,

3 Garcia, M. Racist in the machine: The disturbing implications of algorithmic bias. World Policy Journal, 33(4): 111-117, 2016 4 Breck, E., Cai, S., Nielsen, E., Slib, M., and Sculley, D. The ML test score: A rubric for ML production readiness and technical debt reduction. In 2017 IEEE International Conference on Big Data (Big Data), pp. 1123-1132, Dec 2017. doi: 10.1109/BigData.20178258038. 5 Google PAIR Lab. The What-If Tool: Code-Free Probing of Machine Learning Models, 2018