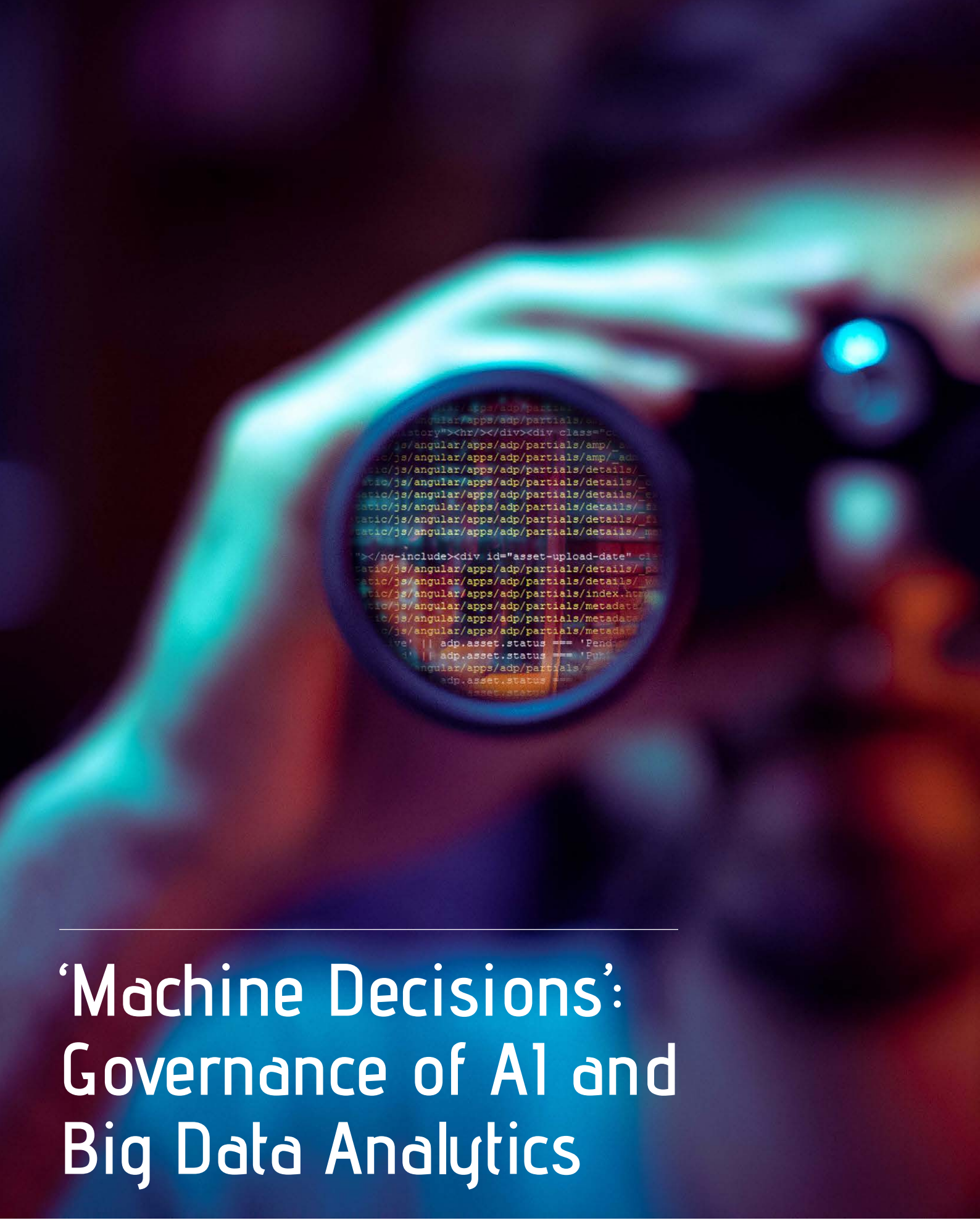


‘Machine Decisions’: Governance of AI and Big Data Analytics

CRO FORUM

Contents

1 Executive summary

1.1. Introduction

As in a number of other sectors, insurers are developing and deploying increasingly complex Big Data tools, first for post-sales purposes such as claims handling and customer service, but also potentially for sales, pricing and underwriting. The CRO Forum published a paper in 2017 on **Big Data & Privacy**, which raised a number of important governance questions. More recently, the CRO Forum has examined how to adapt existing governance to cover such tools, including machine learning algorithms, and what reasonable steps can be taken to 'know what's happening inside the black box', ensure reliability and avoid unfair discrimination.

We consider that guidance on how to apply a model risk framework in this field is timely as applications are evolving fast and firms and regulators are reflecting on their response. Big Data and AI present particular challenges due to complexity, self-calibration, autonomy and the potential for unexpected results and unforeseen impacts. We have explored these issues with the use of case studies and a group comprising data scientists, model risk experts and governance experts. The paper focuses solely on the (re)insurance industry and is meant to provide a practical overview for risk managers.

Section 2 of this paper outlines the context and overall area of focus. This section also probes external stakeholder implications including regulatory focus and possible reputational risk if things 'go wrong'.

Section 3 describes the main types of models and diagnostic tools. It explores their development and reminds the reader that governance processes need to keep pace. It provides a taster to some techniques that can be used to challenge model output and when these might be useful. Section 3 lastly emphasizes that a combination of techniques are likely to be needed to properly understand and grapple with model output in the light of the risks articulated, the tools employed and the purpose of the model itself.

1.2. A number of key insights emerged

Section 4 explores some case studies. It is structured in a way where the reader can engage upfront in a key principle or guideline. The case study then develops this idea further and allows the reader to explore why the area of focus is important and critical. The key principles are the main insights that emerged from our case study discussions. They include the following:

- It is vital to invest sufficient effort upfront to define the problem the complex tool / modelling technique is expected to solve in pursuit of a business objective. The more powerful the tool, the more important this is.
- Human intervention is and will remain critical. Each generation of model will tend to give more precise or selective results than its predecessor, and it can be tempting to accept the output and declare that the black box works. But machine learning models are highly non-linear in nature with complex interdependencies, and so they tend to be extremely difficult for humans to supervise. A key challenge is identifying when and how humans need to intervene and what triggers are in place to prompt their intervention. Examples are explored again to prompt the reader and provide some live test cases where human intervention is or was critical.
- Training needs are emphasized and explored in this section. This should apply at all levels in an organization. It is felt that executive buy-in to the purpose and the risks is key to good governance in respect of complex models and data driven tools.

Section 5 explores the complex area of ethics and bias in data and how such bias, if not understood and managed, can cause significant issues for an organization. The firm's approach should be informed by its value system and overall risk appetite. We therefore recommend that Management and Boards ensure the values are clear and the bias and ethics risks are understood so that appropriate measures can be embedded in the organization. Additional links are provided to assist the reader in exploring this item further.

Building on the sections above, Section 6 of the paper gives an example of an overall model governance framework for Big Data tools, and provides a checklist that the reader can reference when needed. Key here is that model governance techniques and frameworks that exist today do not need to be fundamentally altered, but can be enhanced and adjusted to meet the evolving needs of complex tools and machine learning developments.

Potential risks, importance of human oversight and key recommendations:

Potential risks

A number of risks could arise from complex tools, Big Data and AI, and need to be managed:

- Risk of algorithmic bias
  - Quality of training data – danger of magnifying inherent bias in data, e.g. accidental racial, social or gender profiling
  - Clarity of goals for an algorithm and testing against the company's values
- Risk of lack of explainability
  - Not auditable and cannot justify outputs
  - Risk of loss of human expertise, skill and insight
- Risk of repeated, systemic or runaway error
- Risk of social exclusion if insurance becomes unaffordable for specific profiles
- Reputational, regulatory or legal risk of bad individual outcomes or impact on privacy

Importance of human oversight

Humans remain accountable for machines and algorithms; therefore we need to design explainable tools with useful 'windows into the black box', so we can:

- Justify the output and how we use it, by explaining why and how as well as what
  - Demonstrate fairness, compliance, ethics and non-discrimination by the tool
  - Better understand the tool's limitations and risks so we can design controls that fit
- Sense-check model output and get the tool to show what moves the dial
  - Access insights from the tool to share between functions (Claims, Underwriting, Portfolio Management, Reinsurance, Actuarial)
  - Develop the next generation of technical experts
  - Understand and preserve our core intellectual property in an accessible way

Summary of key recommendations:

- Adapt and extend existing model governance to fit Big Data tools and their uses
- Use explanation methods that are simple enough, and describe the key model limitations
- Ensure human oversight over
  - Setting up the algorithm's goal, or the problem for it to solve
  - Ensuring that the input data being used is high-quality, clean and appropriate, with bias minimised
  - Validating the model outputs to ensure that a suitable solution has been found
  - Validating the model outputs to identify inappropriate bias against vulnerable groups
  - Designing triggers for human intervention, e.g. an unusual localised spike in volumes
- Ensure the firm's values and ethical framework are clear, regularly reviewed and can be applied in practice through appropriate controls and input and output testing
- Document for relevant stakeholders how any critical trade-offs between accuracy vs bias or auditability vs privacy have been decided

2 Background and approach taken

2.1. Approach and observations about use

In the last few years, many human activities have shifted rapidly to the digital dimension. This trend also brings a tremendous amount of digital information which offers great opportunities to the insurance sector. More information increases the appetite both for new data to be used and to apply new technologies to it. These technologies can be used to build better and more robust predictive models. Predictions can be made not just of the future risk but also of customer behaviour and other questions. Model complexity is increasing and the demands of interpreting and challenging model output are also becoming more difficult.

The aim of this document is not to focus on the technical aspects of Big Data tools and quantitative analysis associated with the application of the tools. It is not intended here to document the detailed techniques used or associated algorithms. The goal is to focus on the governance of Big Data tools and to develop high level principles that companies should take into account when using these methods.

The principles are meant to serve as a template or set of guidelines to facilitate the effective application of Big Data tools. Moreover, they aim to limit discrimination and bias which is deemed crucial to ensuring effective and equitable system design and service delivery.

The CRO Forum working group has been composed of representatives of various insurance companies that have some experience of using these tools. The goal was to share experience with new tools, their use and implementation in practice. This area is constantly evolving, and currently live application of AI decision-making is seen particularly in the Claims function. This includes but is not confined to claims journey automation, claim optimisation or fraud detection.

Complex tools are established in the pricing area where optimal price leads to better risk selection and supports overall profitability. However, despite the strong predictive power of the tools, in sales and pricing the usage is often limited as a support tool and the final decision remains to a large extent with the human being. This may evolve and change over time.

2.2. Reputational risk

The use of advanced tools, machine learning techniques and artificial intelligence leads to results which might not be fully comprehensible at an initial view. Making decisions or pricing products based on such model outputs could lead to strategic, operational or reputational risk. These risks should be managed rather than blocking the use of such tools. Moreover, risk mitigating activities should be in place to guarantee a certain quality standard. In particular, it is important that a certain degree of interpretability of model outputs is guaranteed, to help avoid accidental racial, social or gender profiling.

For example, consider the price of a standard insurance contract which is derived by some machine learning technique. If the pricing algorithm can use all available data as input the final price could depend on parameters which might raise ethical questions or challenges e.g. the wealth of a person. Even if the input parameters are constrained, some unknown correlations to other parameters could lead to such a dependency. Following this example through: if we assume income is not known by the model the postcode could indicate that the insured person lives in a less affluent neighbourhood. If this issue becomes public and is perceived as unfair, a large reputational damage could arise.

For this reason, it is key to challenge whether model results are fair and explain outcomes in terms of risk and other appropriate goals. Some examples are stated in the following chapters and explored in more detail. However, it is essential that a sophisticated system of governance exists. Some model weaknesses might already be known, but risk arises from unknown deficiencies. For this reason, quality assurance techniques must be in place and human intervention at the right time is important.

2.3. Regulatory Interest

The discussion on the use and governance around AI and Big Data Analytics is also of interest to regulators. For example, in 2018 Charles Randell, the chairman of Britain's Financial Conduct Authority, gave a speech with the title *"How can we ensure that Big Data does not make us prisoners of technology?"*[1]. He addressed some items to be considered when new technologies are using Big Data, artificial intelligence, machine learning and behavioural science. He left the question open if the industry or the regulator should take the lead in such questions. In 2017 the International Association of Insurance Supervisors published a paper on *"FinTech Developments in the Insurance Industry"*[2]. This paper also concerns innovations like Big Data, machine learning and artificial intelligence and their impact on the regulator.

The European Insurance and Occupational Pension Authority established an "Insurtech Task Force"[3] which analyses the use of Big Data and its related risks. Furthermore, the description of the task force states that as a next step the use of complex tools (artificial intelligence, machine learning) could be assessed and how they should be supervised could be explored. As a side remark, this task force might also assess how complex tools and Big Data can be used for supervisory purposes.

These examples show that this is an area of growing focus and interest for regulators. However, industry should take a lead, as the prime users, and this paper should help the industry to probe this question further.

[1] https://www.fca.org.uk/news/speeches/how-can-we-ensure-big-data-does-not-make-us-prisoners-technology

[2] https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=2ahUKEwiYuf-7oePfAhVUAxAIHYRyBusQFjABegQICBAC&url=https%3A%2F%2Fwww.iaisweb.org%2Ffile%2F65625%2Freport-on-fintech-developments-in-the-insurance-industry&usg=AOvVaw3EI-4cxPRSUBbEwfonRYb6

[3] https://eiopa.europa.eu/Publications/Administrative/InsuTech%20Task%20Force%20Mandate%20-%20BoS.pdf

3 Summary of types of Big Data analytics in use

Analytics at its basic level has always been used in the insurance industry and predictive models have been used for a significant period of time. While techniques employed and data used has gradually evolved over time, the pace of change in the past decade has been significant, driven by:

- An increase in the amount of data generated and how much data is available for analysis.
- Computing Power
- The emergence, particularly in the past few years, of new analytical techniques partially in response to the new data opportunities presented.

3.1 Differences traditional regression and machine learning

Before discussing the differences between traditional regression and machine learning methods it is important to note that both aim to describe a relationship between input (x) and output (y), using a function (f(x)):

$$f(x)=y$$

Linear vs Non-Linear

Traditional regression models are used by insurers for example in accepting clients or pricing (by estimating for example the likelihood that a specific event will occur). Traditional regression models are linear in nature, either directly (linear regression) or indirectly (logistic regression or generalized linear models). Machine learning models can be linear or highly non-linear by nature and are therefore better able to capture complex dependencies. On the other hand, it is more difficult to understand what happens under extreme conditions.

Supervised vs Unsupervised

A key difference in machine learning models is whether they are "supervised" or "unsupervised". Supervised models typically seek to ascribe input data to an existing set of output categories (e.g. matching images of animals to an existing list of animals) whilst unsupervised models seek to identify similarities in data and place records into categories that have not been pre-defined. This makes the models inherently more difficult to interpret and the success and rationale for the model harder to validate.

Few vs Many Parameters

Another potential difference can be found in the number of parameters tested and in the risk of over-fitting a model based on the training data. In traditional regression, the number of risk drivers can be manually limited via statistical tests to prevent overfitting. Given the linear behaviour of these traditional models, the modelled parameter based on a training set of data can be validated against a different testing dataset to ensure that the initial model and parameter values selected are still predictive. Machine learning models typically take a different approach, trying to use as much input data as possible and testing many different model iterations and interaction terms between the different risk drivers before a final model is derived. In many cases, different approaches need to be taken to avoid over-fitting including cross-validation (similar to creating training and test data sets), manually removing features that don't make sense to a user and "early stopping" model iterations at the point in which the model performance on the testing dataset begins to worsen.

Strict Assumptions vs Unconstrained

Based on statistical theory, traditional regression models have strict statistical assumptions (e.g. independence, homoscedasticity and exogeneity) that each can be tested and confirmed. Working through a set of statistical tests can therefore show that the model is working as intended. Statisticians will also understand how data was collected, the distribution of data they are using for modelling and the rationale for the statistical technique they are using to model the data. In combination, this makes model governance reasonably straightforward as it follows a common set of model checks to be performed to give the reviewer confidence in the model, even if in practice these assumptions do not always strictly apply to the dataset in question.

Machine learning models on the other hand are goal seeking and simply aim to mimic specific behaviour in the data, without any prior assumptions on relationships between variables. The model is free to discover patterns that best model the training data. This can make it feel like a "black box" which, as long as it models the data well, will not question "how" the model has been derived. The difference between the traditional techniques and machine

learning makes certain governance elements (e.g. assessing input data and rigorous testing of output results) much more critical when reviewing machine learning models.

3.2 Weighting the pros and cons of the Big Data Analytics algorithms

Ensuring that these analytical models are appropriate either from a financial, ethical, reputational or regulatory perspective is not a new problem. The challenge today is how governance can remain as effective as in the past in an environment in which analytical models are proliferating rapidly, evolving more quickly and increasingly could be deemed as using "black box" techniques that can be hard to interpret.

Infographic **Pros and cons of various common algorithms:**

Generalised linear model (GLM)

Pros	Cons
• Widely used and understood within Insurance and Actuarial disciplines • Relatively easy to interpret	• Difficult to deal with very wide data sets (i.e. >500 features) • Can't handle non-numeric data • Slow model build due to human involvement in feature selection

Level of human involvement

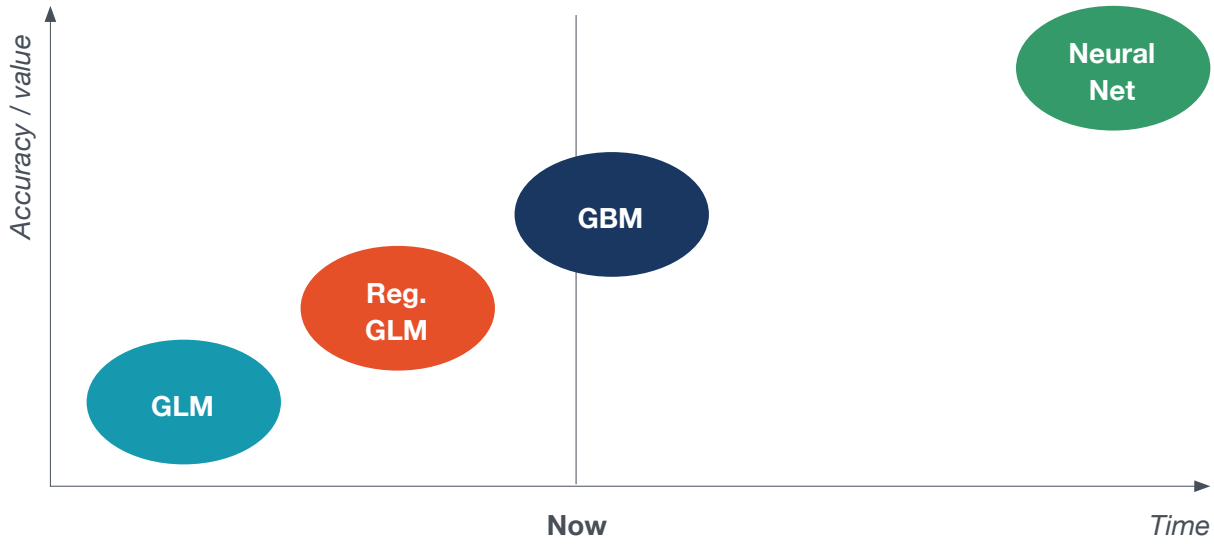
- High level of human involvement
- Typically involved at each stage: data preparation, model setup, model fit, feature selection and model refinement
- Skilled human input is essential for a good model

Regularised GLM

Pros	Cons
• Extension of GLM • Regularisation enables an element of automated model build and simplification • Can speed up model development and avoid GLM overfit	• Elements of automation reduce the opportunity for human insight in feature selection or data engineering • Tooling is currently less advanced than GLM fitting tools

Level of human involvement

- Medium level of human involvement
- As per GLM, but feature selection is performed by the machine rather than the human



Gradient boosting machine (GBM)

Pros	Cons
• Very effective algorithmic technique for regression and classification against a non-linear underlying process • Machine learning so resulting model is objective • Accuracy and speed advantages over GLM	• Relatively new technique, so less well understood amongst Insurers and Actuaries (risk of being poorly applied) • Lack of mature enterprise tooling • Less interpretable than GLMs

Level of human involvement

- Medium to low level of human involvement
- Primary human involvement is in the initial data prep and problem setup

Neural network

Pros	Cons
• Can give the most accurate results when large volumes of training data are available • Capable of performing internal feature engineering • Can transfer learnings between domains	• Difficult to know the most appropriate network structure • Very difficult for humans to interpret • Less well developed outside vision and language applications • Can give very confident but incorrect predictions

Level of human involvement

- Low level of human involvement once appropriate network architecture is known
- Computer training can be long and require specific hardware

Why machine learning models are not black boxes and how to think about and apply interpretability is usefully discussed in the following paper 'The Mythos of Model Interpretability' (https://arxiv.org/pdf/1606.03490.pdf).

It is a given that alongside machine learning techniques, there must be a simultaneous step change in validation techniques and governance rigour if the inherent risks associated with these developments are to be controlled and the full business benefits available from these techniques are to be realised.

The infographic outlined on the previous page shows the progression of modelling techniques in relation to the:

- Different types of models being used over time
- Their pros and cons and key trade-offs
- How they are used and what the role of the human is.

This is intended to simply prompt the reader to think about the pace of change and whether their overall model governance processes remain aligned with this change agenda.

In addition to the above techniques, "ensemble models" are becoming increasingly prevalent. Ensemble models combine the best models within a class to make a "super model" that is more predictive than any given model built in isolation. While from a modelling perspective this is desirable, it makes the challenge of interpretation even harder without appropriate techniques to interpret the ensuing results.

3.3. Tools for insight and validation

As data analytics progresses, tools for model validation are evolving too, which is vital if risk is to be controlled and model outputs are to be understood. As more decision-making is transferred to computer algorithms, there will need to be a greater human emphasis on:

- Setting up the problem for the computer to solve; and
- Ensuring that the input data being used is high-quality, clean and appropriate, with bias minimised; and
- Validating the model outputs to gain assurance that a suitable solution has been found.

A number of techniques have been developed that can provide insight into the decision making process of black-box algorithms. Any of these techniques used in isolation will have inadequacies in respect of model validation, but skilful human interpretation of a variety of different techniques will provide insight and assurance that should form a robust basis for model validation.

In this paper, we have introduced four such validation techniques from the many available. The first two techniques (global variable importance and partial dependency plots) help explain the overall impact of an input feature (e.g. gender) on model predictions while the last two (LIME and SHAP) help explain the contribution an individual factor has on a specific prediction rather than the overall impact.

The techniques listed are widely available among both open source and commercial modelling products.

This is an active area of research and available approaches are regularly evolving, as illustrated by the multiple techniques being applied in the H2O.ai[4] machine learning software as just one of the many machine learning software toolkits currently available[5].

3.3.1. Global variable importance

Global variable importance (aka. feature importance) ranks the overall impact of features on model predictions. In figure 1 on the next page, this shows how important key features of the Titanic would have been in predicting survival rates. In this example, the individual's gender would have been the most important factor in determining whether a person survived.

3.3.2. Partial dependency plots

Partial dependency plots show how the overall prediction varies over the range of values for that feature. In the same Titanic example (see figure 2 on the next page), we can plot survival rates by age which shows that children had a higher survival rate than adults.

[4] http://docs.h2o.ai/driverless-ai/latest-stable/docs/booklets/MLIBooklet.pdf

[5] http://www.datarobot.com/wiki/insights/

3.3.3. Prediction Explanations: LIME

LIME[6] is a technique that explains the individual predictions (or local explanations) of any complex non-interpretable model in an interpretable and faithful manner. It does so by training a simple and interpretable model locally around the prediction (aka surrogate). The simple surrogate model is typically a sparse linear model, but other models such as GLM have been proposed in variants such as k-LIME. This helps the user to understand the impact of the locally important feature in explaining a single result (e.g. what variables were important, and how much, to predict Ms Smith's survival rather than the global variable importance for the total population i.e. all passengers).

3.3.4. SHapley Additive exPlanations (SHAP)

An alternative to LIME that can also explain an individual prediction of the model is SHAP[7]. It explains a result in terms of the weight of different features in giving rise to that prediction. It connects game theory with local explanations, uniting several previous methods and representing a consistent and locally accurate additive feature attribution method based on expectations. It uses game theory to allocate the "pay-out" for each data point amongst the feature variables so that the weight of the "pay-out" represents the importance of each feature.

3.3.5. Other techniques

There are many other techniques that help aid model understanding, such as visualization. The following reference provides a good overview of some of these and explains in a different manner the techniques mentioned above: *https://towardsdatascience.com/explainable-artificial-intelligence-part-2-model-interpretation-strategies-75d4afa6b739*

Explainable AI (XAI), Interpretable AI, or Transparent AI is a very active research topic, which means that the above-mentioned techniques should not be considered as definitive and we should expect new results[8], such as limitations, and new techniques to appear in the near future. The interested reader should refer to publications from conferences[9] such as FAT pointed to below.

Figure 1 **Example variable importance**

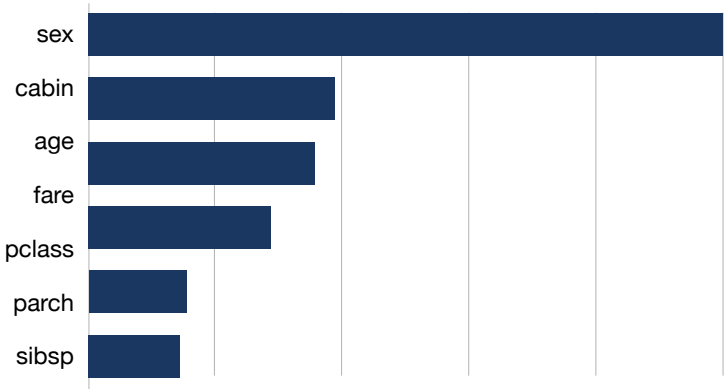
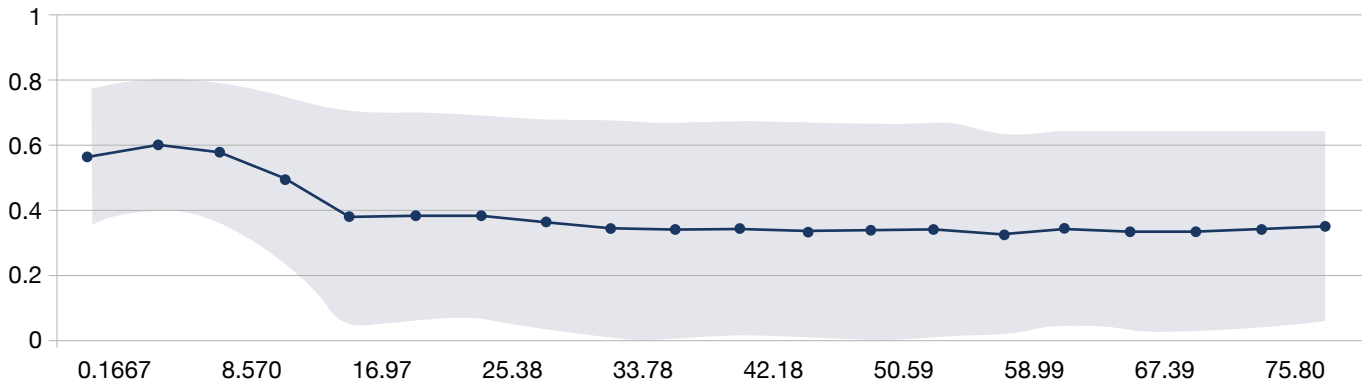


Figure 2 **Example partial dependency** (mean response with ± 1 standard deviation, shaded for numeric data and bars for categorical data)



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

[7] https://github.com/slundberg/shap

[8] https://axa-rev-research.github.io/

[9] https://fatconference.org/

4 Key learnings from case studies

This section provides insights and key learnings based on a number of case studies. Key principles discussed are:

- The importance of the problem statement
- Keeping humans in the loop
- Training needs.

4.1. Problem statement

4.1.1. Explanation of key principle

In Big Data Analytics, the **problem statement** is the phase of up-front design before data exploration and before the model is built. It is the first and most important step of a machine learning project. It is even more vital when the problem is complex.

Before formulating the problem, it is important to perform an opportunity discovery. This involves deriving business value from data collected and owned by the organization. This step is not about data exploration but is about finding potential use cases and applications by looking at which type of data is available. The use cases found are then prioritized and the phase of defining the problem statement can start. While not unique to the subject of this paper this is a critical step in the view of the working group in particular where machine learning is used.

4.1.2. Rationale why this key principle is important

While every problem is different and there is no standard method for formulating a problem statement, it is important to follow several steps before implementing a machine learning algorithm.

First of all, the problem needs to be defined. This includes:

- Formulating the problem statement in a simple and clear manner – even in one sentence
- Defining the task that need to be completed and associated performance metrics (i.e. define what is success)
- After discussing with subject matter experts, listing some intuitions about the problem, i.e. which behaviour in the variables can explain well the outcome. This might highlight that the data to be used is not optimal and provide ideas for the sources of data that could be used
- Trying to link the problem to existing problems. Other problems can help highlight limitations in the phrasing of the problem such as time dimensions and conceptual drift (where the concept being modelled changes over time). Other problems can also point to algorithms and data transformations that could be adopted to spot check performance

Among all these aspects, **defining the performance/success metrics is key**. For classification purposes (e.g. predicting a 0 or a 1), it is critical to define if the goal is to maximize accuracy, minimize false negatives or minimize false positives. Depending on which of these goals is key, the performance metrics to use will be different (accuracy, precision, recall, ...).

The second step is to explain why the problem needs to be solved:

- Explain what the expected benefits are if the problem is solved
- Define how the solution will be used by the end-user including the definition of the requirements of the tool (as in regular software development projects). It is crucial to perform this step by working directly with the end-user from the beginning. Using agile development methods (e.g. iterative development) will improve the efficiency of this step as it is not always possible to define all the requirements from the start.
- Perform an analysis of the envisaged solution to see whether using Big Data Analytics is necessary. In some cases, a simple rule-based program could be more suited if the problem is simple in nature.

The third step of the problem statement is the phase of prototyping and experimentation which is performed before the final model is built. This step will highlight the operational difficulties of solving the problem. For example, the experimentation phase could show it is necessary to acquire more data.

The example provided on the next page illustrates the importance of having a clearly defined and articulated problem statement.

4.2. Humans in the loop

4.2.1. Explanation of principle

Specifically in the area of supervised learning, companies are experimenting with the replacement of traditional regression models by machine learning models. In Section 3 we described the differences between traditional models and machine learning. When replacing

Example

Cancer detection using machine learning

A common example on why it is key to define success in statistics is the detection of if an individual has a cancer.

For cancer detection, a false positive, also called 'Type I' error, is when an individual is predicted to have a cancer when it is not the case. A false negative, or 'Type II' error, is the contrary, i.e. when an individual is predicted to not have a cancer while it is actually the case. It is obvious in this case that the cost of a false negative is significantly larger than the cost of a false positive. For a false negative, the individual might remain undiagnosed and die. For a false positive, the individual will likely live.

In that case, the success metric to choose must include the minimization of the 'Type II' error rate. If that is not the case, the cancer detection might not achieve the purpose for which it is designed and potentially cause real issues.

traditional models by machine learning, keeping humans in the loop becomes increasingly important.

4.2.2. Rationale why this principle is important

In this section we point to some of the complexities using several case studies, and then discuss what this means for the need to have humans in the loop.

In traditional regression, the interaction between the dependent variable and the independent variables is usually clear and thoroughly tested. In machine learning models the number of independent variables is usually much higher and non-linear behaviour and interaction terms make this interaction even harder to trace.

Machine learning models are, just like traditional regression models, trained on historical data. If a certain bias in the historical data is present, this can typically be resolved by removing the corresponding independent variable.

Another problem encountered when working with machine learning models is that it is often hard to predict what will happen in extreme cases that did not manifest in the available history. These problems are further illustrated in the examples below.

The bias as described in the Amazon case, is a model bias resulting from biased data. From a modelling perspective, the model is functioning well, since it mimics the data. Although the model cannot identify the bias, humans can. It is therefore pivotal to have humans periodically monitoring the modelling results to apply common-sense and identifying possible sources of bias, particularly those that would be socially unacceptable.

Note that even if a model is not biased, it might pick up a bias during model use. From that perspective it makes sense to identify possible sources of bias and put automatic controls in place, triggering human intervention.

One approach to tackling gender bias as adopted by LinkedIn is to build separate models for males and females and then show the top 5 male and female candidates alongside each other. Whatever approach is taken, it is important to collect information regarding the possible sources of bias, though that might require collection and use of sensitive information such as for example race. Data such as gender and race is often collected by organisations specifically for the purposes of testing for bias.

The example below shows that machine learning models are not always able to correctly identify situations that deviate too much from historically observed events. In the Tesla case described below, the model did not correctly identify the truck as a vehicle or obstacle and appeared to have ignored it.

Since this kind of situation is not known in advance, it is important that such situations are identified and adequately dealt with. This kind of issue is addressed in the field of adversarial machine learning, which aims to mislead the model by searching for boundary cases that are not correctly identified.

Example

Amazon - gender bias

A recent example where the limitations of machine learning became apparent is at Amazon, where the recruitment selection algorithm showed a bias towards hiring men. Given that Amazon historically hired more men than woman, this bias is present in the data and a machine learning or regression algorithm (that tries to mimic the historical data) will likely inherit this bias as well.
The initial solution was to remove gender as an independent variable of the model. The problem in this case with machine learning as opposed to traditional regression is that you can take out a specific risk driver, but that the machine learning will automatically work around this by taking proxies this risk driver (in case of the gender example this is the use of certain words in a resume that are associated with a specific gender).

Although openly communicated by Amazon, this is likely a phenomenon that occurs more widely. One approach that could be followed is to compare a machine learning model with a traditional regression model, as in the latter a specific risk driver can be excluded more easily.

Example

Tesla - autopilot

One of the big challenges in artificial intelligence is the self-driving car. Tesla is well known for its autopilot, but does require the driver to remain vigilant at all times. Nonetheless, according to Tesla, Tesla achieves a far lower accident rate than other cars[10], having crash rates 4 times lower if autopilot is not engaged and 7 times lower if autopilot is engaged. Nonetheless accidents do happen, such as for example the crash in 2016 in Florida, where a Tesla on autopilot did not intervene when the car hit a trailer of a semi-truck that was crossing the road.

The challenge with systems such as autopilot is that the amount of visual input that needs to be processed is enormous and that accidents typically occur in abnormal situations. As such the focus from humans should be more on extreme cases and scenarios where the model might behave unexpectedly. Significant effort should be put into identifying and analysing these scenarios.

[10] https://www.tesla.com/blog/q3-2018-vehicle-safety-report

4.3. Training needs

4.3.1. Explanation of key principle

Complex models, like other models, are designed for a given purpose and framework. They only provide relevant predictions under this framework. This is a risk for the end-user who does not have the technical skill set and can potentially use a tool for a purpose for which it is not designed. This statement is true for existing complex tools, even though they are not using techniques such as machine learning.

4.3.2. Rationale why this key principle focussing on training needs is important

Several options can be envisaged in order to limit the risk of misusing complex tools:

- **Training the end-user:** It is the most trivial solution. However, this might not be sufficient as it is probable that someone, someday, will use the tool in an inappropriate manner.
- **Limiting the freedom of the end-use to the minimum:** For example, defining a lower and upper bound for each numerical parameter and not accepting parameter values that are not within these boundaries in which the model assumptions are valid.
- **Hiring people to make the link between the model designer and the end-user:** This kind of position is sometimes called 'data analyst'. They usually have a technical background but are more business-focused than data scientists. For example, the 'data analyst' can use the model and propose business or marketing ideas to the sales manager. The two skills of the 'data analyst' (technical and business) ensure the risk of misuse of the model is reduced.
- **Creating a data-driven culture:** This is changing the way people work in the company, finding business problems and allowing them to develop proofs of concept (even if many of them fail) to solve business problems. In terms of human resources, this includes organizing basic machine learning / data science courses in the company for everyone, building awareness about data science and closing the data literacy gap. The goal is that everyone in the company can be part of this change of paradigm.

In addition, AI and complex tools are used and will be used more and more in the future in decision processes and overall company governance. Therefore, the Boards of Directors and Executive Committees should also receive training for them to understand the limitations of these complex and new models. Transitioning to a data-driven company should receive executive buy-in and support.

Example

Google

In 2009, one of Google's former executives, Marissa Mayer (before becoming CEO of Yahoo!), launched a project to test 41 different shades of blue for its Google advertising links. This led one of the Google team leaders to quit, being "tired of debating such minuscule design decisions". However, it was later reported that this minor optimization led to increase Google's ad revenues by $200m in a year. This illustrates that having a full buy-in and support from executives for a data-driven project is compulsory. While investing time and money in such a project may appear unnecessary, its impact on Google's revenues was significant and would not have been possible without a sponsor from the top.

5 Applying an ethical framework to the use of Big Data and AI in insurance

Use of Big Data Analytics and AI within insurance has the potential to generate business advantage and lead to positive outcomes for clients. On the other hand, the use of AI may heighten the risk of systematic misuse due to the models' reliance on historical data which may include embedded human bias. Using biased historical data as 'training data' in models can reinforce bias in an accelerated and widespread manner which gives rise to ethical and legal risks for the firm, and reputational risks for the industry as a whole.

Illustration:
A risk-assessment tool developed by a privately owned company, was used to decide whether to grant people parole and ended up discriminating against African-American and Hispanic men.

The offering of differential treatment to different groups of individuals, based on demographic characteristics (e.g. age), where there is clear statistical evidence on cause and effect relation, is accepted practice within insurance (subject to what may be legally permissible, such as the requirement to eliminate gender pricing in the EU in [2011]). An important consideration going forward will be to assess whether there is ongoing justification for such differentiation. It is possible that certain types of differentiation may be perceived as discrimination based on moral and values criteria such as avoiding racial discrimination, although such criteria may change over time. Therefore, 'justifiability' may be judged not just from the perspective of what is defensible by legal standards, but also what is acceptable within a firm's own ethical framework (see below).

Illustration: Some credit card companies in the US started cutting cardholders' credit limits when charges appeared for marriage guidance counselling, since marriage breakdown is highly correlated with debt default. In such cases firms should assess whether such differentiation is ethically defensible even though it may be *legally permissible*.

A simple yardstick is to treat all individuals in a fair and explainable manner. Having such an approach leads to *fair* outcomes and firms may consider achieving fairness as a minimum goal. Firms may choose to go a step further to build models that constrain outcomes for certain vulnerable demographic groups that they want to protect. The definition of what is classed as a 'vulnerable group', would depend on the firm's own ethical values. Not all sub-groups would be seen as vulnerable or ethically sensitive, e.g. a firm might decide actively to protect pregnant mothers or new parents from deselection or high prices, but not actively to protect students.

Firms can lean on descriptions put forward by the Financial Conduct Authority (FCA) in making this determination. Between 2014 and 2015, the FCA published two papers, *'Consumer Vulnerability'* and *'Vulnerability Exposed'* which present some pointers to insurers to identify vulnerability and good practices to prevent discrimination with respect to vulnerable groups. It suggested a list of risk-factors (not intended to be an exhaustive list) that could guide firms in identifying vulnerabilities:

- Long-Term or significant illness
- Carers
- Older people
- Low basic skills i.e. individuals who struggle with literacy / or numeracy, with or without presence of a formally diagnosed disability
- Job loss or unemployment
- Bereavement

The following descriptions have been put forward by the FCA to assist firms in understanding the different ways vulnerability could occur:

- *"A vulnerable consumer is someone who, due to their personal circumstances, is especially susceptible to detriment, particularly when a firm is not acting with appropriate levels of care."*
- *"The situations and circumstances of 'vulnerable' individuals are diverse, complex and dynamic; the experience of vulnerability is unpredictable, and it can change over time."*
- *"Vulnerability has not just to do with the situation of the consumer. It can be caused or exacerbated by the actions or processes of firms."*

The Senior Management and Board of a firm will find it useful to have established

an ethical framework that defines which personal characteristics (whether lifetime or situational) the firm wishes to control. The ethical framework can be reflected in product and customer strategy, and can inform design decisions and can be used to set parameters and testing strategies for modellers.

5.1. Managing ethical risks

The rest of this section sets out guidance to help insurers understand and mitigate the ethical risks from the use of AI models. This section does not intend to prescribe an ethical framework for firms, instead it focuses how firms' own ethical frameworks may be applied to the governance of AI models. As with all areas of risk management, a proportionate risk-based approach is useful for dealing with these risks. In general, ethical risks could be more material compared to more established approaches based on factors such as those listed below. These include:

- Extent to which the AI models aid decision making
- Severity and likelihood of the impact of the decisions on individuals or groups of individuals, particularly those that may be viewed as 'vulnerable groups'
- Complexity of the model
- Strength of redress mechanisms for customers

In the segment below, we break down the ethical considerations relating to use of AI models by different stages of model development, namely purpose, means (models and data) and outcomes:

Purpose

Purpose of building an AI model could have direct implications on severity and probability of bias

Means

Factors that insurers could consider to minimise bias when deciding on input data and design of the model

Outcomes

How the outputs of the model may be validated to minimise bias, and what considerations should be fed into design of redress mechanisms

5.2. Ethical purpose

When choosing to incorporate use of Big Data and AI in modelling, an ethical purpose is key. Ethical purpose may be defined as one that's based on the principle of non-maleficence i.e. 'do no harm'. Below is a list of considerations when dealing with the purpose of AI models from the perspective of ethics:

- It is recommended that firms clearly define the purpose of building an AI model. Purpose may impact the choice of technology, and the severity of impact of the model on stakeholders.

  Illustration: An AI model that uses customers' personal attributes to calculate credit score to approve/ decline loan applications has the potential of more severe financial impact on customers as compared to one that is used to improve speech recognition of customer calls. Outlining the purpose of the model clearly at the outset will improve firms' ability to apply a proportionate approach to mitigating ethical risks.

- Clearly articulating and linking the outcome metrics to the purpose is likely to generate beneficial outcomes for the business and customers alike, and also lead to better understanding of the impact of the model and how it is measured.

  Illustration: A firm may build a chat-bot to cross-sell products and to improve customer engagement. The transactional aspect of the chat-bot may be disguised in interfaces that appear to offer a diagnosis or information to customers. In such a case, firms should clearly set outcome metrics (e.g. sales and user satisfaction) to link to the model's purpose. This will lead to a more holistic measurement of impact of the model on business and customers.

- When using AI models to automate steps in the process, there may be a risk of staff redundancy in medium to long term. Therefore, it is recommended that firms make concrete efforts to reskill staff well in advance of operationalising such automation and also consider including staff in the newly designed process (e.g. in the ongoing validation of model outcomes).

5.3. Means

A key source of bias in AI models stems from the quality of data that is fed into the model. Therefore, efforts to minimise bias from AI models would likely have a strong data management component. Below is a list of considerations when dealing with data and design of AI models from perspective of ethics:

- Firms may want to proactively inform customers about the use of AI and the type of data collected from them. The implications of collection and use of their data can be made known to customers in an easy-to-understand manner.

  Illustration: Firms can provide customers online data management tools such as privacy or personal data dashboards. Customers may use such tools to review what they choose to share on on-going basis instead of being provided the choice once in the process.

- Firms should ensure that quality of data sets is adequate and regularly screened for bias, particularly with respect to vulnerable customer categories. If the historical data is found to be biased, firms should adjust the data to ensure that bias is eliminated or minimised.

  Illustration: The city of Boston released an excellent 'StreetBump' smartphone app which drew on accelerometer and GPS data to help passively detect potholes. Unfortunately, due to low smartphone ownership amongst lower income groups and elderly, the data sets were missing inputs from the parts of the population who had the fewest resources and, consequently, the results were biased.

- Firms should try to increase the 'auditability' of AI models. This may involve the following components:
  - Firms may need to maintain an audit trail of historical training data. In a Big Data environment, this could potentially translate into additional investment in efficient storage and retrieval mechanisms.
  - Firms could try and improve on 'intelligibility' of all models i.e. strive to make models less of a 'black-box'. Models that have a material impact on customers should be designed to enable tracing of individual decisions to model inputs.
  - In case of complex AI models, firms should foresee the training requirements of technical and customer-facing staff.
- Model bias is more likely to be an outcome from biased training data as opposed to model error. Models may tend to be less 'intelligible' compared to input data, therefore, it is recommended that firms try and address issue of model bias by correcting training data. If this fails, then firms may require to constrain the model to mitigate bias.
- When designing the model, firms may have critical trade-offs to consider e.g. accuracy vs. bias, auditability vs. privacy. It is recommended that firms document these trade-offs in as detailed a manner as possible and communicate decisions taken with respect to them to relevant stakeholders.

  Illustration: Firms that design a model to make outputs fully traceable on an individual basis may find that they are required to expose customers' personal data to staff dealing with escalation procedures. Depending on the type of data being collected, customers may not be comfortable with this even though it leads to better outcomes for them from a redress standpoint.

5.4. Outcomes

The complex nature of AI models is such that, despite best efforts to minimise bias, the models could still have unintended consequences. Below is a list of considerations for insurers when dealing with model outputs and outcomes from an ethical perspective:

- Firms should clearly set out output metrics that are being targeted via the AI models and assess these metrics against firms' own ethical framework.
- As in the case of input data, the model outputs may need to be regularly (perhaps continuously) screened to test for 'bias', particularly with respect to sub groups that are defined as 'vulnerable' from the firm's view point. Automated triggers (Section 4.2.2) requiring human intervention may be built in with regards to the vulnerable categories to drive more equitable decision making from the tools. Where necessary, firms can consider actively constraining the outputs of models to protect vulnerable categories. The frequency and granularity of this checking should be 'risk-based' i.e. linked to the severity of impact of the model output in customers and staff.
- It is good practice to incorporate a human element in analysing the results from the screening of model outputs. Firms may consider involving a wider group of people (e.g. consumer groups and customer facing staff) in the screening process. Using a diverse range of view-points could improve firms' ability to test for bias.
- Unintended bias in model output may be difficult to eliminate entirely. Therefore, firms may consider designing effective information and redress procedures to ensure fair and justifiable outcomes.
- Effective remediation mechanisms may require complete traceability of model outcomes, especially in regulatory regimes where replicability is mandated.

In general, insurers may want to evaluate their firms' code of conduct to see whether it is still valid in an environment of Big Data Analytics and AI. Firms should also try and generate board-level awareness on ethical risks surrounding the use of AI so that the topic gets due priority.

To ensure that, as an industry, we stay on the right track, a human-centric approach to AI is needed, keeping in mind that the development and use of AI should not be seen as a means in itself, but as having the goal to improve business outcomes and human well-being. Achieving this human-centric approach would require insurers to consult with a wide group of stakeholders (e.g. consumer groups, technology companies, legal advisors, government and public sector representatives) on an on-going basis. One insurance firm created a cross-functional ethics forum and learnt that it was necessary for them to deal with the topic on an iterative basis instead of a one-off check at the point of development or deployment.

A second firm has established a working group to set principles for the ethical and transparent collection and processing of customer data; their discussions revealed the difficult balance between maintaining high standards of ethical data use while still providing sufficient flexibility for innovation.

Depending on the market, it is possible that insurers and technology companies have more technical capability relating to use and risks of AI compared to regulators. Therefore, the industry can consider proactively building capacity at the regulators' end to improve outcomes for the industry participants as a whole.

References that maybe helpful are:

- Vulnerability exposed: The consumer experience of vulnerability in financial services
- Consumer Vulnerability
- European commission draft guidelines for trustworthy AI
- Monetary Authority of Singapore (MAS) principles to promote fairness, ethics, accountability and transparency
- FCA view on ethical implications of Big Data
- UK Government's Data Ethics Framework
- OECD expert group on AI
- Principles for Accountable Algorithms and a Social Impact Statement for Algorithms

6 Model management framework for Big Data Analytics

As set out in Section 3, Big Data Analytics involves the use of increasingly complex techniques/models. The realm of models ranges from traditional 'Linear regression' models to the latest and most sophisticated 'Deep Learning' models. Nevertheless, use of models is not new to the (re)insurance industry.

A variety of models (with varying degrees of complexity) are used throughout the life cycle of insurance products. For example, Forecasting, Pricing, Valuation, Capital allocation and Asset Liability Management. Consequently, 'Model Risk' has been a topic of interest for firms and regulators alike and a lot of effort has gone into understanding and mitigating these risks through appropriate management of models.

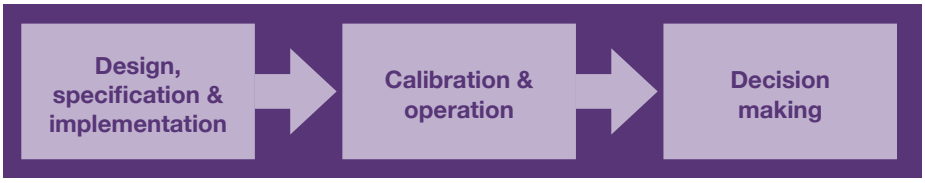
In the context of Big Data Analytics, models/tools could be deployed in almost every function, moving beyond the traditional areas mentioned above. These tools help in enhancing operational efficiency through automation and competitive advantage through business insights. Some examples of newer areas are automated Underwriting, automated Claims handling and 'Chatbots' in Customer service.

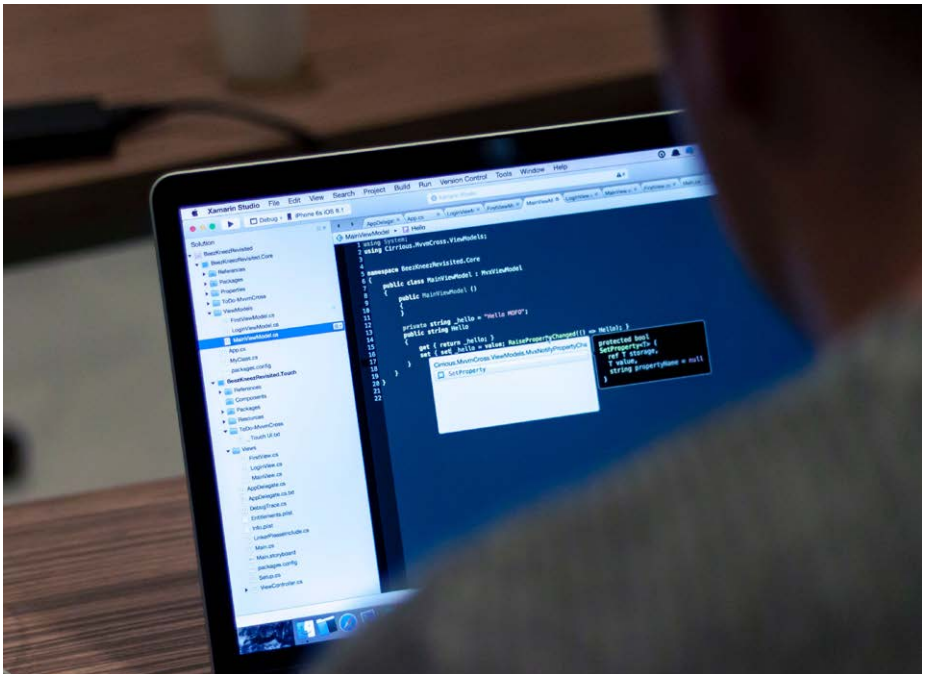
In addition, Big Data Analytics introduces firms to a whole new range of models and techniques that were previously unused, e.g. Deep Learning, Random forests, Support Vector Machines, etc., bringing with it newer challenges.

6.1. Modelling process and associated risks

The CRO Forum's publication, *Leading practices in model management*, March 2017, comprehensively covers the topic of model risks and provides best practices for model management. The publication mainly focuses on 'traditional' models, used in the insurance/reinsurance industries, describing the fundamental modelling process and the associated risks. Figure 3 illustrates the important stages in a typical modelling process (as discussed in *Leading practices in model management*).

Figure 3 **Modelling process**



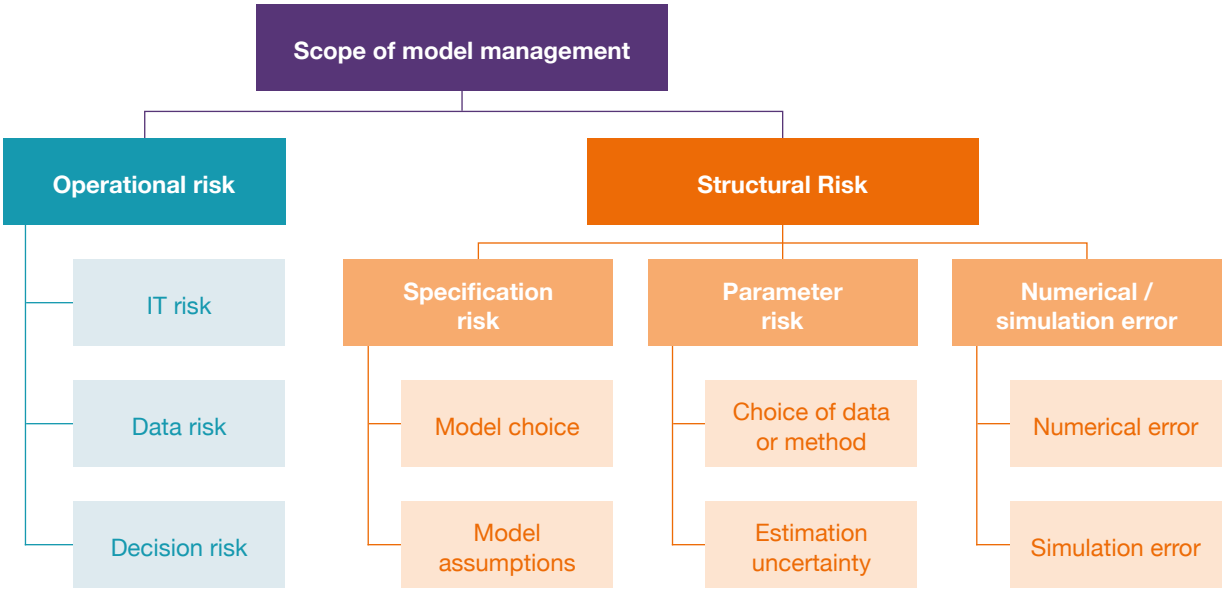


It is interesting to note that the underlying modelling process is fundamentally the same for 'Big Data Analytics' as well. Contrary to the general hype around fully self-learning models (also called Artificial General Intelligence), humans are still an integral part of developing and maintaining even the most sophisticated of models currently deployable. Consequently, the types of risks associated with the models are also not too different from before. However, the way the risks present themeselves in this context is different as there is a huge difference in the kind and volume of data that could be used, sophistication of the algorithms used and the range of business problems that the models can help to resolve.

The fundamental modelling process and the associated types of model risks are very similar to traditional models, allowing the existing processes and governance to be adapted to address specific challenges posed by 'Big Data Analytics tools.

An appropriate model risk management framework involves identifying the relevant risks and charting a corresponding mitigation plan. Figure 4 Illustrates the various risks identified as part of model management.

Figure 4 **Risks in the scope of model management (reproduced from *Leading practices in model management*)**



The risks in figure 4 (for further explanation refer to *Leading practices in model management*) are applicable to Big Data Analytics as well.

Limitations in the technological infrastructure, such as design inadequacy or inappropriate implementation, and gaps in controls/ governance ecosystem lead to Operational risks. Structural risks are caused by the inherent limitations of the models which are simplified representations of the real world. As explained in Section 3.1 the main difference between traditional (statistical) models and Big Data Analytics models (such as Machine learning) is that these new models do not assume any specific relation between input variables, rather discover and reflect the patterns inherent in the data.

Hence, 'Data Risk' and 'Decision Risk' are critical in the context of Big Data Analytics and are discussed further in this section.

Data risk:
The need for data to be 'fit for purpose' has become even more critical with the use of 'Big Data'. Availability of large amounts of data from a wider variety of sources and at a more granular level, means, it is ever more important to understand what data and data features/ attributes are being used and for what purpose. With sophisticated techniques such as 'Deep Learning' it is challenging to explain how the input data attributes impact the model results, leading to 'Black Box' perception.

Data protection legislation such as EU's GDPR holds firms accountable for their use of personal data. This, coupled with the risk of unintended biases in decisions (against or favoring certain sections of society) and use of attributes perceived as discriminatory (e.g. ethnicity), poses questions around an organisation's ethics. As outlined in Section 2.3, this could put firm's reputation in jeopardy and have regulatory implications.

In the Amazon Gender Bias case study in Section 4.2, while the data was manipulated to eliminate gender bias by removing the particular attribute, the presence of proxies to gender (e.g. certain words in the resume) which was not removed led to the biased decisions.

This illustrates the need for extra care in handling big data by considering the direct and indirect implications of using

certain data. Traditionally model users have focused on clean data ('garbage-in, garbage-out') and fit-for-purpose data, whereas Big Data Analytics requires a mindset change requiring the model developers/users to consider wider ramifications of using certain data, even though the data may be 'clean' and 'fit for purpose'.

While 'clean' and 'fit for purpose' data is fundamental to any modelling, the variety of data that could be potentially used within Big Data Analytics necessitates careful consideration of the wider implications of using that data.

(Relevant references include the publication from The CRO Forum, "Big Data & Privacy: unlocking value for consumers, CROs in a changing environment", that addresses the challenges of use of Big Data analytics).

Decision risk:
Decision Risk arises when business decisions are based on inappropriate use of a model. For example, due to inappropriate calibration, outdated models, inappropriate interpretation of the model results or lack of understanding of the limitations of the model. As mentioned earlier, 'Big Data Analytics' tools could be deployed in a wide range of business contexts as compared to traditional models. For example, underwriting, claims management and customer service. Hence, a whole new breadth of decisions are now driven by results of models.

The 'decision risk' associated with Big Data Analytics tools is similar to that of traditional models. However, Big Data Analytics tools lead to decision risks in new ways. For example, lack of desired level of interpretability (explainable outcomes) leads to ill-informed business decisions that are hard to justify. The use of certain sensitive data potentially leads to biased or discriminatory decisions.

From a regulatory standpoint, with legislation such as GDPR, firms can no longer hide behind the 'complexity' of their models to justify their decisions. There is demand for more accountability as firms use more complex models for decision making.

The generic nature of Big Data Analytics tools means they can be used in a variety of business contexts (where data is available); thus it is important to deliberate on the implications of those decisions, at the design stage itself.

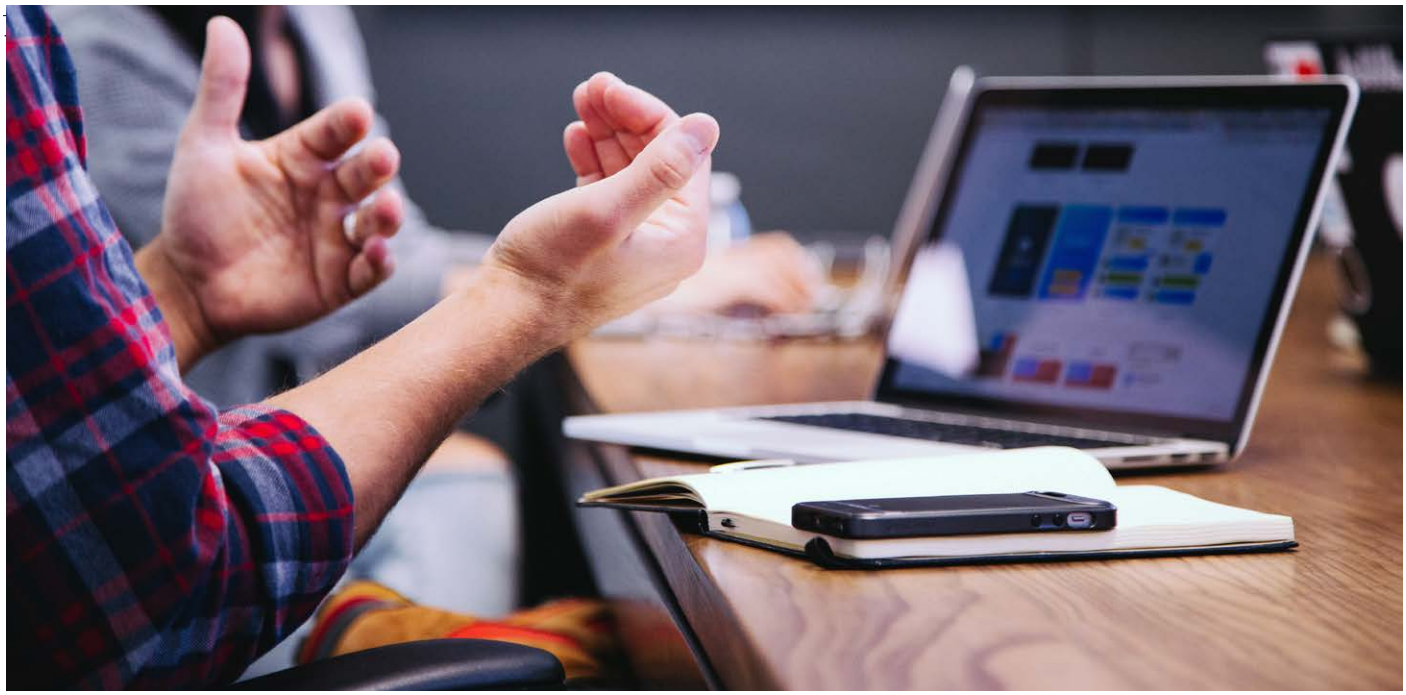
6.2. Model governance

'Model Governance' encompasses a framework for the use and maintenance of models, validation of models, and the adequate disclosure of model assumptions and limitations. A robust model governance supplements the other internal controls and governance frameworks in an organization. For example, given the additional risks associated with use of personal data under Big Data Analytics, 'Data Governance' becomes a vital extension of a robust Model Governance framework. Such extensions help consider risks holistically thereby providing effective assurance.

The principles of traditional Model Risk Management continue to be relevant for these newer tools as well. The specifics of how the principles are adapted in this context could vary. Some of the crucial elements of managing risks that are relevant in the context of Big Data Analytics are discussed further.

Classification of models
A key element of a robust framework is the classification of the models based on risk rating. Many criteria could be used for such rating, for example:

- Whether model results are for internal or external purposes



- Potential customer impacts
- Potential financial impacts
- Potential regulatory or strategic implications
- Potential reputational risks

For Big Data Analytics, additional criteria to be considered would be:

- Social sensitivity of data inputs used
- Accuracy vs interpretability trade off

Typical risk rating classes used are "Critical", "High", "Medium" and "Low ". This risk classification then reflects the relative importance of the models within an organization, thereby allowing appropriate focus to be placed on each of the design, development and maintenance stages of the model. For example, "Critical" models may be required to undergo extensive/rigorous governance requirements, whereas "Low Risk" models would suffice to demonstrate basic model development and maintenance hygiene.

Model in this context includes underlying data, accompanying algorithm and the business decisions. This is quite important, as changes to any of the components (mentioned above) could lead to considerable change in the risk perception and consequently the risk rating of the model. For this reason, model (re)classification should be a regular/periodic exercise, rather than a one time effort.

For example, an automated underwriting decision model to reject a prospective insurance applicant maybe classified "Critical", due to the potential reputational and regulatory consequences of systematic discrimination against certain classes of individuals. However this model may be re-classified as "Medium"/"Low" risk if the business decision is changed to accepting applications automatically and allowing a human expert decide on rejections.

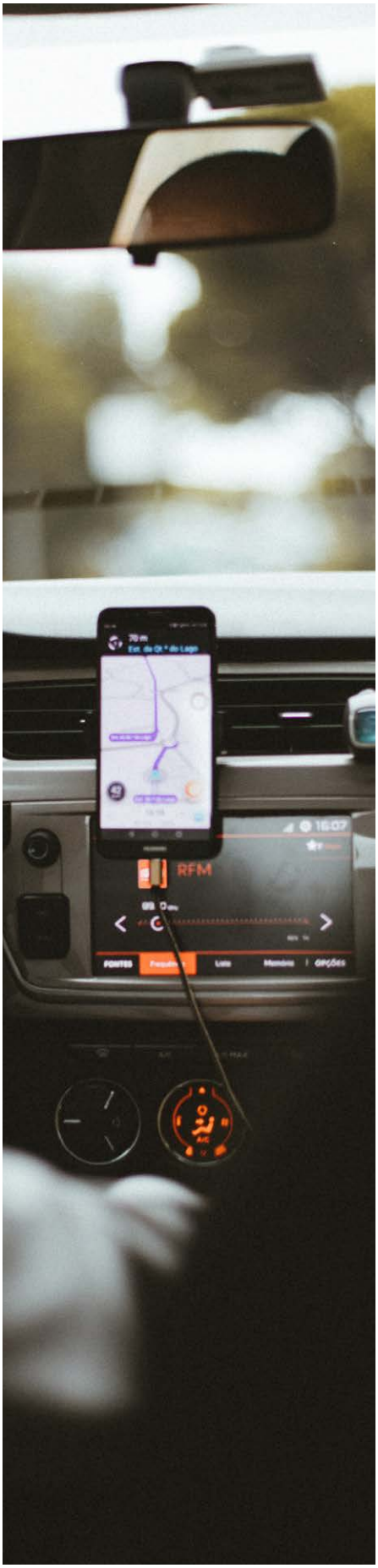
Appropriate classification of the models based on risk rating allows firms to focus on the most critical of the models and enables well-informed decisions.

In order to perform Big Data Analytics, a plethora of sophisticated algorithms are available with varying levels of 'interpretability', the ability to explain the results (see chapter 3). Interpretability is crucial in order to demonstrate a clear understanding of the interactions between the inputs and the outputs of the algorithm. Especially in the context of insurance/reinsurance business, decisions that could deprive individuals or sections of society of basic services need to be justifiable. It is critical to be able to demonstrate that the decisions based on the model results are driven by data (with due consideration to sensitive attributes) and are not arbitrary.

Choosing a model that delivers the desired level of accuracy while maintaining the required level of interpretability (of the results) is critical. Having the classification in place (as described previously in this section) aids the choice of model, achieving acceptable balance between the model accuracy and interpretability. This is business context specific and potentially requires regular monitoring in order to ensure the validity and relevance of the 'balance' on an ongoing basis.

For example, in the automated underwriting decision model (insurance applications) mentioned previously, interpretability may be preferred over accuracy (especially with legislations like EU GDPR) while in case of a targeted retention drive (to improve persistency in an insurance portfolio), an accurate lapse predictor may be preferred over interpretability.

Achieving an appropriate balance between model 'accuracy' and 'interpretability' while developing solutions helps firms to optimize risk and reward, i.e. the risk of taking decisions based on inexplicable model results.



Human oversight

Big-data analytic tools have a number of limitations: they provide data-driven, context specific solutions, enabled by availability of large volumes of data from existing and new sources. The underlying algorithms are designed to reflect the base data, mirroring the complexities and the biases (intended or unintended) alike. Consequently, results of these tools are only as reliable and appropriate as the data used to develop the underlying models. Moreover, these tools are restricted in their application beyond the originally intended context and/or changing 'context'. For example, an automated underwriting system that performs well in Europe may fail in Asia or an extreme/unforeseen input could result in meaningless output.

These challenges amplify in the context of tasks sensitive to 'reputational' or 'regulatory' risks rendering such tools unfit for deployment. This could restrict an organization from harnessing the full potential of such sophisticated tools. An appropriate level of human oversight at various stages of the development goes a long way in maximizing the utility of such tools.

Human centric design: As highlighted in the 'Amazon Gender Bias' case (Section 4), while it's easy to remove sensitive attributes such as gender, ethnicity etc from the training data it is very hard to avoid the use of proxies to such attributes[11]. However having a 'Human Centric' design approach could have helped avoid the bias to a great extent. i.e. as recruitment is a sensitive decision, instead of complete automation, a semi-automated approach of picking top 5 resumes each among men and women separately and taking them forward through manually procedure from thereon could have been considered.

Feedback mechanism: A major software firm designed a female chatbot with its own Twitter account that could learn about the world by conversing with its users. It had to be shut down within a short span of time as pranksters trained it with racist, sexist, and fascist statements[12]. A Human centric approach considering optimizing not only for speed and algorithmic accuracy, but also for user conduct and biases encoded in data could have helped. The organization was able to act quickly as there was a feedback loop with humans involved that helped identify the issue and shut the system down.

Trigger mechanism for human intervention: In case of unforeseen input values/ events, a thorough consideration of the acceptable range of outputs along with triggers to draw human attention is essential. For example, in the case of automated underwriting a threshold may be defined such that whenever the confidence in the output is less than the threshold, manual underwriting is triggered. In the aftermath of London Blasts in 2017 Uber was slammed for being too slow in turning the surge price[13] leading to a lot of bad press. A careful consideration of the 'cap' on prices while designing and human monitoring for unexpected surges could have helped trigger an alarm when prices rise significantly.

Incorporating appropriate human intervention at crucial stages in the model design goes a long way in minimizing unexpected model results.

A comprehensive risk consideration is critical in developing a robust Big Data Analytics model design that mitigates risks and produces reliable results under most circumstances. To achieve this, firms may devise strategies based on the robustness of their existing governance frameworks and the level of maturity in adoption of Big Data Analytics.

For example, a firm building up analytics capabilities may have to sensitize the data scientists around the finer risk aspects (reputational, regulatory) and the domain specific considerations, while the domain experts may have to be trained to understand the advantages and disadvantages of the various algorithms.

In the absence of any formal guidance on data analytics model development, firms could consider having a 'governance checklist' (sample provided in Appendix) for the model developers and business owners to refer to at the design stage of the solution itself. This not only helps mitigate risks but also creates better risk awareness within the firm. Over time as more models are developed and deployed in business, this checklist could be evolved into a formal best practice guideline or standards to transition into BAU.

A cross functional team consisting of data scientists, domain experts, model risk experts, data officer, regulatory experts and representatives from any other team (if required) could be set up to evaluate critical or high risk models to be deployed within the firm, discuss latest technical and regulatory developments and to create risk awareness within the firm.

[11] https://in.reuters.com/article/amazon-com-jobs-automation/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idINKCN1MK0AH

[12] https://www.technologyreview.com/s/603194/the-biggest-technology-failures-of-2016/?set=602944

[13] https://www.independent.co.uk/news/uk/home-news/london-terror-attack-uber-criticised-surge-pricing-after-london-bridge-black-cab-a7772246.html

7 Summary and key conclusions

To conclude, the main recommendations of this paper for the governance of machine decisions and big data models can be summarized as follows:

- Adapt and extend existing model governance to fit Big Data tools and their uses (Section 6.2); this would benefit from
  - Risk-assessing the models and establishing a cross-functional team to evaluate critical and high-risk models
  - Training modellers on use risks, training users on model limitations, training executives, and putting in place mechanisms to ensure cross-functional communication (Section 4.3)
- Develop explanation methods as an explicit objective alongside the model itself; the explanation should be simple enough, and describe the key model limitations; explanatory tools themselves are evolving alongside models (Section 3)
- Ensure the firm's values and ethical framework are clear and regularly reviewed and can be applied in practice to protect selected characteristics or vulnerable groups, through appropriate controls and input and output testing (Section 5)
- Ensure human oversight over the following (Section 4 and throughout)
  - Setting up the algorithm's goal, or the problem for it to solve, carefully and precisely to fit the business goal and to limit unintended consequences
  - Ensuring that the input data being used is high-quality, clean and appropriate, with bias minimised
  - Validating and sense-checking the model outputs to ensure that a suitable solution has been found
  - Validating the model outputs to identify inappropriate bias against vulnerable groups
  - Invest effort to design a comprehensive set of triggers for human oversight and intervention, e.g. an unusual localised spike in volumes
- Document for relevant stakeholders how any critical trade-offs between accuracy vs bias or auditability vs privacy have been decided (Section 5.3)

Machine learning, artificial intelligence and big data analytics are exciting and powerful tools with great potential to benefit customers and insurance firms. We believe that by following the steps outlined here, firms can continue to develop and deploy these tools safely and successfully.

Appendix 1
Sample checklist of model governance questions for Big Data analytics tools

In order to achieve an effective risk management framework, it is imperative to encourage comprehensive risk consideration in the modelling process. To this end, sample questions that could be addressed as part of the various stages of the process are provided below.

1.1 Design, specification and implementation

Business context

- Does the modelling goal appropriately reflect the business challenge?
- Do we understand what business decisions would be made based on the results of the model?
- Is the modelled output metric appropriate for the given business challenge?
- Have the regulatory and reputational risk implications of the business decisions been considered?

Appropriate and adequate data – "Fit for purpose data"

- Is the data available, appropriate and adequate to answer the business question?
- Is there appropriate data governance and data management processes in place?
- Are sensitive input data attributes handled carefully? i.e. those that could be perceived as discriminatory, such as, gender, religion, ethnicity or their proxies

Appropriate choice of model

- Is the selected model appropriate for the business problem and data?
- What trade-off between model interpretability and accuracy is desired?
- How is adequate testing/hold-out performed to eliminate 'overfitting'?
- Does the model perform poorly on training data? (Statistical Bias)
- Is the model uncertainty within acceptable limits?
- Are the model predictions resilient to noise within data?

1.2 Calibration and operation

Re-calibration

- Is there an agreed process in place for updating training data and retraining models on an ongoing basis?
- Is there a business continuity plan in case of problems in updating?

Change management

- Have the changes to the business processes following deployment been fully considered?
- Has the new process been tested and is contingency / dual running in place?
- Is it ensured that changes to the business processes are monitored?

1.3 Decision making

Automated decisions

- Are the potential implications of automating the decisions well understood?
- Is there an easy way to shut down automatic decisions, in case of any issues?
- Is there a contingency plan to move to manual decision making?
- Are relevant alerts in place to draw human attention to potential anomalies?
- Is sufficient human oversight in place?

Bias and fairness

- Has the firm identified particular attributes or vulnerable groups that it seeks to protect from bias or unfair treatment?
- Is there a process to test if the output decisions are unbiased/fair, e.g. to avoid accidental gender, social or racial profiling?
- Is it possible to demonstrate to stakeholders (regulators, clients, staff, shareholders) that the decisions are unbiased and fair?
- Are the decisions tested at regular intervals for possible biases/unfairness?
- Is the decision or the basis of the decision in conflict with the values of the company?

Explainable results

- Can the results of the model be fully described and inspected for individual records?
- Are decisions traceable? Can it be explained why a decision was reached?
- Are simple explanation methods for a complex model available?
- Is the right level of human oversight in place?
- Are the model limitations well understood?

Reproducible & auditable results

- Can the model training process and model results be replicated?
- Is the model appropriately documented such that a qualified third party can replicate it?
- Are the model and the process fully auditable (inputs, modelling, outputs and governance)?

(The categorization above is representational, some of the questions could apply equally to more than one stage of the modelling process.)

Disclaimer

The CRO Forum is supported by a Secretariat that is run by KPMG Advisory N.V.

Laan van Langerhuize 1, 1186 DS Amstelveen, or
PO Box 74500, 1070 DB Amsterdam
The Netherlands
www.thecroforum.org