MINISTRY OF SOCIAL DEVELOPMENT
TE MANATŪ WHAKAHIATO ORA

Data Science Guide for Operations

Model Development Lifecycle

MSD Reference: A13094446

October 2021

Contents

Introduction

There are many academic courses and resources to help the data scientist maximise new and emerging uses of data. The academic application of these is crucial in an operational setting, however it's not enough to ensure data is used to maximise opportunity without also effectively managing potential risk and harms.

As the need to manage potential risks and harms has increased, new principles and support frameworks have been developed in Aotearoa. Examples include the Māori Data Sovereignty Principles[1], Ngā Tikanga Paihere and the 5 Safes Framework[2], the Principles for Safe and Effective Use of Data and Analytics[3], the Data Protection and Use Policy[4], and the Guidelines for Trusted Data Use[5]. There are also several different ethics review boards available for government departments to use when applying new and emerging uses of data.

Organisations usually have internal processes and frameworks to manage and document their approach to the issues around privacy and ethics. A good example is the Ministry of Social Development's (MSD's) Privacy, Human Rights, and Ethics framework (PHRaE)[6].

While these are crucial resources, most haven't been written with the data scientist in mind. This makes it difficult to know which resource to use and how to apply the principles in practice. In short, the publication of these resources alone isn't enough to manage potential risk and harms.

The intent of this document is to help the data scientist practically apply their academic knowledge and support frameworks.

It's a practical guide to help implement new and emerging uses of data in an operational setting (what we refer to as **operational algorithms**).

Using this document

This document assumes sufficient academic knowledge of data science techniques. It's a practical guide for the data scientist implementing operational algorithms.

Implementing operational algorithms requires effective collaboration between many people and teams. The data scientist is a key part of the development process and the first stage in managing risk.

This document covers:

1. data science methods in operation – the work you'll do in your team
2. data science integration with your organisation – the work you'll do with other teams
3. data science application of principles and frameworks to manage potential risks and harms.

[1] Māori Data Sovereignty Principles Te Mana Raraunga

[2] https://data.govt.nz/toolkit/data-ethics/nga-tikanga-paihere/

[3] https://www.privacy.org.nz/assets/New-order/Resources-/Publications/Guidance-resources/Principles-for-the-safe-and-effective-use-of-data-and-analytics-guidance3.pdf

[4] https://dpup.swa.govt.nz/

[5] https://www.aisp.upenn.edu/wp-content/uploads/2019/08/Trusted-Data-Use_2017.pdf

[6] https://www.msd.govt.nz/about-msd-and-our-work/work-programmes/initiatives/phrae/index.html

Data science methods in operation

Idea formulation, selection and planning

Overview

Operational analytics is a key component in supporting automated (or semi-supervised) decision-making for organisations. To make effective use of analytical tools, an organisation must understand what business opportunities can potentially benefit from the application of operational analytics, and how to select and prioritise these opportunities. This is critical to a good return on investment in analytical solutions and to ensure those solutions are fit for purpose.

The use of advanced analytics to solve business problems (including machine learning and artificial intelligence decision-making systems) has grown rapidly in recent years. Organisations have invested heavily in developing their data and analytics capabilities and feel a need to justify the investment through the creation of analytical solutions to all business problems. A potential pitfall of this situation is that it sometimes leads to over-engineered solutions where a simpler and cheaper solution would have worked. The need to justify investment shouldn't drive the application of analytics – the nature of the business problem itself should drive the solution.

This section of the guide explains how to avoid such pitfalls and how to apply operational analytics to the right business opportunities. It covers:

- types of problems well suited to analytical solutions
- properties of a good selection and prioritisation process
- key steps in an example process.

It is important from the outset of the project to get the appropriate level of governance in place, including consideration of legal, ethical and privacy issues.

These will take time and consultation to get right, but there may be significant delays if the requirements need to be revisited in the final stages of sign-off.

Roles and responsibilities

The required roles in analytical idea formulation, selection and optimisation are:

- business owner
- data scientist
- ethics and privacy specialist.

There are likely to be other people who discuss alternative solutions, including process and IT staff. However, here we focus on staff whose job it is to identify business problems, define analytical solutions, and prioritise problems and solutions.

The role of the business owner

The business owner will involve the appropriate people from the business as required. These may include frontline staff, branch or regional managers, and senior level decision makers. Their key tasks are to:

- identify the critical business problems and opportunities for improvement
- clearly explain why these are critical and why the business requires solutions now

- explain who will benefit, and how much they will benefit
- sense-check potential solutions
- decide which ideas should proceed into development.

The role of the data scientist

The key tasks of the data scientist are to:

- listen carefully to the business problems and understand why they are problems
- identify which problems do and do not have potential analytical solutions
- formulate potential solutions
- refine solutions and do initial assessments of viability
- help the business forecast the true costs, risks and benefits associated with analytical solutions.

The role of the ethics and privacy specialist

The key tasks of the ethics and privacy specialist are to:

- understand the goal of any solution
- do an initial ethical and privacy assessment of solutions (this will likely not involve using the full version of your organisation's privacy framework)
- help the business forecast risks associated with analytical solutions.

Main ideas

Types of problems well suited to analytical solutions

The typical features of a business problem that might be solved using operational analytics are:

- high-volume, low-value decision-making
- where a fast or immediate response is required
- where a response must be available 24 hours a day, 7 days a week
- where there are doubts about the consistency of human decision-making.

There are other, more specialised, problems that need specific kinds of analytical solutions. For instance, black box models, like deep neural networks or boosted decision trees, might be better suited where accuracy is paramount, and transparency and interpretability of the model are not important. Some models may be best suited for specific applications – for example, image recognition is almost exclusively the domain of neural nets, and text generation and natural language processing are the domain of certain niche text analytic algorithms.

Properties of a good selection and prioritisation process

The hallmarks of a good selection process for projects that suit advanced analytics solutions are outlined below.

The process should be led by the business

Business teams have the best understanding of what their opportunities and priorities are. Therefore, business problems need to be formulated by business teams and communicated to the analytics teams to see if these can be solved in an efficient way. This doesn't mean that analytics teams can't initiate business change, in fact proof-of-concepts and demos from analytics teams can create new ideas and opportunities from business teams. However, the process of change should be initiated by the business and be driven by their needs.

The process isn't about finding problems analytics can solve

The process is about prioritising and solving business problems. There are several ways these can be dealt with and analytics is only one of those possible solutions. The business problem should drive the solution and not all business problems need to be addressed through analytics. In many cases, the opportunity can be realised with business process change or a new set of business rules without the use of analytics.

Other possible solutions should be considered

Artificial intelligence (AI) machine learning and advanced analytical solutions tend to be hard to build, implement, interpret and maintain. Therefore, other solutions must be considered alongside analytics. If the advantage offered by an analytical solution is only marginal compared to other solutions, then the trade-offs must be carefully considered, and the solutions should be prioritised based on the expected returns or benefits from each.

The most intricate solution isn't always the best

Simpler solutions might be cheaper, more interpretable and carry far less risk than more sophisticated and complex solutions. For instance, in many day-to-day analytical problems, a linear regression may be more interpretable and easier to maintain than a deep neural network, for a small trade-off in model accuracy. Complex algorithms are best kept for business problems where no other simpler solution can be reasonably applied.

Steps in an example prioritisation process

Step 1: The business identifies its most important problems

The business clearly articulates what questions need answering, or what opportunities it would like to pursue to obtain good outcomes for the service users and the organisation. These may be questions of improving efficiency, or targeting a new subpopulation for services, automating a business process or other similar opportunities. The desired outcomes must be quantified; these might be measures of customer satisfaction, fairness of process, transparency and interpretability, reducing human-intervention, or speed of decision-making etc. This metric will help decide the scale of the opportunity and the extent of desired change.

Step 2: The business leads a discussion of possible solutions

The business leads a workshop where members from various teams in the organisation discuss potential solutions and pick candidate approaches. These groups will include not just the analytics team, but also specialists from policy teams, IT and business leaders that could propose various solutions that may or may not include analytical modelling approaches. The discussions should focus on prioritising solutions in terms of impact, complexity, ease of implementation and maintenance, cost, and transparency.

Step 3: If analytics are proposed, perform a viability assessment

The business performs an initial viability assessment of the proposed analytical model that focusses on the following aspects:

- availability and quality of data
- ethics and privacy considerations of using that data and the overall solution
- technical viability of the solution

- IT implementation of the solution and ongoing maintenance
- people and process change required to make the solution work in practice.

This viability assessment determines any potential problems with the analytical solution before any large investment is made. Concerns and mitigations regarding viability must be documented and added to potential risks to the project.

Step 4: Develop a business case that covers the risks, costs and benefits

The results of the initial assessment, including identified risks, are documented in a business case. The assessment also helps to quantify costs, which will likely include:

- an ethics and privacy assessment
- data preparation
- designing and building the analytics solution
- documentation
- internal or external review
- deployment and testing of the solution
- people and process change
- development of suitable communications
- early-life support
- ongoing maintenance.

Benefits from the project may be in the form of efficiency gains which are relatively easy to determine in monetary terms, however not all benefits may be easily valued in dollars. There may be intangible benefits such as speed, improved trust, transparency, fairness and customer satisfaction. The business case must clearly state the expected benefits and propose measures that may be used to value both tangible and intangible benefits.

Another key consideration is that analytical solutions tend to have a high initial cost and lower (but not zero) ongoing costs, which means the return on investment will be seen only over time. The life expectancy of an analytical solution will need to be determined and the ongoing maintenance costs should be factored in to determine the overall return on investment.

Finally, like any new solution that’s deployed into a practical situation, there’ll invariably be small tweaks once it is deployed. These will need to be made in the early-life support phase, immediately after the solution goes live. This differs from ongoing maintenance in that the changes could be more effort intensive. This means the cost of early-life support must also be factored into the business case.

Model selection and optimisation

Overview

Selecting the best analytical model from several candidate models will depend on several criteria, of which accuracy is only one. Transparency and ease of implementation are the other main criteria and can be of critical importance for an analytical algorithm. Fairness should also be used to assess the list of candidate models (see *Fairness and bias*).

As there are plenty of resources available that discuss model accuracy in detail, this section will emphasise practical aspects that need to be considered during model selection, and only refer to aspects of accuracy that relate to model implementation.

Roles and responsibilities

The required roles in model selection and optimisation are:

- data scientist
- business owner
- analytics owner.

The role of the data scientist

The data scientist plays the most important role in the model selection and optimisation. The key tasks of the data scientist are to:

- short-list the modelling methodologies best suited for the context
- identify the model evaluation metrics
- perform model training and validation
- create tuned candidate models
- work with the business owner to identify the best model for the business context
- document the model tuning and selection process, along with any assumptions made
- complete the sign-off documentation.

The role of the business owner

The role of the business owner is to support the data scientist with the business perspective and contribute to model selection by comparing the model metrics for the candidate models. The key tasks of the business owner are to:

- provide the business perspective on the model purpose and maintenance
- prioritise the model metrics based on their importance in the business context and the associated trade-off with accuracy.

The role of the analytics owner

The analytics owner is responsible for:

- sign-off of the approach to model selection and optimisation.

Main ideas

No single model type is always the best

Each model type has strengths and weaknesses. The model that is best-suited will depend on the business context, the type of problem and the available data. It also depends on the weighting given to transparency and fairness, ease of interpretability and implementation. Table 1 compares several model types that are often used and demonstrates the trade-offs between different properties.

Table 1: Properties of common predictive models.

Type of model	Transparency and interpretability	Ease of implementation	Accuracy tendency[7]
Business rules	High	High	Low
Simple tree	High	High	Low/medium
Linear regression	High	Medium	Medium
Polynomial regression	Medium	Medium	High
Gradient boosted tree	Low	Low	High
Random forest	Low	Low	High
Neural network	Very low	Low	High

It is important to note that transparency and interpretability doesn't just refer to how easily something can be described to the general public, it also refers to the ability of the model developer to understand and explain the model's behaviour. It's important for the model developer to perform sense-checks on the model outputs and ensure there are no unexpected idiosyncrasies in the behaviour. For this reason, it's common to begin by building simple, easy-to-interpret models.

The ease of implementation and maintenance of a model will depend on how it is being implemented. The values in Table 1 relate to when the model is being implemented in a separate platform to the one it was developed in. For instance, if a model was developed in an analytical tool, such as SAS, Python or R, but implemented in a core business system, it will be harder to manage and maintain than if the same platform is being used throughout. This setup is common because analytical tools rarely have the necessary service levels and support to be relied on to make real-time decisions.

Note that the ease of implementation is likely to be high for all model types if the same system is used throughout.

The accuracy of different models will also depend on the relationship between the inputs and the target variable. If it is a reasonably smooth relationship then regression and neural networks will tend to work well, whereas if there are big jumps or discontinuities, the tree-based methods will tend to perform better.

In practice, since the bulk of the effort in building a model is in the preparation of the data, the general advice is to build several candidate models irrespective of other considerations for the purposes of benchmarking and comparison. For instance, even if neural networks should not be used in a business context for reasons of transparency, it might still be helpful

[7] '*Accuracy tendency*' rather than '*Accuracy*' because the actual accuracy depends on the business problem and data.

to build one as a comparison model to determine the differences in model performance and accuracy.

Appropriate measures for model selection

Each type of predictive error has different implications in practice

Measures of accuracy such as R-squared and the misclassification rate treat the different types of errors (eg false positives and false negatives) as if they are of the same importance. However, in practice, different errors will tend to have different real-world implications.

For instance, consider an analytical model for benefit fraud that flags potential cases of suspicion. A false negative would lead to non-detection of an actual benefit fraud. On the other hand, a false positive may lead to wrong conclusions regarding a beneficiary which may have much larger personal implications for the beneficiary. There might be a greater need to reduce this error at the expense of compromising on the false negative error rates or vice versa. The data scientist needs to take these outcomes into account when evaluating the accuracy of a model and appropriately weigh the different types of errors (see *Fairness and bias*).

Hold-out datasets should be used to evaluate models

Whether it be operational models or hypothesis testing, it is always important to use part of the data as a hold-out set. Any data that's used in training a model or formulation of a hypothesis must not be reused for validating the model. It is important to note that this requirement is more subtle than simply discarding the data used to train a model. Any data that was used to make decisions regarding the model design must not be reused, even if the data wasn't used to train the model.

A hold-out dataset is always recommended to enable the data scientist to optimise the complexity of the model and avoid overfitting. Conventional machine learning approaches recommend using training, validation and test datasets, and the relative sizes often depend on the nature of the business problem and type of validation.

In situations where there are only a limited number of past observations, there are still ways to test the model using independent observations through random sampling. These approaches include only using a training and validations dataset or using random bootstrap sampling or cross validation to make the most of the limited data available.

Steps in selecting the best model

Step 1: Construct a grid of candidate models and their attributes

To select the best model out of many candidate models, it is good practice to construct a grid of candidate models and their attributes. Table 2 shows a sample grid. It must be noted that the fairness evaluation will typically only be done for the best models because this evaluation can require a larger effort.

Step 2: The business owner and data scientist choose based on the attributes

The business owner and the data scientist must work together to objectively choose the best model, based on the weighting given to each attribute. The process must be systematically documented, and the assumptions clearly stated. It's likely that simpler models may have higher chances of being selected unless the differences in accuracy are very large and accuracy is of primary importance.

Table 2: Example grid for evaluating models (values depend on the data and business problem).

Model ID	Model type	Weighted accuracy	Level of transparency	Ease of implementation	Fairness
A1	Simple decision tree	0.83	Very high	High	Medium
A2	Decision tree	0.84	High	High	High
B2	Logistic regression	0.85	High	Medium	Medium
B4	Neural network	0.88	Very low	Low	High

Data preparation

Overview

The largest effort spent during model development is in data preparation. The transformations applied on the data are key to the success of the analytical model. These transformations, in turn, depend on the data scientist's understanding of the business context, biases in the data and quality considerations. While there are several 'automated' data cleaning and preparation programs available in the market, the highly contextual nature of data preparation often requires the data scientist to directly design these transformations.

The accuracy of an analytical model usually depends more on data preparation methods than on the model type or the model tuning procedures.

Some of the key decisions that must be made in preparing the data are as follows:

- Is the historical data relevant to the business context, or are there marked differences in the business process that need to be accounted for?
- Are there inherent biases in the data, and how will these affect the model outcomes?
- What variables should be used as inputs to the model and what is the target (dependent) variable?
- Are there useful proxy variables that can be constructed to represent unobservable variables?
- Are there any errors or missing values that should be accounted for?
- What transformations need to be applied to the data?
- Are there any influential outliers?
- Is there a need to account for rare occurrences or imbalances in the data?

As with most of the decisions made in building an operational algorithm, the answers to these questions rely on understanding the business objective. For some analytical modelling problems, a literature review of similar case studies may also inspire new ideas for data preparation.

This section discusses the common themes to be considered while preparing data for analytical modelling. This isn't an exhaustive prescription of how to perform data preparation – it's only intended to showcase commonly encountered scenarios in operational analytics.

Roles and responsibilities

The required roles in data preparation are:

- data scientist
- business owner
- analytics owner.

The role of the data scientist

The data scientist plays the most important role in data preparation. Their key tasks are to:

- conduct exploratory analysis of the available data and identify assumptions that need to be made
- work with the business owner and subject matter experts (SMEs) to understand the business context of each variable to be used in the model
- identify and short-list the variables to be used as the independent and dependent variables

- create and implement data transformation mappings and associated business rules
- document the assumptions and data transformations
- complete the sign-off documentation.

The role of the business owner

The role of the business owner is to provide the data scientist with the business perspective on the data. The business owner will also need to ensure that SMEs are available to explain the variable meanings and context of data collection. Their key tasks are to:

- provide the business perspective on data to be used
- review the data transformations to ensure the transformed variables adequately represent the purpose of the model.

The role of the analytics owner

The analytics owner is responsible for:

- sign-off of the approach to data preparation.

Main ideas

A representative dataset is critical

The two key criteria for the data to be used in creating an analytical model are:

- it is representative of the population that the model will be applied to
- it is of sufficient size.

In many cases, these criteria create a trade-off between each other. Often, the easiest way to increase the sample size would be to include data from further back in time. However, it is more likely that older data won't be representative of the current business context. In other cases, the opposite problem exists – the dataset is too large to be practically used. This is a better problem to have. If this is the case, a representative sample should be taken.

The required size of a dataset depends on the business context and the intended model type. The representativeness/size trade-off should be considered, and compromises must be made regarding model type depending on the amount of data available.

Categorical variables with many possible values or with rare values tend to push up the size of dataset required by increasing the dimensionality of the dataset. Several data transformation techniques exist to encode such variables and mitigate these problems at the expense of accuracy or sample size.

Even with very large datasets it may be possible that there are parts of the input space where there is too little data for the algorithm to be reliable. In this situation, the algorithm shouldn't be used, and human decision-making should take its place.

Use of proxy target variables

In many cases, the target variable that is to be predicted is not readily available due to limitations in the available operational data. In these situations, a proxy variable is often used in place of the actual target variable for the purposes of training the analytical model.

Examples of proxies that are commonly used include:

- Past outcomes of a business process (eg past decisions) are used as a proxy for what the outcome should have been. In this case, the decision, based on perceived risk/benefit, is considered a proxy for the actual risk/benefit that resulted from the decision.

- A payment or indirect outcome can be used as a proxy for a real-world event or outcome. An example is the use of an unemployment benefit receipt as a proxy for the state of being unemployed.
- Restricting an outcome to a specific time window when the true outcome can occur over a much longer time period.

Whenever a proxy variable is used, the rationale must be clearly articulated and documented. The associated documentation should also include a commentary and an analysis of the limitations of the proxy variable. The goal of the analysis is to understand how to interpret the analytical model output and what impact the use of the proxy variable will have on the model outcomes.

Proxy input variables are both useful and dangerous

Proxy input variables can be very useful when building algorithms using administrative data because the underlying quantity may be unobservable or not available as part of the data collection mechanisms. However, it is important to understand the limitations of these proxies to represent the true quantity and any systematic differences must be considered and documented.

The meaning of these proxy variables may also differ based on the context in which these are used. The input from the business owner or SMEs can be valuable in better understanding these proxy variables and what they represent, as well as any biases that the data scientist should be aware of while building the model or interpreting the results.

Care must be taken to ensure that none of the input variables are inadvertently acting as proxies for variables of concern that could bias the outcomes from the model. Common examples are gender or ethnicity, which might be overloaded with other unobservable effects. Any such variables should be carefully dealt with, and the relationships to the model outcomes should be studied.

Data errors can affect the algorithm in many ways

It is inevitable that data will contain some errors. These errors may be random or systematic. Systematic errors can lead to bias and are more difficult to treat (see *Fairness and bias*).

Data errors in the historical data used to build the model will affect the operational accuracy of the model and will have an impact across the whole population. For each variable used in the model, the different sources of error should be considered, mitigated (if possible) and documented.

It is also important to note that the historical data used for training and validating a model is often subject to correction and cleaning of errors, which have an impact on how the model works. The model expects similarly cleaned data when it is used for scoring. For this reason, it is mandatory to apply the same data preparation rules on the data to be scored when the model is in production.

While in production, the model is guiding decisions on individuals or groups. Errors in the data to be scored will lead to real-life impacts. This error is of high concern since it can lead to a different outcome for the individual or group. The data scientist should investigate the frequency of errors in the data that is scored and ensure that any variables that are vulnerable to a lot of errors are removed or cleaned appropriately.

Missing values need to be addressed in both the training and scoring data

Missing data is a common problem and many imputation approaches have been developed to mitigate it. The exact method to be used depends on the extent of missingness, the reason behind missing data (whether it is random or not), and the business context. Several automated methods to deal with these exist, but the choice of technique should be a judicious consideration after studying the data.

The imputation will need to fill in missing values in the scoring data as well. Analysis of the missing values in the scoring data should be carried out to ensure they have the same source as missing values in the training data.

There is also a human consideration when addressing missing data. For instance, if the model replaces an application process that used to involve a person interacting with the applicant, there may be several changes or discussions that lead to capturing and correcting the missing data. However, these nuances disappear when the model replaces the human element. This may lead to new patterns of missingness that weren't noticeable in the historical data and lead to unexpected behaviour of the model. It is necessary to carefully understand the process of data capture that may lead to missingness in data.

Transformations can help, but pragmatism may be more important

Variable transformations can often help increase the accuracy of the algorithm by reducing the influence of outliers or approximating a normal distribution. For many algorithms, ensuring numeric features are of similar scale avoids those with high values having undue influence. However, it is possible that the use of transformations may make the model harder to explain. In this case, the data scientist may need to focus on the pragmatic solution and make compromises rather than strictly satisfying theoretical norms. A good example is the use of principal components analysis to reduce multicollinearity and extract usable information from a dataset and use it for modelling. Doing this complicates the ability to interpret the model and makes it unusable for explaining model behaviour. It is important to balance pragmatism with theoretical considerations and accuracy of analytical models.

Stability tests can uncover multicollinearity and outliers

Multicollinearity and outliers can affect the stability of equation-based models. The data scientist must use standard tests (such as the dffits, dfbetas, vif and collin options in SAS), particularly when using a regression model. These allow the identification and removal of high-leverage points. It will also allow any stability issues, caused by using two highly correlated input variables, to be identified and addressed by removing one of the variables.

Treating imbalance in target variable values

In situations where there is huge imbalance in the target variable values, or in the case of predicting rare events, the accuracy of analytical models can be improved by under-sampling or oversampling the data. This is particularly true of machine learning algorithms used for predictive modelling, such as neural networks. There are more sophisticated techniques for treating imbalanced datasets, like the SMOTE algorithm that creates synthetic minority class data for better balancing between the target variable values.

Fairness and bias

Overview

The fairness of algorithms and their usage is a rapidly growing concern in operational analytics. This is closely linked to the topic of algorithmic bias but is part of a wider generalisation of the concept. Even unbiased algorithms can be used in an unfair way or result in unequal outcomes due to differences in the underlying rates for different populations. Fairness of an algorithm can be measured in several ways, each quite different from the other and sometimes with conflicting definitions. It's rarely possible to satisfy all measures of fairness at the same time. Achieving a higher degree of fairness in an algorithm will also involve a trade-off with accuracy.

Until recently, accuracy was often the only consideration in model evaluation. However, recent concerns that algorithms can reinforce inherent prejudices has led to new considerations for model evaluation, including measures of fairness and transparency. One of the main criticisms of algorithm-based decision-making is that it can 'bake in' existing biases in the system and propagate them into the future. This concern must be acknowledged. Most datasets are inherently biased in some way and machine learning or statistical approaches are designed to find these systematic patterns in the data. This means there is a need to carefully study the data and understand these biases before automated model building approaches are applied to administrative data.

It is important to understand the key difference between fairness and algorithmic bias. While bias focuses on the model alone, fairness relates to the overall outcome of the algorithm, the business process, the underlying prevalence in the populations and the data generation processes[8]. Algorithmic bias tends to receive closer attention because the roles involved in the algorithm development often have a narrow focus on the outcome of the model alone, rather than the overall business context and operations. Also, it tends to be easier to address algorithmic bias than designing a 'fair' algorithm, because the definition of 'fairness' may change depending on the context. If the context changes, for example a model initially designed to target assistance is used as part of a fraud detection system, the notion of fairness may also change.

The measures of fairness to be considered in an analytical model will be driven by the business goals. In some cases, accuracy will still be the priority, particularly if this already allocates resources or services to under-represented subgroups in a fair manner or if the underlying rates are reasonably well-balanced across various subgroups. In most other cases, it is necessary to consider specific fairness measures as part of model evaluation.

A few important fairness metrics[9] are listed below. While this is not a complete list, these are a good starting point for defining measures of algorithmic fairness based on business context.

- Equal threshold (the same threshold is applied to everyone even though the accuracy, false positive rate and false negative rate may be different for different groups).
- Parity (ensuring the same proportion of each population are classified by the model as positive). For example, the same proportion of males and females classified as likely to receive services.

[8] Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: a critical review of fair machine learning. arXiv:1808.00023v2 [cs.CY]. https://arxiv.org/pdf/1808.00023.pdf

[9] Verma, S., & Rubin, J. (2018). Fairness definitions explained. 2018 ACM/IEEE International Workshop on Software Fairness. Gothenburg, Sweden. https://www.ece.ubc.ca/~mjulia/publications/Fairness_Definitions_Explained_2018.pdf

- Equal odds (ensuring the true positive rate and false positive rate will be the same for each population).
- Equal opportunity (ensuring the true positive rate is the same for each population).
- Equal accuracy (ensuring the same percentage of people will be correctly classified for each population).

Roles and responsibilities

The required roles in ensuring algorithms are fair are:

- data scientist
- business owner
- privacy and ethics specialist
- analytics owner.

The role of the data scientist

In addition to developing the model, the data scientist has a key role in ensuring transparency and fairness in the modelling process. Their key tasks are to:

- identify, quantify and document bias
- create a fairness overview, which explains in lay-person’s terms what measure of fairness was chosen and why
- communicate the risks to the analytics owner and business stakeholders
- work with the business stakeholders to formulate mitigation options for addressing fairness and bias
- implement the identified changes
- fill in the sign-off documentation.

The role of the business owner

The role of the business owner is to identify the implications of the bias in the context of the business problem and articulate the goal of the business. Their key tasks are to:

- prioritise the fairness measures for the business problem, and the associated trade-offs with accuracy
- provide the business perspective on the bias and its implications
- provide input into the decision about which of the mitigation options suit the business context.

The role of the privacy and ethics specialist

The role of the privacy and ethics specialist is to ensure the use of the data is consistent with organisation’s policies and gather the information required by the PHRaE framework. Their key tasks are to:

- assist in categorising the variables based on degree of concern
- assist in identifying proxies
- assist the assessment of biases and what remedial measures to take
- complete the bias and fairness related sections of the PHRaE.

The role of the analytics owner

The role of the analytics owner is to formally accept the approach to ensure fairness. The role requires an adequate understanding of the potential biases, the mitigation approach, and the impact of the trade-offs made to achieve this. The key tasks of this role are to:

- understand the biases, the mitigation approach and the impact
- probe the team with relevant questions, ensuring the risks are well-understood and within reasonable thresholds
- sign-off the approach to fairness.

Main ideas

This section discusses the steps to assess and ensure algorithmic fairness. These steps are intended as a guide and should be adapted to suit the context of the business problem. At each step, a peer review is recommended to ensure quality and robustness of the approach. Also, not all roles listed earlier would be required in each step.

Steps in assessing fairness and bias

Step 1: Determine which fairness measures are to be addressed

Roles involved: data scientist, privacy and ethics specialist, business owner, analytics owner.

Depending on the fairness considerations identified in the analytical modelling approach, the project team will choose the fairness measure that needs to be optimised. Any trade-off with accuracy or other modelling aspects should also be included in this decision. The fairness measure chosen will heavily depend on the business context, the underlying data and the desired outcome from the model.

Step 2: Determine the degree of concern for each variable

Roles involved: data scientist, privacy and ethics specialist, business owner, analytics owner.

The data used in the analytical model could be a major source of concern in ensuring algorithmic fairness. Each variable in the data needs to be carefully studied to determine what it represents in the business context and how it could bias the outcomes.

Variables can be categorised into groups based on their potential impact to fairness:

- High concern – variables where bias is not acceptable.
- Understanding differences – variables where differences are present and need to be understood, but bias on that variable is not prohibited in the context of this business problem.
- No concern – variables where differences in outcomes are expected and of no concern.

Table 3: Variable classifications.

Degree of concern	Characteristics
High concern	No significant bias is acceptable. Variables that could be acting as proxies for unobservable characteristics will need to be identified. Extra care required while modelling minority subpopulations.
Understanding differences	Bias is acceptable (and sometimes desirable, such as when services are directed towards traditionally disadvantaged groups). The bias will need to be understood though. The bias may influence the final model or its implementation. Variables and their proxies can be used as inputs into the model. If majority (or traditionally advantaged groups) are found to be advantaged by the algorithm, then it may be advisable to move the variable to the 'High Concern' category.
No concern	Input variables that do not act as proxies or add any known biases. Often include undisputable/unbiased facts about a subject's past.

The degree of concern for each variable comes from the PHRaE framework and depends on the business goal. For example, if the goal is to offer extra services to support a specific subpopulation then there may be no variables in the 'High concern' category. If the business goal is to take away services based on subpopulation characteristics, then there may be several variables in the 'High concern' category.

Protected variables such as ethnicity, gender, age and sexual orientation will always fall into the 'High concern' or 'Understanding differences' categories. A few examples illustrating this have been added below.

Example 1: The Youth Service NEETS (YSN) algorithm

MSD's YSN algorithm is used to identify clients who can be offered extra services[10]. The client has the option to decide not to use these additional services with no impact on their entitlements. The goal of the algorithm is to merely suggest these services to the user and does not enforce services or exclude individuals.

It is likely that the YSN model will target a larger proportion of services to Māori, Pacific people, and women than their proportion in the general population. This form of bias is likely to be acceptable given that the business context is to offer extra services tailored for these groups. If, for example, Māori made up a very large proportion of the target population for these services it makes sense for the business to consider a service design that better suits the needs of Māori, or offer a version that is specifically designed to help Māori find employment.

In this example, variables like ethnicity and gender may fall into the 'Understanding differences' category, provided there is proper consultation with the business stakeholders and privacy and ethics concerns are addressed.

[10] https://www.msd.govt.nz/about-msd-and-our-work/work-programmes/initiatives/phrae/youth-service-for-neet.html

Example 2: A benefit fraud algorithm

A benefit fraud algorithm might be used to identify and flag cases of potential benefit fraud. If variables like ethnicity, gender or age changed the probabilities that a case could be flagged then this would be cause for concern. The algorithm would not allow the clients to opt out of the impact from this model.

In this example, variables like ethnicity, gender and age fall into the 'High concern' category. The team involved in the design of this algorithm would need to carefully consider the differences between groups (if there is a difference), analyse what drives those differences, and identify if any biases in past decision-making might be contributing to the difference.

Step 3: Analyse the relationship between variables of concern and model outcome

Roles involved: data scientist.

Once the variables are categorised, the next step is to estimate the extent of bias caused by these variables. A few exploratory tests can quickly shed light on this relationship, and these are provided in Table 4. The goal of these tests is to determine if there is a relationship between the protected variable and the observed or predicted outcome.

Test 1 is performed prior to modelling, and the subsequent tests are done after the Potential Final Model (PFM) has been identified. A binary target variable has been assumed here, but similar tests would apply to continuous target variables as well.

Table 4: Tests to find a relationship between variable of concern and observed or predicted outcomes.

	Test	Goal
Test 1	Frequency analysis - variable of concern versus target.	Determine if there are observable historical differences.
Test 2	Frequency analysis - variable of concern versus predicted decision (based on a realistic threshold).	Determine if there are significant differences in the decisions made by the algorithm.
Test 3	Frequency analysis - variable of concern versus accuracy (with breakdowns into Type 1 and Type 2 errors).	Determine if there are any differences in accuracy for different values of the variable. Determine if there are any differences in Type 1 and Type 2 error rates.

The results of these tests will demonstrate **one** of the following outcomes:

1. There is no relationship between the variable of concern and the observed or predicted values. This implies that the model is likely to be free from bias resulting from the variable.
2. There is evidence of a relationship that needs to be further investigated.
    a. There is a relationship between the variable of concern and the predicted values, and this relationship is *consistent* with the relationship between the variable of concern and the observed values.
    This would indicate that the bias in the analytical model results from the *bias in the input data* (a few possible options to address these are discussed in Step 4).
    b. There is a relationship between the variable of concern and the predicted values, and this relationship is *not consistent* with the relationship between the variable of concern and the observed values.

This would indicate that the *modelling process* may be introducing bias to the outcome. The next step would be to isolate the part of the process causing the bias. To do this, a second round of tests are used (shown in Table 5).

Table 5: Tests to isolate which part of the modelling process may be introducing bias.

	Test	Goal
Test 4	Build model using only the variable of concern and the same type of model as the PFM.	See how the observed differences translate into a model of this type. This is easier for simple statistical models.
Test 5	Add variable of concern to the PFM and assess whether this variable has a significant impact on the model.	Determine if there are historical observed differences that can't be explained by other input variables.
Test 6	Build model using the same input variables as the PFM but make variable of concern the target. Repeat using a stepwise regression (or any easily interpretable model).	Determine if any of the input variables could be acting as proxies for the variable of concern and which input variables are the best proxies. The stepwise model is used in this case so that it's easier to conduct this determination.
Test 7	Build model using only the variable of concern as the independent variable, with the predicted probabilities of the PFM as the target.	Determine if the observed differences found in Tests 1 and 2 are translating into the predictions of the model.
Test 8	Build model using only the variable of concern, with the predicted decision (for a realistic threshold) of the PFM as the target.	Determine how the application of the threshold affects the relationship between the variable of concern and the outcome of the model.

Step 4: Think about the practical implications and possible solutions

Roles involved: data scientist, privacy and ethics specialist, business owner.

To understand the implications of the bias, the results from the previous step need to be combined with the subject matter expertise of the business owner and the organisation's processes. Some of the questions that need to be addressed are:

- Is there a strong relationship between the variable of concern and the target?
- Does the strength and direction of this relationship have serious impact in the context of the business problem?
- Do the tests listed in Step 3 provide insights into why the observed relationship exists?
- Could the observed relationship be caused by biases in past decision-making processes?
- Are any of the independent variables acting as proxies?
- Could these proxies be influenced by biases in past decision-making processes?
- What are the implications of an incorrect outcome for the individual/group?

Bias tends to be entrenched in society and in the administrative data we collect. In a model with many predictors available, the impact of a protected variable such as ethnicity is likely to be expressed through correlations with other variables such as social deprivation[11]. Removing the protected variable from the model may not address the bias, and if you do need to make an adjustment based on that variable it effectively becomes part of the model anyway.

If the team agrees that an unacceptable level of bias exists in the model outcome, then recourse needs to be considered. There are several approaches that can be used to address these, either on the data itself or on the analytical model. Some options are listed below but the appropriate one will depend on the business context:

- Exclude any observations that are thought to be biased.
- 'Correct' any observations that are thought to be biased.
- Exclude any proxies that are thought to be biased.
- Adjust the parameter estimates (or equivalent parameters) manually.
- Use different thresholds for different groups.

Sometimes there may be no data available on the variable of concern or the data that is available is unreliable. For example, ethnicity may not be known or may be unreliable. This should be factored into the discussion and, even if no data exists, it is important to show that the possibility of bias was considered.

It is not tenable to argue that the existing process (based on human decision-making) includes bias, therefore the algorithm should also be allowed to include bias. Using algorithms to remove bias, or continually reduce it, should be the goal.

Step 5: Decide on approach for each variable of concern

Roles involved: data scientist, privacy and ethics specialist, business owner, analytics owner.

Once the team has agreed how to address biases in the analytical model, the appropriate options, impact and conclusions need to be presented to the analytics owner. The aim of this step is to ensure the analytics owner has a complete understanding of the reasoning behind the suggested approach and its business impact. This understanding is a pre-requisite for the analytics owner to sign off the model as being ready to go into production. They must be confident that the potential biases have been identified and accounted for. If the analytics owner is not convinced of the approach to bias mitigation, new data must be found, or the algorithm will have to be revisited.

If required, the analytics owner must consult with internal and external analytics and business experts to ensure the approach is acceptable. If the risk of bias is considered high, an external technical or legal review can be considered.

The documentation created as part of this presentation will form the core of the fairness documentation in the technical report.

Step 6: Implement the selected solution

Roles involved: data scientist.

The agreed and approved solution is then implemented by the data scientist and any potential shortcomings and assumptions are documented for reference.

[11] Williams, B., Brooks, C., & Shmargad, Y. (2018): *How Algorithms Discriminate Based on Data they Lack: Challenges, Solutions, and Policy Implications.* Journal of Information Policy, Vol. 8. https://www.jstor.org/stable/10.5325/jinfopoli.8.2018.0078

After the solution is implemented, the tests from Step 3 are repeated to ensure biases in the variables of concern are adequately addressed by the approach. If differences are found, this may call for revisiting the approach, and the process might need to be repeated from Step 3 onwards.

Step 7: Final approval for the analytical model

Roles involved: data scientist, privacy and ethics specialist, business owner, analytics owner.

Once the solution has been implemented and found to be successful in addressing questions of fairness, and all associated documentation has been created, the analytics owner will need to determine if the concerns have been adequately addressed.

A summary of fairness will need to be included in the pack for final approval.

Model and process maintenance

Overview

Operational algorithms usually require a large, upfront investment by the organisation. The benefits, on the other hand, accrue slowly over time. The upfront investment will only pay off if the algorithm is successful for a long period of time and if it does not need to be rebuilt regularly. Therefore, it is necessary to have an estimate of the life expectancy of a model upfront so that the return on investment can be estimated before a project starts. This is the first key piece of information needed to effectively monitor and maintain an algorithm.

In addition, the benefits will only be achieved with careful planning during the building phase, learning during early-life support, and regular maintenance when the algorithm is in production.

The overall success of the algorithm will be affected by changes in the data, business processes, priorities and goals, as well as the external environment. These changes are difficult to predict at the outset, so part of the model maintenance involves ensuring the algorithm adapts to the changing business requirements.

This section outlines different update approaches, when they should be used, and what the monitoring and maintenance plan should contain.

Roles and responsibilities

The required roles in model and process maintenance are:

- data scientist
- business owner
- analytics owner.

The role of the data scientist

The data scientist has a key role in determining the model and process maintenance setup. Their primary tasks are to:

- collate the information required to estimate the algorithm's life expectancy
- calculate the value of the algorithm over its lifetime and write up the results of this discussion
- formulate the approach to the algorithm update
- write the monitoring and maintenance plan
- fill in the sign-off documentation
- carry out the monitoring and maintenance plan.

The role of the business owner

The business owner provides a key perspective on the life expectancy of the algorithm and provides other critical information. Their primary tasks are to:

- provide understanding of the benefits of the algorithm and the value of those benefits
- provide understanding of the likelihood and nature of upcoming changes within the business
- understand the monitoring and maintenance plan and what it means for the business.

The role of the analytics owner

The primary tasks associated with this role are to:

- lead the discussion of the algorithm's life expectancy
- sign-off the monitoring and maintenance plan
- ensure the resources are available and accountable for delivering on the plan.

Steps in maintaining and monitoring an algorithm

Step 1: Set a realistic life expectancy

Roles involved: analytics owner, data scientist, business owner.

One of the key requirements for an algorithm is the life expectancy required by the business. This should be addressed during the design phase. The requirement is often not explicitly specified but it needs to be. The factors that will feed into this decision include the following:

- The upfront and ongoing cost of building and maintaining the algorithm.
- The value of the benefits the business derives from the algorithm. These benefits could be direct cost savings or more difficult to value benefits, such as faster decision-making and service user satisfaction. The benefits need to be measurable before and after the model goes live so the change in value can be quantified.
- The speed of change of the business problem.
- Other upcoming changes in the organisation.

The analytics owner will seek input from the data scientist to help estimate costs and the business owner to understand the benefits and evolution of the business problem. Other enterprise-level business owners should be consulted about upcoming changes in the organisation and whether benefits to one area will result in costs in another. This will help produce a more realistic view of the cost/benefits, and in turn, generate a better understanding of the model's life expectancy. If the benefits can't be realised before the business problem starts to evolve then that indicates the model shouldn't go ahead.

Step 2: Choose an update setup based on speed of evolution

Roles involved: data scientist.

In the government context, most business problems evolve only very slowly or not at all. The speed of this evolution will dictate what type of setup is required for updating the algorithm. The three options are self-learning, slow evolution and no evolution. The data scientist will be responsible for determining which update method is appropriate.

Self-learning

A self-learning setup continuously feeds data into the algorithm and updates the model on the fly. This setup is the one that tends to get the most attention but it's also the least-used option.

Advantages:

- The algorithm can evolve quickly to account for changes in the data or business problem.
- The algorithm doesn't need a scheduled maintenance update.
- Those implementing the algorithm can't defer thinking about updates and must be fully prepared for updates to happen from the outset.

- The updating process is tested as part of the initial deployment when the most funding and resources are available.

Disadvantages:

- The algorithm constantly changes, making it significantly harder to explain how it works.
- There is a risk that the algorithm will learn erroneous patterns due to data errors and other issues.
- Any adjustments for bias are difficult to make as the bias constantly changes.
- It's harder to track which version of the algorithm has been used for any given decision.
- Only the model automatically adapts. The business process that goes with it does not.

The result of these disadvantages is that there needs to be a significantly more robust monitoring process in place to ensure the model is accurately predicting outcomes.

Slow evolution

Slowly evolving algorithms are static models that are periodically updated. This is the most popular approach in the government sector because it balances stability against the ability to adapt to changes in the business problem or data.

Advantages:

- Adapts to changes in the data and/or business usage.
- Easier to explain how the model works since it is not constantly updating.
- Easier to say which version of the model was used to make a particular decision.
- Easier to find, and adjust for, biases in the data or algorithm.
- Tends to have a longer life expectancy than the no evolution approach.

Disadvantages:

- Adaption may be too slow.
- Scheduled updates need resources and may not be prioritised.
- The update process may not be fully developed and tested during the initial development and implementation of the algorithm.

When using the slow evolution approach, at least one update should be done before the algorithm goes live. That ensures the update process is fully developed and tested before the algorithm is used.

No evolution

No evolution algorithms are static models that don't update. They suit business problems that are constant over time, such as rules determined by legislation. In practice, there is a risk that a slow evolution algorithm (see above), with no effective monitoring and maintenance plan, will become a no evolution algorithm. If this happens, the algorithm will need to be rebuilt every few years.

No evolution models still need regular performance monitoring.

Advantages:

- Simpler to implement.
- Easier to explain how the model works since it does not update.
- No version control needed.

- Easier to find, and adjust for, biases in the data or algorithm.

Disadvantages:

- No adaptation to changes in the data or business problem.
- No mechanism to stop accuracy dropping over time.
- Can easily be forgotten about, eventually leaving an algorithm making poor decisions.

Step 3: Build a monitoring and maintenance plan

Roles involved: data scientist.

The monitoring part of the plan is essentially the same irrespective of the updating setup. It should include the following:

- A monitoring report that is automatically generated and sent to the data scientist for checking to ensure there are no problems with the model. This report needs to show:
    - any significant changes in the distribution of outcomes from the model
    - any significant changes in the inputs to the model.
- A periodic, detailed check of the algorithm.
- A periodic assessment of the accuracy of the algorithm, and confirmation it's still solving the business problem.
- A clear method by which the users and/or business can notify the data scientist that they have observed an issue or have some concerns.

The two periodic steps are included even for a self-learning setup because, even though the model continuously updates, the process of the data scientist and the business assessing its performance only needs to happen at certain times. The monitoring report will need to be more comprehensive for the self-learning algorithm since both the data and model are changing.

The maintenance part of the plan will depend on the updating setup. It should include the following:

- Ad hoc maintenance to fix problems identified by the monitoring or the business. This involves fixing immediate problems, such as a problem with the flow of data into the model or with the flow of results out of the model.
- Periodic maintenance to address slow-moving issues. This involves slow degradation in the accuracy of the algorithm or small changes in how the algorithm is used by the organisation.
- A schedule of times for periodic maintenance.
- The person or role responsible for both ad hoc and periodic maintenance.
- A plan for communicating changes with stakeholders, service users and the public. With many of the algorithms implemented to date there are very rarely communications after the initial implementation. This can have a significant impact on the true transparency of the algorithm since the initial documentation will often no longer be accurate.

Decommissioning an operational algorithm

As with any IT product the algorithm will eventually need to be decommissioned, often to be replaced by another algorithm. The two main goals in this process are:

- to ensure a smooth transition for staff, service providers and service users
- to learn from the algorithm and what did and didn't work in practice.

A decommissioning plan will need to be developed to ensure these goals are meet. This plan will include:

- the goal of the algorithm
- why the algorithm is being decommissioned
- who will approve the decommissioning
- the date of decommission
- what the algorithm will be replaced with (if anything)
- if the decommissioning is because of shifting to a new platform, eg moving from on-premises to the cloud, then a migration plan should also be provided.

The data scientist can be responsible for the plan but will need to consult many different stakeholders to create it. Stakeholders include service users, service providers, communications, the data warehousing team, service delivery (IT) and records management. Additional IT support may be required if you have hardware that needs to be decommissioned, firewalls or whitelists to be updated, revoking certificates and so on. The earlier you start this process, the smoother the transition will be.

The plan should be reviewed by the relevant stakeholders and agreed upon. Once it is agreed, a list of activities detailing how the decommissioning will take place should be agreed along with which team is carrying out the tasks.

Stakeholders will be notified once the decommissioning process is complete. The document should be filed along with key documents such as the technical report, plain English explanation, any reviews etc, with a link to this information included in the communications. Any datasets and code should also be archived with their location noted in the decommissioning report. Your records management advisor will be able to take you through this process and ensure you comply with legislation. The code base should also be archived so that changes can no longer be made.

External review

Overview

External reviews are a key part of identifying and managing risks associated with operational algorithms. The two key goals of using an external review are:

- to answer specific questions about the algorithm where the organisation is concerned (or conversely to probe the algorithm for weaknesses where there are specific business, ethical or legal worries)
- to give management certainty that the algorithm is something they can stand behind (algorithms are large, complex, highly visible investments that senior management should expect to receive questions on).

Three different types of review are common for an operational algorithm, both before deployment and after it has been in operation for some time:

- Legal (L) – ensuring all legislation is complied with.
- Ethical (E) – the impact on people of data and algorithm use.
- Technical (T) – accuracy and validity of the models.

There are significant overlaps between these review types. The tables below give examples of the questions an external review may raise. Questions are arranged into basic, detailed and very detailed, and are categorised into the type of review that will address that question. It's unlikely that a review would look at all the questions below in detail, so settling on the scope of the review is important.

The decision about whether an external review is required or not will depend on the risks associated with algorithm, both for service users and the organisation. All high-risk algorithms should go through an external review process. However, if the risk is considered low and the potential impact on service users is low, an external review is unlikely to be needed.

One key difference between reviews of operational algorithms and other analytical work is the timing of the review. For operational algorithms, the review should be done earlier in the process. This allows any required changes to be implemented, tested and communicated, which is a much larger process for operational algorithms. The earlier timing also helps in the management of identified risks.

Roles and responsibilities

The key roles in an external review are:

- analytics owner
- data scientist.

The role of the analytics owner

The analytics owner has a key role in the external review process. Their primary tasks are to:

- liaise with the Technical Advisory Group to determine what, if any, external reviews are required
- determine the scope of the reviews and key questions
- procure the review
- ensure the reviewers have access to the necessary people and resources
- ensure the results of the review are actioned

- sign off the external review and resulting actions.

The role of the data scientist

The primary tasks associated with this role are to:

- provide documentation and other resources for the external reviewer
- answer the reviewer's questions
- action any changes that result from the review
- fill in the sign-off documentation.

Main ideas

Basic questions

The basic questions require Yes/No answers that focus on the process that was used. They are designed to show that the essential elements have been considered without delving into the detail. For organisations with a comprehensive privacy and ethics framework, such as MSD's PHRaE, most of the legal and ethical issues will be gathered in one place, but at other organisations these issues will be more distributed.

Table 6: Key basic questions an external reviewer may ask.

Question	Type of review
Is the sign-off process that has been followed clear?	All
Is there one person (usually at the Deputy Chief Executive level) that's responsible for use of the algorithm?	All
Has the privacy and ethics process been completed?	E
Is there a monitoring and maintenance plan in place?	T
Has transparency been considered?	T, E
Has fairness been considered?	T
Have affected parties been consulted?	E
Is there documentation of how the algorithm was built?	T
Is there documentation of how the algorithm was implemented?	T

Detailed questions

These questions go beyond simple Yes/No answers and look at the quality of work done. However, they usually won't involve looking at any data or code in detail.

Table 7: Key detailed questions an external reviewer may ask.

Question	Type of review
What are possible public reactions if this was to end up on the front page of a newspaper?	E
What advice did our in-house legal team give and was it followed?	L
What processes have been followed to ensure this work can be re-run in the future?	T
Have peer reviews and tests been carried out?	T
Is there an audit trail of what work that was carried out (eg in Jira tickets or some other mechanism)?	T
Is the frequency of the planned maintenance reasonable?	T
Will there be a suitable dataset available for updating the model?	T
Does the layperson's explanation accurately represent the technical version?	T
Is the layperson's explanation understandable by the service providers and service users?	T
Is the fairness measure used appropriate given the business problem?	T, E
Have the te ao Māori opportunities and risks been properly considered?	E
Have a range of approaches been tried and the most appropriate one selected?	T
Is the algorithm accurate enough to solve the business problem?	T
Has the algorithm been properly tested for biases and has appropriate action been taken to correct for biases?	All
Have the implications of incorrect predictions for individuals been properly considered? For groups of people?	All
Is the technical documentation sufficiently detailed to allow a data scientist not involved in the project to understand what was done?	T
Is the implementation documentation sufficiently detailed to allow the algorithm to be maintained and updated?	T
Is the code clearly commented?	T

Very detailed questions

These questions involve looking at the code, data and results of each used model. Only the most important models will receive a review of this detail and the review is likely to be targeted at specific areas of concern. Many different questions could be asked so only a limited number of example questions are given in Table 8.

Table 8: Key highly detailed questions an external reviewer may ask.

Question	Type of review
Are there any legal precedents for the use of data or a model of this type?	L
Has additional thought been given to vulnerable populations, such as children, victims of abuse and others?	E
Do the benefits of the model outweigh the risks (including social and psychological risks)?	E
Is there true consent to use the data to build a predictive model?	E
Has social license been considered?	E
Was advice required from an external legal team and was that advice followed?	L
Are there any concerns with data sovereignty, including te ao Māori opportunities and risks?	E/T
Have the code, data, secrets and other credentials been appropriately stored?	T
Is the test coverage appropriate for the model?	T
What issues were identified during testing and were they appropriately addressed?	T
Has the selected model been correctly optimised to be accurate in practice?	T
Could the accuracy of the model be improved?	T
Have prohibited variables and their proxies been used/not used in an appropriate way?	E

Steps for an external review

Step 1: Determine the goal of the review

Select one of the two goals in the overview above.

Step 2: Determine the type of review

Select whether the review will be legal, ethical or technical in nature.

Step 3: Determine the scope of the review

Select questions from Table 6, Table 7 and Table 8 or formulate your own similar questions. There are many possible questions for an external reviewer to look at so having a clearly defined scope is key to getting the best value out of your review process.

Step 4: Procure the review

Operational analytics is a very practical application of analytics to a real-world business problem. Selecting a reviewer with practical experience is critical to getting useful questions and advice.

Step 5: Implement and document actions based on the review

The data scientist and analytics manager should promptly respond to each question or issue raised by the reviewer. Any actions taken or not taken as a result of the review need to be documented. This is so the organisation can clearly show what the reviewer identified and how it responded.

Step 6: Decide if the review and response will be made public

Releasing the review and response to the public is a good way to build trust in the algorithm and process followed. For some types of model, it won’t be possible to release the full review. However, it should be possible to release any information the review contained on the process that was followed.

Working with the business unit receiving your service

Successful integration of new and emerging uses of data requires a data scientist to work with other parts of their organisation. These can be separated into two groups, the business unit they are delivering to, and the business units who support them to do this.

Model and process design

Overview

All algorithms are a combination of a model and business processes that integrate the model with the rest of the workflow. Some processes come before the model and are involved in collecting the needed data or triggering the algorithm to make a decision. Other processes follow the model and include informing the service user of the decision and triggering other processes.

A key component of this combination is how human and automated decision-making integrate. The integration ranges from fully automated to augmented human decision-making. In practice, it’s more common that the algorithm deals only with the high-volume, low-complexity decisions while people make the low-volume, high-complexity decisions. Ensuring that automated and human decision-making work smoothly together is critical to good user experience and making sure the overall approach is ethical.

The business has a key role in designing how the algorithm and processes work together since they best understand the workflow and how service users interact with the process. The design should bring together input from the business, the data scientist, service providers and service users.

The impact of a service or decision process, including a model’s role, may also be subject to formal evaluation. Like modellers, staff in research and evaluation teams need sufficient data to do their job and too often those needs are overlooked.

This section details the three common approaches to integrating automated and human decision-making:

- Fully automated decision-making.
- Partially automated decision-making.
- Augmented decision-making.

The strengths and weaknesses of each approach are discussed, as well as some of the ethical and technical issues commonly encountered.

Roles and responsibilities

The required roles in model and process design are:

- data scientist
- the business owner, and SMEs.

The role of the data scientist

The primary tasks associated with this role are to:

- work with business experts to determine the appropriate integration approach

- deliver the modelling results, to an automated system or to a human, where they can be used directly or picked up and passed on by another IT system
- help the business set thresholds
- convert the value gained from the model into a usable format for the business
- help the business to understand, control and monitor both the human and automated components of augmented decision-making
- fill in the sign-off documentation.

The role of the business owner (and SMEs)

The primary tasks associated with this role are to:

- bring the business perspective to the complexity of decision-making and the specialist skills of a person when required
- help the data scientist develop a systematic approach for streaming to the decision maker (human/automated)
- understand, control and monitor augmented decision-making
- sign off the approach to model and process design.

Main ideas

Fully automated decision-making

Fully automated decision-making is where 100% of decisions are made by the algorithm without human intervention. This approach is often accompanied by a human review or monitoring to ensure the algorithm is operating as expected. Fully automated decision-making is usually limited to simple problems of low importance or where extreme speed is needed. Examples include most business rules, decisions required while a service user is interacting in person with a staff member and, in the business world, automated trading systems which do trades in fractions of a second.

Advantages:

- Potentially very fast.
- Can be available 24 hours a day, 7 days a week.
- Low ongoing cost.
- Free from human bias.

Disadvantages:

- Can only be applied to simple problems.
- No option for a service user to explain their issues to a person and feel heard.
- Multiple ethical issues if the decision is an important one (lack of human oversight, the ability to opt out is not always available, and usually there is no other service alternative, which can raise issues of whether there is true consent).
- Doesn’t evolve with social and ethical changes (risks locking in the status at the time the algorithm is built).
- Can be difficult to update since obtaining a dataset to retrain the model is difficult.

In the New Zealand government sector, the only type of decision that would be fully automated is one that people don’t think of as an algorithm, such as business rules or other rule-based decisions. In these cases, there is no doubt about the outcome based on the inputs.

Key considerations when implementing fully automated decision-making

Even business rules are a type of algorithm (although they may be based on business knowledge rather than data). Therefore, they should go through the same ethical and fairness testing as any other algorithm. It is likely there are many rules-based algorithms in use, in both the public and private sector, that are potentially biased and haven't been subject to any ethical process.

The monitoring and maintenance approach must be more thorough for fully automated algorithms since only service users interact with the system. Without staff involvement, there is more opportunity for a problem to go undetected for a long period of time.

Partially automated decision-making

Partially automated decision making combines the strengths of automated and human decision-making. These algorithms can automatically make many simple, low-impact decisions. This allows human decision-making to focus on more complex decisions - decisions that are likely to have a big impact on the individual service user. The risk of a significant negative impact on an individual is further reduced by only allowing the algorithm to make positive decisions (where the algorithm would have made a negative decision, the decision is made by a person instead). In this situation, the algorithm is acting to fast-track positive decisions. The claims approval process at ACC is an example of this setup in practice.[12]

When using partial automation, the organisation must choose which decisions are made by humans and which by the algorithm. This is typically controlled using a threshold, which measures the complexity, uncertainty and/or impact of the decision. The organisation can move the threshold in response to changing conditions so it has some control without needing to change the algorithm.

Advantages:

- All the advantages of full automation for simple decisions.
- Ethical risks can be managed by using targeted human decision-making.
- Human decisions can be used to retrain the model (being mindful of possible human bias).
- Concentrates limited human resources where they are needed most.
- Service users can explain their issues to a person.
- Can evolve with social and ethical changes.

Disadvantages:

- More complicated than other approaches because it is a combination of two decision-making processes.
- The threshold needs to be set at an appropriate level.
- The volume the algorithm processes needs to be sufficiently large to justify the investment of building a partially automated decision-making system.

Key considerations when implementing partially automated decision-making

The organisation must ensure the human decision is not always the same. For example, all human decisions could be 'declines'. In this case, the algorithm is making all the decisions. Another key consideration is the influence of the algorithm on human decision-making (see *Augmented decision making*).

[12] https://www.acc.co.nz/assets/im-injured/ef79338f63/claims-approval-technical-summary.pdf

During development, fairness testing must be done across the range of realistic threshold values to ensure the organisation can move the threshold without impacting fairness. The data scientist should also show the business what impact different thresholds have on accuracy. A common risk is when operational pressures, such as limited FTEs for human decision-making, increase the threshold without the business having clear visibility of the affect it can have on accuracy.

Maintenance and monitoring should consider both the decisions made by the algorithm and those made by humans to ensure the system as a whole is working well. One option is to have an ongoing quality assurance process where a small proportion of decisions go through both the automated and human approaches. This can also make updating the model easier since it increases the size and coverage of the decisions available to retrain the algorithm.

Augmented decision-making

Augmented decision-making is when the algorithm provides information to enhance a human decision-making process. An example of this is the use of personality tests in recruitment: the test gives the person making the decision information which they can use, or they can ignore. This approach is the most complicated of the three shown here since it attempts to combine the advantages of human and algorithmic decision-making in every decision.

This approach is most applicable when every decision is complex or when every decision could have a significant impact on the service user's life. In this case, there are no simple decisions that could be automated, so a skilled human is needed for all decisions. The algorithm's role could be to help by summarising a large amount of information that a human would struggle to deal with, or by acting as a backup to the human decision to ensure no red flags are missed. It also is used where the human can gather, often directly from the client, information not available from the administrative data. As a result, augmented decision-making is a popular buzz word in medicine where patients can have very long histories and where the results of mistakes can be fatal.

Advantages:

- Combines human and algorithmic inputs for complex decision-making.
- Can increase the speed of human decision-making.
- Service users can explain their issues to a person.
- Can evolve with social and ethical changes.
- Concentrates limited human resources to the part of each decision where they are needed most.
- Can act as a backup to ensure nothing is missed in an important decision.

Disadvantages:

- Each human decision maker can use the computer-generated information differently.
- Human decision makers can become overly reliant on the computer-generated information so that it is effectively a fully automated system without any of the ethical considerations being properly covered.
- Human decision makers can ignore the computer-generated information.
- All decisions can be subject to human bias.
- Decisions are slower.
- Decisions can only be made when someone is available to make them.

Key considerations when implementing augmented decision-making

The success or failure of an augmented decision-making process comes down to understanding, controlling and monitoring how the computer-generated information influences the human decision.

Understanding is important because it drives what information is displayed to the decision maker and in what form it needs to be. Data scientists are gaining a greater understanding of the importance of how information is displayed, as can be seen with the emergence of data visualisation and storytelling. In this case, the data scientist needs to display the right information for every individual decision.

Controlling is important because it allows for greater consistency from one decision maker to the next. It can also be used to limit the reliance of the human decision maker on the computer information (eg by only showing a comparable computer decision after the human has entered their initial decision). In this part of the process, timing can be a key tool.

Finally, **monitoring** is a significant factor because the influence of the computer-generated information can change over time. This could be due to the user gradually becoming more and more reliant on it, or it could be due to new users being hired and not being trained to the same knowledge level as the original users.

In summary, augmented decision-making holds great potential to help with complex decisions, but it also comes with risks that must be effectively managed.

Consultation and co-design

Overview

Organisations benefit greatly by involving frontline staff, service providers and service users in the design of their processes and tools. This is particularly true of operational algorithms. Algorithms are based on data that gives only a limited perspective of complex real-world events and activities, so must be calibrated against the actual experiences of those involved.

Each organisation has its own consultation and co-design process. As these are constantly evolving, they won't be covered here. Instead, this section focuses on what the data scientist should learn from the consultation and co-design process:

- Understanding the data.
- Understanding the process.
- Understanding the best way to be transparent.
- Using all that information to build better models.

The key principle is that a high-quality algorithm relies on the data scientist understanding the stakeholders involved in the decisions and how their experiences are captured in the data.

Roles and responsibilities

The required roles in consultation and co-design are:

- data scientist
- the business owner (and SMEs).

The role of the data scientist

The primary tasks associated with this role are to:

- understand the data from the perspective of other stakeholders
- learn the implications for modelling
- understand the process and design the model input and output to make the best use of the process
- build the model with transparency in mind
- make optimal use of the existing knowledge and theories of stakeholders
- fill in the sign-off documentation.

The role of the business owner (and SMEs)

The primary tasks associated with this role are to:

- help the data scientist to understand the meaning and potential weaknesses of the data in the real world
- formulate theories and ideas about what factors influence the outcome
- sense-check the model results
- help the communications team understand how transparency can be achieved (see *Transparency and communications*)
- help the data scientist understand the process that fits around the model (see *Model and process design*)
- sign off the approach to consultation and co-design.

There are also roles for other external stakeholders such as service providers and service users. Their tasks will be like the first four listed for the business owner.

Main ideas

Understanding the data: it represents real people and real experiences

Working with service providers or frontline staff that input the data gives the data scientist the opportunity to understand:

- what each value in the data means
- how missing values are created and how they should be imputed
- what data to include or exclude
- what definitions best fit the target variable
- what derived variables may be useful
- what data can't be trusted
- when the data is input (some may be added after the prediction needs to be made, so can't be used)
- what data is missing.

Understanding the process: the process and the model must work together

As detailed in the *Model and process design* section, the model must fit in with the business process to form a successful algorithm. For example, the model may make one decision but the business process will decide whose application gets sent to the model for a decision, how that decision is communicated and what that decision means in terms of the services available to the service user. Frontline staff have extensive experience working with the business process and the people involved. Therefore, they are key to the data scientist understanding the process so that:

- the data used in building the model is representative of the data used in scoring the model (eg it isn't from a different group of people and hasn't been further cleaned)
- the output of the model is in the appropriate form for the broader process to use
- the scope of the decisions the algorithm makes can be limited to only include situations where it is most accurate (eg the algorithm might only make the straightforward decisions while a person makes the complex decisions)
- the algorithm may be tailored to give the business some control over the algorithm without needing to make changes (eg using thresholds).

Understanding transparency: models that mimic the real world are easier to explain

When building an algorithm, the best approach is to start with the variables that are familiar to the people involved and which their common sense or business knowledge indicates should be most useful. This familiarity helps with gaining their acceptance and trust of the algorithm, which is key to changing their behaviour (usually the hardest part of successfully implementing an algorithm, see *Change management*).

It also makes transparency easier to achieve. Working with frontline staff, service providers and service users also gives the data scientist a better understanding of:

- the audience for communications
- what the audience needs to know
- how to communicate with them effectively

- what variables are familiar to them, and what terminology to use.

Understanding context leads to better models

When setting out to build a model it's important to start with a simple, transparent version that can be understood by stakeholders. This means it can be sense-checked to ensure no errors have been made in the data processing. It’s easy to make mistakes and hard to pick them up in complex models.

Stakeholders often enjoy being able to put forward their theories about which input variables are useful and having the data scientist investigate them. Some will turn out to be accurate and useful in the end model and some will be dispelled in the process. One way to explain this process to stakeholders is that they will generally be good at identifying which variables are important and in which direction those variables influence the outcome. However, stakeholders will generally not be as good at estimating the size of the effect each variable has – that is what the data is required for.

Finally, working with stakeholders will give the data scientist a better perspective on what fairness measure should be used and why (see *Fairness and bias*).

Stakeholders get a more transparent process and a more transparent algorithm

Stakeholders should be looking to build their trust in the algorithm by driving the data scientist towards:

- a simple solution
- a solution built on real-world experience
- a transparent model
- a solution that addresses some of their concerns about the process.

Stakeholders should also try to instill in the data scientist and communications team the understanding necessary to effectively communicate with other stakeholders about what the algorithm does, how it does it and what the organisation is trying to achieve.

Stakeholders will be the people whose work and lives are affected by the algorithm. If they do not trust it, then neither the organisation nor the stakeholders involved will get the outcomes they seek.

Change management

Overview

Implementation of an algorithm involves significant technology, process and people change. The most important, and most difficult, of these is people change. Organisations will often have change managers who will be best positioned to address the change. However, the data scientist can help by designing a transparent algorithm around the understanding of frontline staff and other stakeholders.

The data scientist should be aware that people are more accepting of change if they understand what the change is and why the change is being made. This is a good reason to avoid black box models.

A key message in the change process is that the algorithm will be targeted towards low-value, high-volume decisions. This can free up staff to concentrate on the complex decisions that require their expertise.

There are also significant risks associated with people change, especially if the change involves job losses. The data scientist should be careful to leave managing those risks to the professional change manager.

This section details how the data scientist can help the change manager and stakeholders understand:

- why the change is being made
- what the change is.

The data scientist can also help stakeholders by making the model and process work well together, listening to stakeholder feedback, and refining the algorithm during the early life support stage of the implementation.

Roles and responsibilities

The required roles in change management are:

- change manager
- data scientist.

The role of the change manager

The key tasks of the change manager are to:

- understand what the algorithm is doing and why
- clearly communicate what information about the algorithm they need
- give feedback to the data scientist about any information they, or the stakeholders, are having difficulty understanding
- fill in the sign-off documentation.

The role of the data scientist

The key tasks of the data scientist are to:

- articulate why a change is being made and why an algorithm is being used
- make it as easy as possible to understand the algorithm and how it preforms
- provide input into training materials
- refine the algorithm in early life support to incorporate feedback and make it as usable as possible.

The change management approach and implementation will be signed off by the L2 business owner.

Main ideas

Why is the change being made?

To solve a business or service user problem.

In the *Idea formulation, selection and planning* section we talk about how important it is that the business plays a leading role in deciding which problem is being solved. This is also critically important in managing change because the people involved will be much more receptive if the algorithm is solving problems they know about first hand.

The opposite approach – where the analysis team comes up with a solution and tries to 'sell' it to the business – is far less successful because the people involved struggle to understand why the change is being made and therefore resist it. The 'why?' in this case is often that the organisation wants to make better use of its data assets or the data scientist involved think they can do something cool. Neither of these are good reasons to make a change and the end users will struggle to link their experience to these reasons.

Service users will also be much more receptive to change if it's solving a problem for them. For example, the algorithm may make faster decisions so that the service user can have all their issues addressed in one interaction.

The 'why?' question is critical to ethics so will need to be clearly articulated when filling in the PHRaE. When the reason for the change is something that may be unpopular, the organisation needs to be honest about it rather than searching for another reason. The most common example of this is making cost savings by replacing staff with an algorithm. In this case the organisation should simply state that is the reason – there is nothing wrong with a government agency trying to be efficient.

What is the change?

Transparency and simplicity are key

Algorithms are often inherently complicated and difficult to understand – so much so that sometimes even the data scientist that built them doesn't know how they work. However, they don't need to be. Simple algorithms can be almost as accurate as the most complicated ones and may be much more successful in the end because the people involved are more accepting of them. Similarly, people will be more accepting if they can see that a genuine effort has been made to make fairer decisions.

Transparency and clear communications are vital to making the algorithm understandable by, and accessible to, others (see *Transparency and communications*).

A few tactics that might help

1) Base the algorithm on the factors and logic that the human decision makers used – this will make it easier for users to understand the algorithm. The data scientist is simply using data to build on the user's knowledge and quantify already known relationships.
2) Involve some influential stakeholders in building the model. This will give them a chance to see how the data scientist is building the model and to build up trust of the process. They can then communicate their understanding to other stakeholders using their language.

3) Build a demo of the algorithm in Excel. This will allow interested stakeholders to see a much more accessible version than the SAS, R, etc version. They can use the demo to try different inputs and see that it reacts in the expected way.
4) Think carefully about how each prediction is communicated to frontline staff, service providers and service users.

Training and explaining how the model works

As with most large changes, frontline staff will need to be trained how to use the new decision-making system and what they are required to input. This is particularly true of the augmented decision-making approach (see *Model and process design*) where frontline staff need to make a decision based on information from the model.

The tactics above will help with making training easier, but it shouldn't be assumed that staff will find the adjustment easy. For staff who have significant interactions with the model, specialist learner and development resources should be used rather than relying on the data scientist to communicate the changes. The influential stakeholders mentioned above will be invaluable in the design of training materials.

Invest in making the process and algorithm work well together

Users will tend to be more accepting of a change that works smoothly and adapts to their feedback. The *Model and process design* section includes details about making the process and algorithm work well together. To make this a reality, user testing is very important. Even so, the desired smoothness will not always happen straight away. Sometimes only applying the combined approach in a real situation will highlight situations where it doesn't work smoothly.

The organisation should expect to refine both the process and the algorithm in the early life support phase immediately after deployment. This refining will need to be an ongoing feature of the monitoring and maintenance as the business problem will often evolve over time. Managing the transition period during change is something that should be factored in at the beginning of a project and sized according to the scale of the change.

The organisation should also expect the unexpected when it comes to the user's reactions. It may be variable and not what anyone involved expected. That's why it's important to include users in the project from the beginning, have a dress rehearsal to gauge how users will interact with the model, and then have a decent block of time allocated for making changes before go-live.

Overall, change management is likely to be an area of rapid evolution in the future as the ethical considerations, algorithms and people's perception of algorithms change.

Working with business units who provide support

Leveraging tools and processes

Overview

A significant component of the MDL framework is the standard set of processes on project methodology, design, development, testing and deployment. Each organisation will already have processes and tools in place that cover all these areas. The role of the MDL is not to replace those tools and processes but to co-ordinate how they fit together to help the implementation of operational algorithms.

This section looks at the different elements of the MDL and explains how existing tools can be used to help the process.

Roles and responsibilities

The required roles in leveraging tools and processes are:

- data scientist
- deployment manager.

The role of the data scientist

The key tasks of the data scientist are to:

- complete the design work
- manage project tasks and workflow (for example with a Kanban board)
- version artefacts (code, reports, data)
- test the solution.

The role of the deployment manager

The key tasks of the deployment manager are to:

- ensure all pre-deployment steps and quality assurance requirements are met before deploying to production
- merge code back to the trunk
- fill in the sign-off documentation.

The IT implementation and testing will be signed off by the L2 IT approver.

Main ideas

Project methodologies

There are many different project methodologies, ranging from **Waterfall** which is heavy on up-front requirements, to **Agile** which is more iterative in nature, or a hybrid model that is halfway in between.

The Agile way of working is common among IT teams and has been adopted by many data science teams. Tasks are organised using Scrum or Kanban boards, often using a tool such as Jira. The sections below relate to the Agile approach, but the general principles apply to most project management frameworks.

Making the most of Kanban

Kanban works in much the same way for operational analytics as for other applications. A key advantage is that Kanban can be used by a range of teams from a range of different areas of the business. As the sections above demonstrate, there are many different teams involved in getting an algorithm into operation, so having them all use the same tool is a significant advantage.

Below are a set of recommendations for maximising the effectiveness of Kanban. These are similar for algorithms as they are for other applications.

Have a purpose

Have a vision statement on your physical board to remind you of your goals. These can be represented in your work tracking software as an epic or feature rather than a user story. These goals should be SMART and something that most roles can understand. For example, save $10 million in the fraud space by the end of this financial year. This should be the same goal you mentioned in the PHRaE and it should be understood across the whole team.

Maintain your board

Keep your boards up to date with all the tasks relating to building your predictive model. Tools like Jira will also provide an audit trail automatically. This is very useful when revisiting a model that was last worked on a year or more ago. Some organisations install monitors or moveable screens in their space so that the digital board also becomes a physical ever-present reminder of what the focus is for the day.

Minimise work in progress

Nobody can really be focusing on two things at once. Kanban boards are designed around the principle of limiting work in progress. Situations where individuals have several tickets for extracting data, model building and other features in progress at once should be avoided, even if they are not actively working on all of them. Teams should make use of a status such as ‘blocked’ or ‘postponed’ in their workflow to highlight what’s truly going on with a task. Doing this allows other people to see what stage the project is in, even if they weren’t at a standup.

Make the most of your project tracking tool’s features

Tools such as Jira can time track, which can help you get a better idea of how much effort is involved in a task. Links also allow dependencies to be included, such as documentation, which must be complete before deployment. Using such features on bigger projects is advised especially if you’re unsure how much effort a task will need. Links to dependencies are particularly good for larger projects.

Have your board columns reflect your lifecycle

The most basic Kanban boards have three columns: To do, Doing and Done. An alternative arrangement of analytical projects is to arrange the columns to align with standard stages in a model’s development:

To do	Design/analyse	Develop	Test/review	Deploy	Done

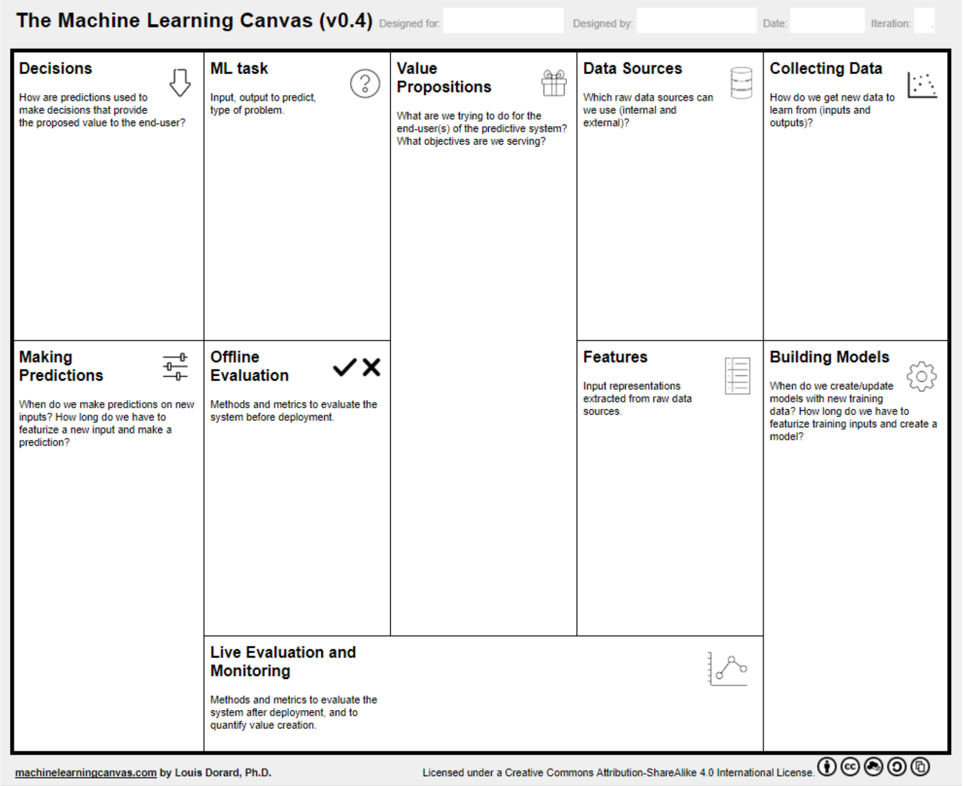
Work generally moves from left to right but given the exploratory nature of analytics it’s possible for cards to move in the other direction from time to time.

Design

Lean canvases are a good way to start your design journey. Traditional lean canvases focus on problems, solutions, key metrics, plus cost and revenue. However, for government, revenue should be thought of more generally rather than fiscally. Once the business' problem, solution and value are understood, a hypothesis canvas can be filled in with an explanation of how you are going to execute this. An example of a machine learning canvas is shown in Figure 1 These canvases can sum up a lot of information on a single A3 and get to the heart of the business problem and how you intend to solve it.

This will also help complete parts of the privacy and ethics documentation.

Figure 1: An example machine learning canvas[13].



The Machine Learning Canvas (v0.4) Designed for: Designed by: Date: Iteration:

Decisions
How are predictions used to make decisions that provide the proposed value to the end-user?

ML task
Input, output to predict, type of problem.

Value Propositions
What are we trying to do for the end-user(s) of the predictive system? What objectives are we serving?

Data Sources
Which raw data sources can we use (internal and external)?

Collecting Data
How do we get new data to learn from (inputs and outputs)?

Making Predictions
When do we make predictions on new inputs? How long do we have to featurize a new input and make a prediction?

Offline Evaluation
Methods and metrics to evaluate the system before deployment.

Features
Input representations extracted from raw data sources.

Building Models
When do we create/update models with new training data? How long do we have to featurize training inputs and create a model?

Live Evaluation and Monitoring
Methods and metrics to evaluate the system after deployment, and to quantify value creation.

machinelearningcanvas.com by Louis Dorard, Ph.D.

Licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Development

Any code that is to be deployed operationally needs to follow good coding practice that's documented well. This includes items such as building modular code that can be reused in the future to decrease development time. It is important to work in with the wider standards of your IT teams – after all, your code will be running on their systems.

In addition, good model development practices are outlined in this document.

Version control

Version control allows for the tracking of changes to files including metadata such as who made changes and when they were made. Tracking changes is particularly beneficial when working as a team or performing model maintenance where the data scientist may not have been involved in the initial development or previous changes.

Version control is particularly important for operational algorithms, which tend to have a longer life expectancy than other analytical solutions. It's also often a requirement that an

[13] Source: https://www.louisdorard.com/machine-learning-canvas

audit trail is kept for automated decisions so that it's clear which version of a model was used to make a particular decision.

Anything that changes over time should be versioned. This includes:

- code
- data
- documents
- passwords and other secrets.

Code

Code is usually versioned using Git, or in some cases Subversion. The workflows and branching methods should be documented and well understood within the team.

Documentation relating specifically to the code and how to install and run it may be kept in the repository, but project documentation and reports, correspondence, and data or results are normally stored elsewhere.

In addition, it's worth linking the commit messages to the ticket numbers for tasks. This makes it easier for the tester to know what code to look at and what code may need to be revisited for potential bugs or changes.

Documents

Project documents such as business cases, privacy assessments, technical reports and sign-off are usually stored in an enterprise filing system. This means managers and policy and privacy teams can see key documents without needing Git or Subversion.

For consistency, Word or Excel should be stored either under code management or in the corporate file system – but not in both.

Data

Most of the data will be managed by the data warehousing team. However, sometimes data may be received in spreadsheets or on iron keys for a particular purpose. These should ideally be managed by the data warehousing team so they can be appropriately versioned and subject to the appropriate security. Having them incorporated into data lineage tools will also help ensure all teams understand what a dataset has been used for.

Passwords and other secrets

These items shouldn't be stored in code repositories. If they are, anyone with access to the repository will have access to the passwords. It also means the work can't be shared publicly, limiting some elements of transparency. There will be more secrets to manage for cloud-based data science solutions that require keys, tokens and certificates to run. This problem is common with other organisations who have a plethora of secrets to manage their analytics.

There are many tools and methods available that manage passwords, credentials and secrets[14,15]. These range from password management tools such as KeePass or Dashlane, through to credential and key management, offered by Azure Key Vault or AWS Secret Manager combined with Azure Key Management Service. A team may have responsibility for managing secrets during development, but the organisation's IT security standards will likely apply for operational deployments.

[14] https://blog.envkey.com/managing-passwords-and-secrets-common-anti-patterns-2d5d2ab8e8ca

[15] https://blog.cryptomove.com/secrets-management-guide-approaches-open-source-tools-commercial-products-challenges-db560fd0584d

There are measures you can implement to try and mitigate passwords getting into SVN. To protect version control systems, make sure your code management tool ignores versioning sensitive files, however ideally passwords shouldn't be sitting in plain text files.

There are also tools to check if passwords have made their way into the version control system. Clouseau[16] allows you to pattern match for passwords, profanity and much more. While it's intended for Git, someone who's good with grep commands in Linux could create their own password checkers.

Testing

Testing can cover many different areas:

- Peer reviewing code.
- Reviewing the model (see *External review*).
- Business interpretation check.
- Fairness and bias testing (see *Fairness and bias*).
- Testing for model drift.

The testing that's specific to data science products is covered in other areas of this document.

Testing common to IT

Unit testing involves testing the individual components of a piece of code. For example, one test might check the data extraction to make sure there aren't any columns missing because of incorrect joins. Having these sorts of tests mean that next time a model project requires it, the same code can be reused.

Code to build models should be modular to make the components more reusable and easier to troubleshoot when something goes wrong. The tests could be as simple as function calls sitting in a script or it might be rolled into an automated testing framework.

Integration tests determine how the components work when they're connected. Integration testing is less of an issue in data science because the components usually have an order of execution. For example, data extract, data transformation, model building and then model assessment. Even within data transformation, the order in which operations are performed is fairly consistent, eg transformation comes before imputation.

As a result, running through all the code in a linear sequence effectively performs integration testing.

Integration testing should be carried out in a dedicated testing environment rather than the data scientist's local machine to ensure the code will still function correctly and no dependencies to areas such as a local C: drive have been introduced into the code.

Another type of testing that's very important for operational algorithms looks at the effect of errors in the scoring data. For example, a model had a flag of 'large benefit receipt' as its main driver and a particular service user had mistakenly been paid a large amount, but the payments had been reversed. If the reversal hadn't been recorded in the data, the automated decision would likely be incorrect.

[16] https://github.com/cfpb/clouseau

Having an audit trail

For typical IT testing, a test exit report shows the test's pass/fail rates as well as a list of defects carried into production. Your organisation's IT team may have a framework that could be available as a starting point for the data scientist.

With testing, it's important to have an audit trail to understand how features have been tested and whether they have bugs. This can be managed through your work tracking system (Jira) since the data scientist will already have access to it – there will likely be a specific ticket or story type for bugs.

Deployment

This section covers deployment from a technical perspective. See the *Governance Guide* and the *Model and process maintenance* section for other components of operationalising a model.

Each organisation will have its own procedures to make an algorithm operational, often mirroring the deployment standards used by IT or data warehouse teams. Some of the general principles include the following:

- Use code management (Git or Subversion) to clearly track which code is deployed.
- Stage deployment through development, test, and production environments.
- Quality assurance needs to occur, and be documented, as the application passes through each environment.
- The code released at each stage should be tagged to allow easy identification of which version of code was running at which time.
- Tags and commit messages should be meaningful and clearly identify changes from the previous version and indicate which environment the code is destined for.
- In particular, 'Merging to the trunk' at the time of deployment will ensure that the latest version of code in the code management is the same as that running operationally. Development of new features or updates would normally start as a branch from the trunk.
- Scheduling the code to run regularly will require identifying dependencies on other programs and data sources.
- Scheduled jobs need at least minimal documentation to help system administrators manage situations where the algorithm (or one of its dependencies) has failed – especially if the algorithm is part of a long chain of related processes.
- To avoid surprises, other teams with programs that interact with the algorithm should be alerted to any significant changes to be deployed, including being involved in the integrated testing.
- Logs from each run need to be complete and saved for an appropriate length of time. Error free logs from development and test environments are a prerequisite for advancing to production.
- Along with the model or algorithm itself there will likely be audit and monitoring scripts which need to match with the latest version of the algorithm.

Transparency and communications

Overview

Transparency is becoming increasingly important in the use of algorithms in government and this trend has resulted in more information being proactively released to the public. This change is partially a recognition that the Official Information Act entitles the media and public to request information about an algorithm. However, it's also due to the public's growing expectation that government and non-government organisations are more careful and transparent about how they use the data they collect.

Given these factors, it's better to think of transparency as an opportunity to reduce the risks associated with algorithms. The risks include:

- the organisation is using data it has but should not be using for this particular application
- the algorithm is biased or unfair
- the algorithm is being used for an application that would be embarrassing if the public found out
- the proper process hasn't been followed for getting sign-off, or ethical sign-off hasn't been obtained
- the organisation has tried to keep the existence of the algorithm secret.

Transparency forces the organisation to address all these issues especially when the algorithm is being implemented. This is when it's easiest to address issues because the resources are available and the decisions being made are fresh in everyone's mind.

Transparency is also an opportunity to build trust. Even for a fraud model, where the exact details of the algorithm can't be released, it should be possible to show that the right process has been followed and that ethical sign-off has been obtained. For other models, the technical documentation, the results of any reviews and a layperson's explanation should be released. It may even be possible for the code to be released via a code sharing website such as GitHub. The key to building trust is for the organisation to always be ahead of the public's need for information and to proactively release more than would be expected.

The recommended approach when developing an algorithm, is for the staff on the project to proceed as if all information about the model will be proactively released (including the code and the process followed). This limits the risk and increases the quality of the work. For example, data scientists are much more thorough about commenting their code when they think that others could be looking at it. The organisation can then decide how much to release while knowing that everything associated with the project is of a standard that could be released.

To co-ordinate the release of information, a communications plan should be part of the initial setup of the project. For important models, specialist communications resources should be used to ensure the information is targeted towards the right audience and easy to understand. Below are outlines of what transparency should look like for different types of model. These outlines start with those algorithms where the least information can be released and move towards greater degrees of transparency. All models should be able to follow one of these three approaches.

Roles and responsibilities

The required roles in transparency and communications are:

- communications specialist

- data scientist.

The role of the communications specialist

The key tasks of the communications specialist are to:

- write a layperson's explanation of the algorithm
- provide a recommendation
- fill in the sign-off documentation.

The role of the data scientist

The key tasks of the data scientist are to:

- ensure the layperson's explanation is accurate
- provide diagrams and statistics for use in the layperson's explanation.

The communications specialist and data scientist need to work together to provide a recommendation about what information will be released to the public and when. The **communications manager** will review the communications plan and material for release. However, the release of information to the public is so important it will be signed off by the **level 2 approver**.

Main ideas

Transparency for a fraud model – some information can be released

Typically for fraud models, organisations look to keep as much a secret as possible so that the people under investigation don't know how to avoid being detected by the model. However, there are some pieces of information that can be released without jeopardising the effectiveness of the model. In fact, if done correctly, the knowledge that there is a sophisticated algorithm detecting fraud can act as a deterrent.

The information that should be proactively released focuses on showing that the correct process has been followed rather than the details of the data used and the type of algorithm. It includes:

- the existence of the algorithm
- evidence that ethical sign-off has been obtained
- evidence that the proper sign-off process has been followed
- what fairness measure has been used
- evidence that affected groups have been properly consulted
- if external reviews were carried out, what those reviews covered and the high-level results.

It's also reasonable to expect that the results of the algorithm are published periodically. For example, that X potentially fraudulent transactions were detected, Y transactions were verified as fraudulent, and this resulted in Z prosecutions. Breakdowns by gender, ethnicity and age would also help demonstrate that the algorithm is fair.

Transparency for an access-to-service model – more information can be released

For a model that controls access to services or payments, it's important to balance the public's right to knowledge with the potential for 'gaming the system'. We would expect that all the information listed above for the fraud model would be released, as well as:

- information about the business/service user problem that's being addressed and the specific goal of using an algorithm (eg faster decisions, cost savings)
- information about the data that was used to build the model and what's been done to anonymise the data prior to its use
- the variables used and a general idea about how each one affects the outcome - the latter will vary from one situation to another but may include the direction of the effect or the relative importance of the different variables
- the type of algorithm used (eg decision tree, neural network)
- the tests that were conducted to ensure fairness and what the results of those tests were
- any adjustments that have been made to the algorithm/data to ensure fairness
- information about how the model and the process work together (see *Model and process design*)
- basic information about the model maintenance plan (such as the frequency of updates)
- information about the consultation and co-design process.

The level of detail will depend on the importance of the decision being made. The form of the communications will also depend on the audience. Ideally both a plain English and a technical version would be released along with frequently asked questions and contact details for someone who can address questions and feedback.

Transparency for an additional services model – even more information can be released

Additional services models are those where the organisation is going beyond its minimum requirement to solve a specific problem by offering people additional services. This could be help accessing health care for groups with high needs, or targeted education and training for those that need it.

In this case, all the information detailed above for an access-to-service model should be released, as well as:

- a clear justification for what problem is being addressed by the program and why
- if there are no concerns about gaming the system, the formula (or similar) for the model
- the variables used and details of how they affect the outcome
- the full technical report
- the code used.

The audience is key to good communications

Typically, an organisation will think of service users as the main audience for communications. This is a good starting point and helps ensure the communications are not too technical. However, the first audience is in fact the staff who need to sign off the algorithm. They need to understand it well enough to know if the risks have been managed appropriately and what questions to ask.

Level 2 approvers are a special case since they have final sign-off. They will need to answer questions from other senior managers and from the Minister. Having good laypersons explanations immediately on hand will be invaluable.

Another internal use for good, simple communications is change management (see *Change management*). Frontline staff and service providers are much more likely to carry out change if they understand the change and why it's happening.

Finally, communications need to be kept up to date. An algorithm could be in use for 10 years or more and the environment could be subject to substantial change during this time. The plan maintenance should include looking at the communications at least every 5 years.

Non-transparent and self-updating models require special mention

For some modelling techniques there is the added complication that they are a black box, even to the data scientist who built them. This makes some components of transparency difficult to achieve. However, most of what is outlined above is not dependent on understanding the model itself and some insights can be gained by using surrogate models – simple, transparent models that replicate the predictions of the model in most cases.

Self-updating models also present some issues because the model is constantly changing. In this case, it's impossible to say exactly how the model works. As for non-transparent models, it is possible to achieve a reasonably high degree of transparency.

Principles and frameworks to manage potential risks and harms

Privacy and ethics

Overview

Privacy and ethics are very important subjects. They make sure data is used appropriately and help surface potential risks with a given activity.

The approach each organization takes to manage these issues will vary from an ad hoc series of decisions to an integrated formal process that is centrally managed. An example of the latter is MSD's existing framework called the PHRaE that is used to answer privacy, human rights and ethics questions. The framework creates a consistent and transparent approach to working through these matters. The acronym PHRaE is used as a generic term to describe an organisation's framework for managing these issues. This section is designed to complement an organisation's PHRaE by specifically looking at:

- making the most of PHRaE
- what outside the PHRaE needs to be considered
- what other existing tools and guidance can be used.

More information about MSD's PHRaE can be found on the MSD website[17].

Roles and responsibilities

The required roles for privacy and ethics include:

- data scientist
- business owner
- privacy and ethics specialist
- analytics owner.

The role of the data scientist

In addition to the development of the model, the data scientist has responsibilities in the privacy and ethics section. The key tasks associated with this role are to:

- complete the PHRaE
- implement the identified changes
- fill in the sign-off documentation.

The role of the business owner

The role of the business owner is to provide the frontline perspective. The key tasks involved are to:

- confirm the goal of the business

[17] https://www.msd.govt.nz/documents/about-msd-and-our-work/work-programmes/initiatives/phrae/phrae-on-a-page.pdf

- provide some input for the service users (for larger projects, discussing this with the actual service providers and service users would be better)
- provide input into the decisions regarding ethical trade-offs.

The role of the privacy and ethics specialist

The role of the privacy and ethics specialist is to ensure the use of the data is consistent with the organisation's policies and to review and provide feedback on the PHRaE. They may also recommend additional content outside of the PHRaE. The key tasks of this role are to:

- ensure the data collection, security and consent policies are clearly understood and followed
- provide a critical perspective on the use of data for the overall business goal
- co-ordinate the completion of the PHRaE checklists with other team members
- give feedback on the PHRaE
- determine whether additional tools such as the Data Futures Partnership Guidelines for Trusted Data Use, Data Protection and Use Policy (DPUP), Five Safes Framework, Ngā Tikanga Paihere and Māori Data Audit Tool need to be completed
- ensure the PHRaE is signed off in a timely manner.

The role of the analytics owner

The role of the analytics owner is to formally accept the content in the PHRaE. The key tasks of this role are to:

- sign-off the PHRaE (in larger, more sensitive, projects the Level 3 owner may have this role).

Main ideas

To make the most of PHRaE

Know why you're building a model in the first place. Feedback from teams is that this is the part they find the most difficult. Having a clear understanding of why you're doing work and what value it will have is important for making quality decisions in the modelling.

The PHRaE should be populated for projects to build operational algorithms.

The PHRaE is something that should be started as soon as possible and revisited regularly, including during the model revision phase. Starting early can identify potential changes that you need to make to your design early on. A good approach is for data scientists to complete a first draft of the PHRaE as soon as possible and seek initial feedback from the privacy and ethics specialist.

It's designed to be a thorough tool, so it's **best to work on it as a team**. The great thing about using people from different areas of the organisation is that they come with different perspectives that may improve your solution.

See if you can find examples of previous PHRaEs to get an idea of what sort of content they contain. It may mean that you get your PHRaE completed faster.

Share your PHRaE widely. It allows others to learn about the PHRaE and ensures the right eyes have assessed your particular project.

There are factors outside the PHRaE to consider

There are several areas where other frameworks can be used to complement the PHRaE. You may wish to use these other frameworks in addition to the PHRaE when:

- you need to quickly write about privacy, human rights and ethics to make them understandable to service users (the Guidelines for Trusted Data Use is good for this)
- the project involves a major redesign of a high-profile service - DPUP in addition to PHRaE can help to clearly articulate the purpose, creating a shared value and transparency
- the piece of work has major impact on Māori service users - the Māori Data Audit Tool and Ngā Tikanga Paihere are more appropriate here as those frameworks cover te ao Māori perspective more so than the PHRaE.

The privacy and ethics specialist can advise whether additional tools should be completed and who to ask for help to complete them. In addition to the PHRaE, you may also need to gain ethics approval from other organisations. This situation is most likely when you are partnering with another organisation or you're obtaining data from another organisation. Parts of the PHRaE allude to this but it's worth knowing this up front so you can incorporate items such as Approved Information Sharing Agreements (AISAs) and Ethics Panel Approval into your project timelines.

While the PHRaE asks a range of questions, it doesn't deep dive into some areas especially when it comes to putting a model into operation. Even with few risks identified, you may not achieve your business outcomes because of how the model is implemented. More details are covered in the *Change management* and *Testing* sections.

For projects with serious ramifications, public consultation can be used to gauge public feedback. What may seem like best intentions can be viewed very differently to the recipient. If the consultation reveals major disagreements or risks and the team wishes to seek external advice, the Data Ethics Advisory Committee[18] may be able to help.

Additional tools that might be useful in this space

Data Protection and Use Policy (DPUP)

The DPUP[19] has four guidelines that are underpinned by principles. The guidelines are purpose matters, transparency and choice, access to information, and sharing value.

While this policy shares some common themes with PHRaE (and is also as comprehensive as PHRaE), it's written with a completely different audience in mind. PHRaE is comprehensive and suited to large organisations with established privacy teams and data science teams. DPUP takes a more holistic view and is geared towards relationships with all stakeholders, including frontline staff and service users.

DPUP is a large undertaking that should be used for projects that involve a service delivery component. However, most of the work is best carried out by those in the operational space, including service design specialists and communication advisors, who can clearly articulate the purpose and provide information in a transparent manner. Example considerations are provided on the DPUP website. The Social Wellbeing Agency is currently working on a toolkit to provide further educational resources.

The data scientist should be consulted and informed of developments in this space as it could raise questions relating to the work they're doing, such as:

- if particular records should be excluded
- what variables should or should not be used

[18] https://www.data.govt.nz/leadership/advisory-governance/data-ethics-advisory-group/

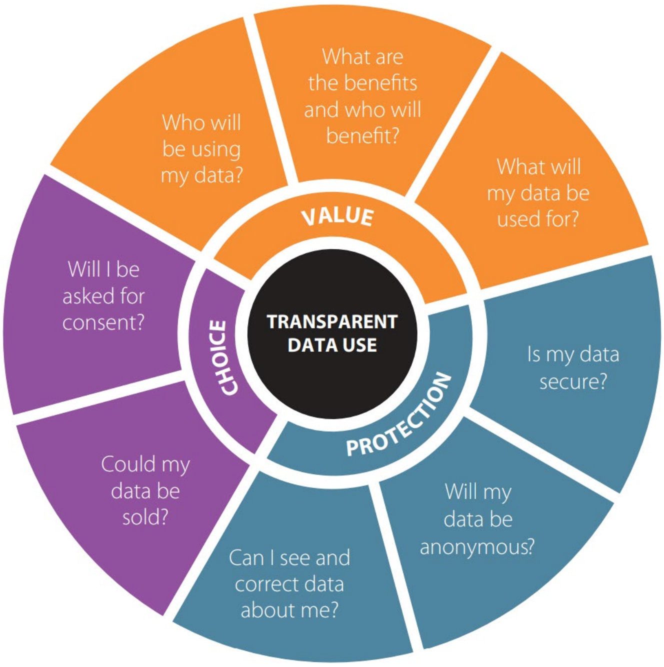
[19] https://dpup.swa.govt.nz/

- who we should engage with throughout this lifecycle to get feedback on the approach and findings
- how the outputs should be produced so that they can be understood by frontline staff and clients using services.

Guidelines for Trusted Data Use

The PHRaE asks a wide range of questions. In situations, such as communicating to the public, a simpler tool might be less overwhelming. Guidelines[20] by the Data Futures Partnership include 8 questions that are asked from an individual's point of view. These are questions asked in some shape or form in the PHRaE.

Figure 2: What New Zealanders want to know about data use. Data Futures Partnership.


TRANSPARENT DATA USE
VALUE
CHOICE
PROTECTION
Who will be using my data?
What are the benefits and who will benefit?
What will my data be used for?
Will I be asked for consent?
Is my data secure?
Could my data be sold?
Will my data be anonymous?
Can I see and correct data about me?

The Five Safes Framework and Ngā Tikanga Paihere

For working with Māori data, Ngā Tikanga Paihere[21] can be used to guide culturally appropriate use of data. This framework was designed by members of Te Mana Raraunga and Stats NZ. The PHRaE will ask about how the system will impact Māori and whether you have a plan for consulting with your cultural advisors. Using a tool such as this will help with your consultation and to ensure you are using data appropriately.

Stats NZ have used this model to provide te ao Māori perspective on all Information Data Infrastructure applications since 2017. More information about how Stats NZ have used this, including follow-ups they have carried out, can be found on their website.

Recently, the National Ethics Advisory Committee has integrated this model into their ethical standards, including minimum expectations around Māori involvement in a project.

[20] https://www.aisp.upenn.edu/wp-content/uploads/2019/08/Trusted-Data-Use_2017.pdf

[21] https://data.govt.nz/toolkit/data-ethics/nga-tikanga-paihere/

Māori Data Audit Tool

This tool[22] was produced by Te Mana Raraunga to assess how readily a project addresses te ao Māori principles described in their charter. Areas it explicitly covers, that the Five Safes and Ngā Tikanga Paihere do not, are governance data sharing agreements and the Treaty principles.

Given that it contains a wide breadth of content, it would be filled out by multiple parties. The data scientist can add information around the data and how they plan to use it. The technical lead/analytics manager would be involved in the questions around information sharing while the cultural capability advisor can help with filling out the other sections.

It is worth noting that this is an area many organisations struggle with. Guidance is evolving but often the starting point is to talk with your Cultural team or Technical Advisory Group (which should have representation from someone who has expertise on the application of te ao Māori principles).

Risk identification, classification and management

Overview

This section is repeated in the *Governance Guide* as it is necessary for all audiences and team members.

Risk management **does not** eliminate risk. Risks are never closed[23] – they exist as long as we undertake an activity to achieve an outcome. We can, however, influence the level of risk in terms of its potential impact or likelihood. The purpose of risk management is to enable informed decisions. It allows us to consciously choose whether we accept a given level of risk or act to reduce that risk.

To ensure that risk management is consistently understood and applied across MDL products, a common approach is needed. This section provides a simple approach to identifying, classifying and managing risk for MDL products.

This doesn't replace risk management policies in your organisation. Instead, it supplements current best practice with a primary focus on the risks associated with operational algorithm projects.

Roles and responsibilities

Everyone working on operational algorithms has a responsibility to ensure risks are identified and managed. While the ownership of risk resides with accountable individuals, all team members involved have a role to play.

For data scientists, the *Data Science Guide for Operations* is a practical guide to avoid risks associated with technical error. These risks need to be identified so they can be managed and owned.

Beyond technical error, there are other risks that can't be avoided that need to be identified and categorised so they can be managed and owned.

[22] https://www.temanararaunga.maori.nz/nga-rauemi#MaoriDataAuditTool

[23] Issues are risks that have already occurred. Issues will usually be dealt with as a project progresses and once resolved can be closed. Any issues that remain outstanding can be thought of as a type of risk. In those cases, the mitigations take the form of planned remedial actions.

The role of the data scientist

The key tasks associated with this role are to:

- use this guide as a first barrier to avoid risk
- ensure identified risks are classified, recorded and escalated appropriately based on the governance framework for the product.

The role of the analytics owner

The key tasks associated with this role are to:

- ensure appropriate governance is set up for operational algorithm products
- actively work to identify and manage risks, including identifying owners for the controls mitigating or reducing any impacts
- ensure identified risks are classified, recorded and escalated appropriately based on the governance framework for the product
- clearly describe the residual risk for any risks that have not been fully mitigated.

The role of Technical Advisory Group

The Technical Advisory Group has a key role to play in ensuring risks and potential harms are identified and appropriately managed.

The group is also responsible for identifying when stakeholder engagement is required, and ensuring it is used in a way that appropriately identifies and manages risks and harms.

Main ideas

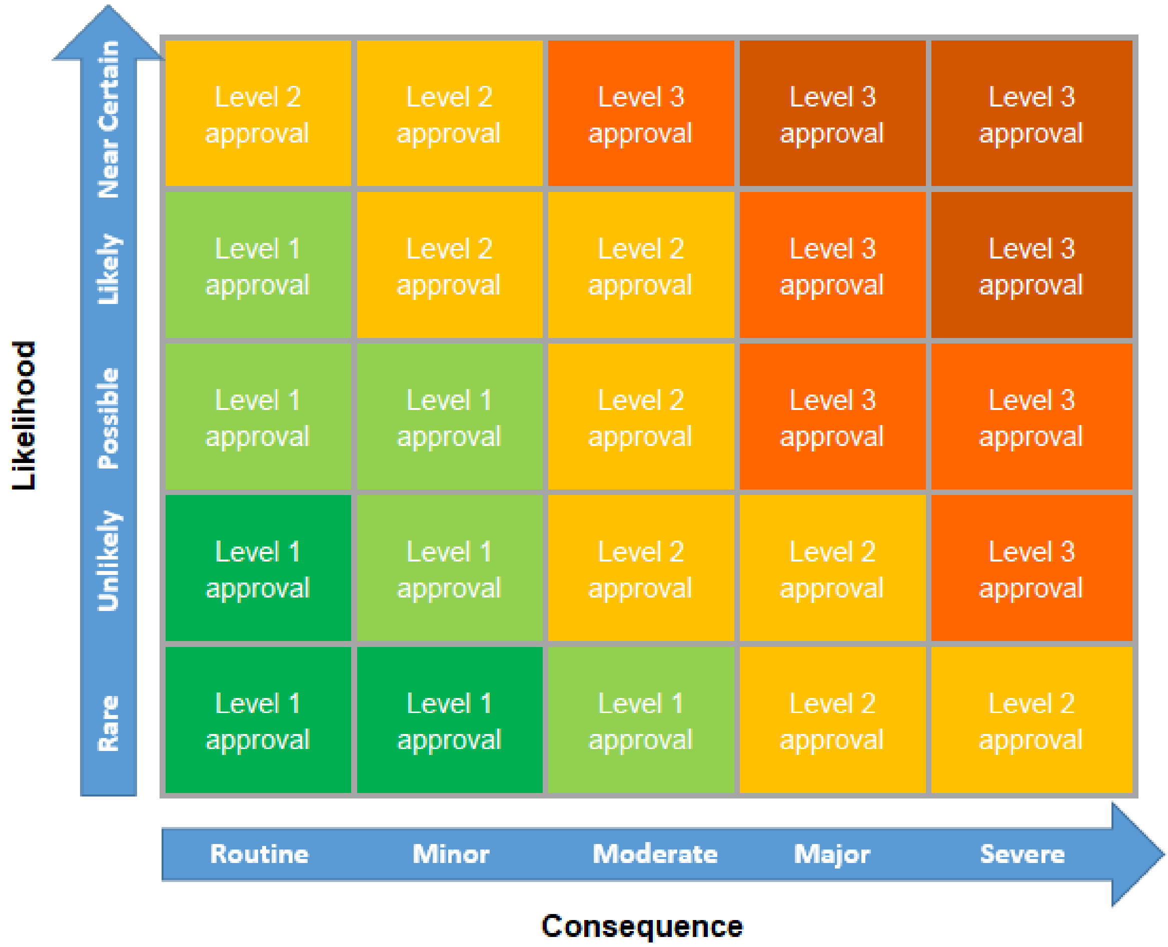
Risks must be categorised

When risks are identified, they must be recorded in a risk register (see the *Governance Guide* for the Technical Advisory Group Terms of Reference) and classified on identification.

Figure 3 provides a risk classification matrix. This is a self-assessment tool.

Impact is the severity of the impact should the risk occur. **Likelihood of occurring** is how likely the risk is to occur.

Figure 3: Risk classification matrix.



Likelihood	Routine	Minor	Moderate	Major	Severe
Near Certain	Level 2 approval	Level 2 approval	Level 3 approval	Level 3 approval	Level 3 approval
Likely	Level 1 approval	Level 2 approval	Level 2 approval	Level 3 approval	Level 3 approval
Possible	Level 1 approval	Level 1 approval	Level 2 approval	Level 3 approval	Level 3 approval
Unlikely	Level 1 approval	Level 1 approval	Level 2 approval	Level 2 approval	Level 3 approval
Rare	Level 1 approval	Level 1 approval	Level 1 approval	Level 2 approval	Level 2 approval

Risk impact

Examples of low risks could include:

- Working with highly skewed or missing data that results in unequal sample sizes and variety. There are technical risks associated with using such data, so it needs to be pre-processed before the application of analytical models – at the same time ensuring the pre-processing doesn't affect the business context of the data.
- Low sample sizes or small effect sizes, or other issues that affect the power of the statistical analysis. These are quite common risks associated with analytical models and require a power analysis to determine what can be realistically achieved from the available data.
- Model tuning could be resource-intensive and take a huge amount of processing time to arrive at a suitable level of calibration. This is dependent on the complexity of the model itself, and the number of parameters to be tuned. Each additional parameter increases time or resources needed for tuning, and sub-par tuning increases the risk of unpredictable model performance.
- The business loses the staff skills required to make the decisions the algorithm now makes. This will expose the organisation to risks should the algorithm need to be turned off (eg due to a legal decision). These risks may not have been considered when the algorithm was developed and subsequently freed up frontline staff who had experience making those decisions to work in other areas.

Examples of medium risks could include:

- Having to make compromises on model accuracy to ensure interpretability (or vice versa). There are chances that an easily interpretable model can have lower accuracy than a more sophisticated, black-box analytical model. These are often contradictory goals that could impact model outcomes.
- Model overfitting – repeated testing and tuning of models – can easily propagate unconscious biases regarding data, especially if the same datasets get re-used. The impact might be unexpectedly low model performance when the model is used for scoring real-world data and may require extensive retraining.
- Noisy data or influential outliers that violate key technical assumptions in a statistical model leading to unexpected model behaviour or poor performance in model metrics.
- The business uses the predictions of the algorithm in a new way that was not considered when the PHRaE was completed. This could result in unethical uses and uses of the data for which the organisation does not have appropriate permission for.

Examples of high risks could include:

- Developing analytical models that were trained on examples not entirely reflective of real-world scenarios in terms of business context or data quality. This could lead to unexpected poor model performance and require significant re-engineering efforts.
- The use of data for purposes where consent or social licence isn’t clear. These could be concerns relating to privacy or ownership of the data, or the use of the data in a context that was not agreed to by its providers. The impact might be that the results from this project may have to be completely redacted, even if the study was well-designed and the results were insightful. There might also be significant legal and reputational impacts for the organisation.
- Use of a variable with a large predictive power, but with inappropriate or questionable meaning in an analytical model. For instance, the use of ethnicity or gender identity might be particularly concerning in certain business contexts but may have predictive power because it acts as a proxy for certain unobservable characteristics. This is quite a common occurrence but will have a significant social impact by propagating biases and a reputational impact for the organisation.
- That frontline staff become overly reliant on the predictions of the model. This is relevant to augmented decision-making and would mean that the model and human decision-making are no longer working as designed. This could have a significant impact on the quality of decisions as well as on the ethics of those decisions, which may rely on human decision-making as a safeguard in certain situations.

Risk likelihood

Table 9 provides guidance on selecting the appropriate likelihood of occurrence level to categorise risk.

Table 9: Risk likelihood of occurring guide.

Likelihood of occurring level	Under what circumstances could the risk occur	When is the risk expected to occur	What controls are in place to prevent the risk occurring	Has the risk occurred before
High	Probable: likely to occur often during standard operations.	The risk is expected to occur within the next 6 – 12 months.	No effective controls or weak controls, eg limited business controls, with no audits performed.	Has happened in the past and no compensating controls have been implemented.
Medium	Occasional: Likely to occur sometime during standard operations.	The risk is expected to occur with the next 1 to 3 years, with a 20%-50% expectation that the risk will occur during the next 12 months.	Minimal controls, eg some business controls, with some audits performed.	The risk has occurred in other similar organisations with similar levels of controls in place.
Low	Improbable: Unlikely to occur or is only expected to occur in exceptional circumstances, such as deliberate fraud or activity beyond control of business actions.	The risk is not expected to occur within the next 5 years, and there is a less than 20% expectation that the risk will occur during the next 12 months.	Effective controls, eg timely business controls, with internal & external audits performed	The risk hasn't occurred in the business or in other similar organisations.

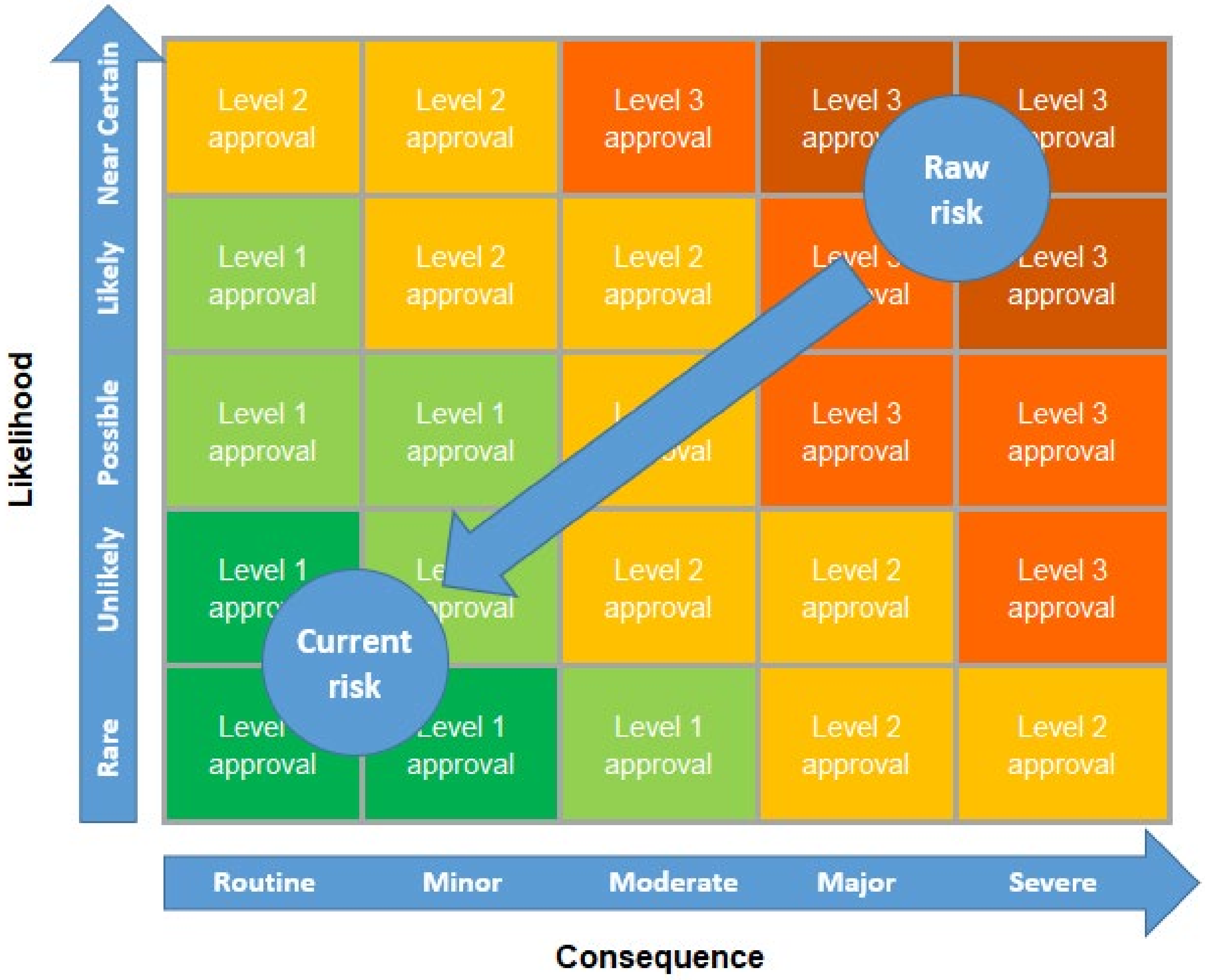
Risks must be owned and treated

Once risks have been identified and categorised, the controls to mitigate them need to be owned. Ownership depends on the category of risk and is assigned based on current risk.

Raw risks are risks that have not yet been treated. Current risks are risks that have been treated. The risk register should capture the category of the risk before treatment/mitigation (raw), and after (current).

In the event the risk moves category after it has been assigned to a control owner, it may be reassigned to a new control owner to reflect the new level of risk.

Figure 4: Risk classification matrix impact of treatment.



Likelihood	Routine	Minor	Moderate	Major	Severe
Near Certain	Level 2 approval	Level 2 approval	Level 3 approval	Level 3 approval	Level 3 approval
Likely	Level 1 approval	Level 2 approval	Level 2 approval	Level 3 approval	Level 3 approval
Possible	Level 1 approval	Level 1 approval	Level 2 approval	Level 3 approval	Level 3 approval
Unlikely	Level 1 approval	Level 1 approval	Level 2 approval	Level 2 approval	Level 3 approval
Rare	Level 1 approval	Level 1 approval	Level 1 approval	Level 2 approval	Level 2 approval

Mitigations and controls need owners too

The responsibility of the person owning a risk is to be confident that the treatments applied to mitigate that risk are effective and will continue to be so. Most of the time that person will not be the one who manages and implements the control itself.

Identifying control owners as part of the risk register provides clarity for the person signing off on a risk and is a way for the control owner to confirm the control is an appropriate fit for the risk it is applied to.

Owners of controls should be able to demonstrate that the control is effective, and that thought has been given to the maintenance and monitoring of the control.

One control sometimes mitigates more than one risk. For example, IT security around firewalls speaks to both the robustness of the delivery system and is an important means of protecting privacy. It can be helpful to build a library of well understood controls and the types of risk they commonly apply to.