Veritas Document 2

FEAT Fairness Principles Assessment Case Studies

Veritas Consortium
December 2020

Contents

01 Introduction

This document contains case study examples of the FEAT Fairness Assessment Methodology applied to customer marketing and credit scoring AIDA systems. It is the second of two documents presenting the FEAT Fairness Assessment Methodology: Document 1 contains the Methodology itself, along with general guidance for conducting assessments and integrating the Methodology with an FSI's existing risk management processes.

The FEAT Fairness Assessment Methodology is generic and intended to apply across different AIDA use cases. However, different use cases of AIDA will have unique fairness considerations in addition to those that apply across AIDA systems. This document focuses on two use cases: customer marking and credit scoring, presenting illustrative examples of assessments of these systems, and specific considerations to assist FSIs in conducting assessments of them. Accompanying these realistic but hypothetical case studies are reflections from two FSIs (HSBC for customer marketing and UOB for credit scoring) that applied the Methodology to their real systems.

The case studies illustrate applications of the Methodology for AIDA systems with different levels of risk. The customer marketing case study in Section 2.3 is an example of a detailed assessment suitable for a higher risk AIDA system, while the shorter case study in Section 2.4 is intended to illustrate application to a lower risk system. For the credit scoring case study, Section 3.3 illustrates a higher risk application of the Methodology. Answers to Part A of these case studies serve as examples of possible "triage" assessments as part of a risk management process (See Document 1 Sections 2 and 3.3).

While illustrating the application of the Methodology at different levels of risk, these case studies make no claim that the systems presented do or do not align with the FEAT Fairness Principles: this is a value judgement to be made by an AIDA System Assessor (See Document 1 Section 2) based on the answers to the assessment questions.

02 Customer Marketing

2.1 Introduction

Marketing is the business of crafting, promoting and selling products (goods or services) [7]. In a market economy, marketing plays an important role in enabling and facilitating the relationship between producers and consumers. Consumers can benefit from information that leads them to buy goods and services that meet their demands, and producers can increase profits by suitably designing and promoting their products.

Marketing is not ethically neutral and can be harmful. For instance, some products are known to have a significant potential for causing harm and yet their production and promotion is legalised in various contexts (such as with tobacco, alcoholic beverages and fast food). Also, even if a product is not inherently harmful, its promotion can be ethically questionable: much of modern advertising relies on deliberate nudging of people's wants and desires [12]. Finally, even if a product is not intrinsically harmful or its promotion manipulative, it may still lead to harm (for instance, a customer may be harmed as a result of defaulting on a loan). This suggests it is important to proactively account for ethical considerations when assessing and designing marketing systems, so there is a higher chance of detecting and mitigating ethical risks.

Marketing involves four key components: product design, pricing, how the product is distributed and made available, and how its existence and relevance is communicated to consumers. These are often referred to as product, price, place and promotion — the "four P's of marketing" [24]. The analysis of marketing in this document focuses on the promotion aspect of marketing, and in particular to whom the promotion is made or allocated (as opposed to its content or messaging).

An important feature demarcating different types of marketing promotions is whether it is possible to control which individuals receive the communication. For instance, advertisements in TV, billboards, radio or printed newspapers aren't capable of selecting which exact individuals are targeted; on the other hand, emails, direct mails, SMS and phone calls can be targeted at the individual level. This last type of marketing is called direct marketing, and it requires access to a personal channel of communication with the consumer (which happens to be the case if the consumer is a registered customer). This document largely focuses on direct marketing systems as those in which AIDA plays the most substantial role, however, many of the considerations presented will also apply to other types of AIDA-driven marketing.

Within the scope of AIDA direct marketing are a variety of systems with different objectives and possible interventions, including

Content Generation

Generating personalised content for customer based on "like-me" profiles and showing personalised offers relevant to the customer

Web and App Personalisation

Leveraging historical and real-time data to customise the web and app pages displayed to customers at different purchase stages as per their interests

Targeted offers

Determining the likelihood of a customer purchasing a product conditioned on receiving a marketing offer such as a discount

Omnichannel Assistance

Supporting customers in their purchase seamlessly across channels.

The next section draws from the variety of possible AIDA direct marketing systems to highlight a set of common properties that define the scope of the presented guidance.

2.1.1 AIDA direct marketing systems

AIDA direct marketing systems typically use one or more machine learning models to predict customer response to a marketing intervention, then target customers according to their predicted response in order to achieve a specific objective. Such systems have the following components:

A population from which selection occurs

This may be existing customers or potential leads, but a data set of these individuals exists and contains tabular information about the individuals that is used by the system for targeting interventions.

A well-defined set of marketing interventions

The system may send an email to selected individuals, or have a human operator call them, or apply a discount to an existing product, or select between multiple different interventions depending on their anticipated impact.

A mathematically precise objective for selecting individuals

The objective may be to select individuals that are likely to purchase the product being marketed (propensity modelling), or to select individuals who would purchase the product if and only if they were selected (uplift modelling), or to maximise profit using either of these formulations, or to have the customer perform some other action like log in to a website.

An algorithmic implementation

This may be automated business rules and/or supervised machine learning algorithms that use data features to select individuals based on the stated objective.

The ability to match individual marketing and outcome (e.g. sales) records

This allows FSIs to trace whether a marketing intervention was followed by an outcome (e.g. the acquisition of a product or service). This link provides a key piece of information that informs the calculation of a variety of relevant performance metrics about the direct marketing system.

Even with these properties, the boundaries of an AIDA direct marketing system may be imprecise or ambiguous. A model might, for example, use as a feature the predictions from another model. Ideally, this secondary model and the data that informs it would also be considered as part of the AIDA system under study, but risk-based judgements on the part of the AIDA System Owner should be used to manage the scope. Generally speaking, AIDA direct marketing systems are trained for a single advertising "campaign", which has a well-defined objective, time-scale, set of products or services, and interventions. This creates a natural boundary for scoping an assessment.

However, some AIDA models for direct marketing may persist over multiple campaigns or products. A general rule that acknowledges the context-dependent nature of fairness and harms is to analyse a system for a particular purpose: if the purpose changes a new assessment is conducted to capture the new context, even if the underlying model is the same.

2.1.2 Resources for assessing AIDA customer marketing systems

Building on the considerations presented in the FEAT Fairness Assessment Methodology (Document 1 Section 3), the next section provides additional considerations specific to customer marketing use cases (with a focus on AIDA direct marketing systems).

Following the considerations, the document presents two case studies of FEAT fairness assessments conducted on hypothetical direct marketing. These case studies are designed to illustrate the application of the Methodology and provide practitioners conducting assessments with concrete examples. The first case study (Section 2.3) is intended to illustrate the assessment of a higher risk system, analysed at a high level of detail. The second case study (Section 2.4) is a lower risk system, analysed more succinctly.

Finally, Section 2.5 presents reflections from HSBC, an FSI that applied the Methodology of one of their real marketing systems in production. The aim of these reflections is to help practitioners identify some of the practical challenges FSIs may face conducting assessments, and suggest approaches to overcome them.

2.2 Methodology considerations for customer marketing

2.2.1 Part A: describe system objectives and context

System objectives

Understanding the design of the system provides helpful context for assessing the other information obtained in an assessment, as subtle design choices may affect the system's operation in unexpected ways. These reflections may be helpful in creating your responses:

01 Who are the cohort of customers or leads from which the system selects individuals for a marketing intervention?

02 What are the natures of the marketing interventions and how are they assigned?

03 If the AIDA system uses a score or rank to select customers, is it using a propensity or uplift modelling-based approach, and what does that approach imply about the system's objectives?

04 Is the intervention the same for everyone, or are there a set of interventions that the system selects between, or is each intervention itself personalised to the customer?

05 Does the system apply eligibility rules for products or services being marketed, or is eligibility estimated as part of the selection process?

06 Does the system have business rules or event triggers in addition to a predictive model or set of models, and if so, how is the precedence of these systems determined?

Harms and benefits

To understand whether a marketing system's decisions systematically disadvantage individuals or groups (FEAT Fairness Principle F1) it is necessary to understand the potential harms and benefits the system causes. Fairness in the operation of the system can then be assessed by examining how the system distributes these harms and benefits. Similarly, one way to justify the inclusion of personal attributes in modelling (FEAT Fairness Principle F2), is to show that such inclusion creates more benefits, fewer harms, or distributes these more equally. Note that benefits and harms are often defined relative to a baseline: FSIs may choose to define the receiving of a special discount as a benefit, or equivalently, they may define failing to receive the benefit as a harm.

Every real system will have unique harms and benefits. These will depend on the marketing intervention, the audience, the product or service being marketed, and the timing, location and context of the campaign amongst many other factors. Undertaking careful

consultation with customers and impacted individuals and groups will help AIDA System Owners understand the potential harms and benefits of their particular system.

Direct marketing systems do, however, have a common structure and mechanism suggesting harms and benefits that may be typical in a financial services context. This section presents some possible harms and benefits as a starting point for analysis. **These harms and benefits may or may not be relevant to any given AIDA marketing system**. Similarly, some harms or benefits may be present in a system, but their magnitude may be deemed insufficient by the FSI to be used as part of an analysis of systematic disadvantage. AIDA System Owners should carefully analyse their system, and with reference to their fairness objectives and risk appetite, determine the relevant harms and benefits. The relevant harms and benefits may or may not include some of the examples listed below, and are likely to include additional items not listed here. For illustration purposes, the example list given below is used to motivate the case study in Section 2.3.

Direct marketing systems, almost by definition, selectively apply an intervention to some customers, usually to increase the chance of them acquiring a product or service in the future. This intervention itself may have an impact on the customer, separate from the action the intervention is trying to get the customer to take. Examples of such "direct" harms and benefits of a marketing system include:

Benefit from receiving the intervention

Some interventions may provide a direct benefit irrespective of whether the customer subsequently acquires the product. An intervention that, for example, provides a voucher for a free meal is valuable in and of itself. Similarly, an intervention that has a customer service member call the customer and provide a one-on-one consultation of their credit needs provides a benefit that customers not selected may have to wait in a phone queue to receive. Even an intervention that provides a discount to the product or service being marketed can still be thought of as providing something of value to the customer, irrespective of their subsequent actions.

Harm from receiving an unwanted intervention

Some interventions may be annoying to people who are not interested in the product or service being offered. Examples include emails or phone calls that appear to be unsolicited and distract or clutter inboxes.

In addition to direct impacts like these there are likely to be additional harms and benefits that also involve the product or service being marketed: if the marketing system causes a person to acquire a product, then some of the responsibility for the product's impacts lies with the marketing function of the FSI. The inclusion of such "indirect" harms and benefits is necessary owing to the cause and effect relationship between receiving a marketing intervention and receiving the harm or benefit. Some individuals will acquire the product if and only if they receive the marketing intervention. In this case, the marketing system is a partial cause of the harm or benefit they receive from the product and therefore the designers of the marketing system share responsibility. Examples of indirect harms and benefits include:

Benefit from acquiring product or service

The product being marketed presumably provides a direct benefit to the consumer, such as a loan that allows them to buy a house or car, or a new credit card that provides easy online shopping.

Harm from a failed application

Some products or services have application processes, and in some cases, submitting an application comes with a cost such as time and effort that may be wasted if the outcome of that application is negative. This is relevant if the marketing intervention attempted to cause the customer to apply. In some cases a failed application for a credit product can be recorded and potentially even make it more difficult in the future to succeed in credit applications.

Harm or benefit from a longer term outcome

Some products or services have longer term outcomes relevant to their impact. Examples include loans, for which it is possible both to default or to repay the loan successfully.

There are likely other harms and benefits that a direct marketing system may cause that are not listed here. To fully realise an analysis of the harms and benefits of a direct marketing system, conducting careful consultation with customers and domain experts is likely to be required.

2.2.2 Part B: examine data and models for unintentional bias

Performance measures

The choice of performance measures entails significant consequences for the operation of an AIDA system, which may not be apparent before deployment. For direct marketing systems, many performance measures are computed with reference to a baseline system. Often, this system is "no marketing system"; a case where customers or leads may still acquire the product or services through their own volition or due to pre-existing marketing. Empirical lift, for example, measures the increase in acquisitions of the product or service compared to a control group. Similarly, profit is also often computed by subtracting the expected profit from a control group. If no control group is available, then observational causal inference techniques must be used to estimate these measures. For an introduction to these approaches, see [26].

Which of the many quantitative measures in the literature is appropriate will depend on the particular system and the AIDA System Owner's objectives. For typical direct marketing systems, the underlying model used for targeting is likely a binary or categorical classifier, estimating either the likelihood to acquire the product or service on offer, or else that likelihood conditioned on receiving or not receiving a treatment. In such a case, typical measures of performance would include:

- empirical lift
- class balanced accuracy
- log or cross entropy loss, or some other proper scoring rule
- area under the curve (AUC)
- confusion matrices and related measures:
  - precision
  - recall
  - specificity
  - negative predictive value

For a review of these and other standard performance measures, see [19, 16].

2.2.3 Part C: measure disadvantage

Quantifying harms and benefits

Relevant harms and benefits, and how they are quantified, will depend on the precise system under assessment, the AIDA System Owner, and the surrounding context. The following section suggests one approach for a particular (example) system, whose harms and benefits are introduced in Part A above.

Example: incidence rates of typical harms and benefits

Consider a direct marketing system selecting individuals for a single type of intervention, with the aim of increasing sales of a loan product. Presume the loan product has an application process, that the customers have not yet applied for the product, and that the future outcome of a customer application is not known to the system.

Let S be a random variable representing the predictions of the targets from the automated decision system under investigation, taking values $S \in \{ 0,1 \}$. For this example assume the predictions are equivalent to the final selections of people made for a marketing intervention. Let A be a random variable denoting the personal attribute that is going to be used to analyse the system's fairness with respect to. This takes values $A \in \{ 0,...,K \}$.

After selection, there is a sequence of possible stages of outcome that may be relevant to a direct marketing system, in this example, where that system is marketing a loan product. To simplify notation the harm and benefit analysis utilises shortened references to possible outcomes (all of which may take values in $\{0, 1\}$):

Outcome	Shorthand
Applied for loan	App
Acquired loan	Acq
Resolved loan (did not default)	Rvd

For direct harms and benefits that are independent of the product or service being marketed the incidence rates of these harms and benefits can be defined as the *probability*, P, they will occur as in the case of the typical fairness measures in the literature [2]. These incidence rates can be interpreted directly as being (implicitly) compared to "no system" in which none of these harms or benefits would occur.

However, indirect harms and benefits arising from the product being marketed may still occur even if the marketing system did not exist: people may acquire the products without receiving an intervention. Therefore a suitable incidence measure for indirect harms and benefits would only count those additional harms and benefits that arose from the system's operation. One such incidence measure is the *lift* rate, defined as the difference between the incidence rate of the deployed system under study, $P_d(\cdot)$, and the incidence rate measured on a control cohort that does not receive the marketing intervention, $P_c(\cdot)$,

$$Z(\cdot) = P_d(\cdot) - P_c(\cdot)$$

These harms can be reasoned about in a counterfactual sense, that is "what is the incident rate difference between the AIDA marketing system compared to the status quo". This generalises the notion of true lift or uplift [19] modelling to measure the harms and benefits of an AIDA marketing system. These measures require access to a control dataset to estimate $P_c(\cdot)$. There are numerous experimental and observational methods for obtaining such datasets, see [6]. A simple means of acquiring this control dataset would be to release the financial products that are to be marketed for a period of time without simultaneously engaging in the marketing intervention that would be the subject of this analysis. The resulting data, under certain circumstances and assumptions, would yield information about the "status quo". This method may not be applicable to all products or marketing interventions, however.

The following are incidence rates defined for the typical harms and benefits listed above. They are presented as probabilities or lift scores derived from probabilities. These probabilities will have to be estimated — this may be achieved by using empirical counts or by using more sophisticated model-based approaches such as classification or regression modeling. For illustrative purposes, the explanation assumes that A = sex and a = women:

Benefit of receiving the intervention:

$P(S = 1 | A = a)$

In the context of direct marketing, this probability can likely be estimated from empirical counts of selection outcomes (and either the bootstrapping or Bayesian methods can be used to yield uncertainty estimates). This is the number of women selected, divided by the number of women in the query cohort. For example, given an incidence rate of 0.2, one could say that "women in the cohort are selected 20% of the time". If compared across groups, this is the same as the demographic or statistical parity fairness measure (see Section 3.5).

Harm from receiving unwanted intervention:

$P(S = 1 | \text{App} = 0, A = a)$

Again, in many circumstances empirical counts of outcomes can be used to estimate this probability.

This is the number of women selected that did not apply, divided by the number of women that did not apply. Given an incidence rate of 0.3, one could say that "women that don't apply for the product receive the intervention 30% of the time". This is simply the false positive rate of the system, and if compared across groups, will lead to a relaxation of the equalised odds fairness measure [15].

Benefit from acquiring the product:

$Z(\text{Acq}=1 | A=a = P_d(\text{Acq}=1 | A=a) - P_c(\text{Acq}=1 | A=a)$

This is the additional number of women that acquired the product, divided by the number of women in the cohort. Given an impact rate of 0.1, one could say that "the marketing system caused a 10% increase in the rate women acquire the product". Empirical or model based methods will be required to estimate this quantity, see [19] for a review of relevant methods.

Harm from a failed application:

$Z(Acq=0|App=1, A=a) = P_d(Acq=0|App=1, A=a) - P_c(Acq=0|App=1, A=a)$ or

$Z(Acq=0, App=1|A=a) = P_d(Acq=0, App=1|A=a) - P_c(Acq=0, App=1|A=a)$

The first term is the change in rejection rate of those who applied, the second is the change in the rejection rate of the deployment cohort. An example of the first measure is the number of additional women that applied for the product and had that application rejected, divided by the number of women in the cohort who applied. Given an incidence rate of 0.1, one could say that "the marketing system caused a 10% increase in the rate at which women had their applications rejected". Similarly to "benefit from acquiring the product", empirical or model based methods will be required to estimate this quantity, see [19] for a review of relevant methods.

Harm from a long term outcome:

$Z(Rvd=0|Acq=1, A=a) = P_d(Rvd=0|Acq=1, A=a) - P_c(Rvd=0|Acq=1, A=a)$ or

$Z(Rvd=0, Acq=1|A=a) = P_d(Rvd=0, Acq=1|A=a) - P_c(Rvd=0, Acq=1|A=a)$

The first term is the change in probability of harm from long term outcome for those who acquired the product, and the second is the probability of harm from the long term outcome for the whole deployment cohort. An example of the first measure is the number of additional women that acquired a loan and subsequently defaulted, divided by the number of women that acquired the loan. Given an incidence rate of 0.05, we might say that "5% more women that acquired loans in the cohort defaulted on them because of the system". Again empirical or model based methods will be required to estimate this quantity, see [19] for a review of relevant methods.

Feedback and long term impacts on fairness

A direct marketing system by its very nature must intervene in the world by attempting to persuade customers to acquire products they may have not otherwise. Through its actions it is changing which groups and individuals obtain products, and consequently the outcomes these individuals and groups experience as a result of obtaining these products. In this case *feedback cycles of systemic disadvantage* can occur.

For example: if an individual from a community that has historically been systematically undersupplied with loans is given a loan, even if that person defaults, the community as a whole may benefit. This is the logic underlying the idea of using different thresholds as a strategy to mitigate unfairness. Lowering the threshold for a group will lead to more loans for individuals for that group, but unless this actually changes their expected ability to pay, the result of lower thresholds may principally be more defaults. This could impact the community negatively, leading to unmanaged debt and higher credit rejection rates over time.

Note that whilst feedback effects such as these may be relevant to fairness considerations in many AIDA systems, work to understand how to measure and control them is at an early stage of development. For recent research into measuring long term fairness in feedback scenarios that may be particularly relevant to direct marketing systems see [22,10].

These reflections may help practitioners develop their responses to the questions in this section:

- Can the definition of harms and benefits capture upstream issues (such as a lack of representation of some groups in the cohort)?
- Can the definition of harms and benefits capture downstream issues (such as the impact on a customer's future spending or borrowing habits of the marketing intervention)?

2.2.4 Part D: justify the use of personal attributes

No marketing-specific considerations.

2.2.5 Part E: examine system monitoring and review

No marketing-specific considerations.

2.3 Synthetic uplift case study

The following is a running example of a hypothetical, simulated, AIDA direct marketing system used for marketing unsecured loans. Please note that this running example:

- *is evaluated at a high level of detail to illustrate the Methodology applied to a higher risk system*
- *is an example assessment of a system determined by a fictional FSI to be higher risk, not guidance for FSIs on the actual risk associated with this example (for more details on the risk-based approach of the Methodology see Document 1 Section 2)*
- *is intended to be a simple illustration of how to use the Methodology*
- *does not represent any AIDA systems in place at any of the Consortium members*
- *should not be taken as guidance for any context- or value-sensitive decision such as choices of fairness objectives, measures, or personal attributes*
- *is not intended to constrain the scope of the Methodology: other uses may have different interventions, products, objectives, and use of AIDA systems*
- *uses simulated data that is not intended to depict realistic statistical relationships or performance measures*
- *has omitted some analyses for the sake of brevity (for example, those relating to harms from default)*

The terms "we", "us" or "our" in the running example refer to the functional author of the assessment and not to members of the Consortium as elsewhere in this document.

A (fictional) FSI with a new unsecured "fast & simple" loan product would like to embark on a marketing campaign to its existing (non-exempt) customer base. The FSI will make a profit from the interest rate payments of this product, but this profit will be offset by the cost of the marketing campaign that is carried out through a call centre operated by a different subsidiary company. The purpose of the marketing system under analysis is to select existing customers for a marketing call to increase sales of the product.

The code to run some this analysis can be found in the following GitHub repo https://github.com/veritas-project/phase1/

2.3.1 Part A: describe system objectives and context

A1 *What are the business objectives of the system and how is AIDA used to achieve these objectives?*

The overall objective of this marketing system is to maximise profits from interest payments, compared to the counterfactual baseline case where the marketing system does not exist — where only "walk-in" and online customers are offered the product.

- The product is an unsecured loan product with a $5,000 set principal and a 7% per annum interest rate (compounded monthly). The repayment period is 1 year, and typically $190 is made in profit on the interest payments from a loan.
- Each marketing call costs the company $20 to execute (including campaign overheads, salaries etc.).
- The marketing selection system uses a true lift [19] predictive model that attempts to rank customers in order of how likely they are to be persuaded by a marketing intervention, when they would otherwise have not acquired the product.
- Another predictive model was created that predicts how likely a customer is to be rejected for a loan application if they did or did not receive a marketing intervention. This model is similar to the lift model, but predicts "rejection lift rate".
- Only customers who applied for the loan and were accepted were considered to have successfully acquired the product.
- The selection process will then balance profit from lift against rejection rates from lift by differentially selecting locals and foreign nationals.
- Training data for these systems has been obtained from a small scale randomised controlled trial and carefully selected experimental data from previous loan products of a similar nature.
- Deployment data are from an initial eligible customer pool of 10,000 individuals. Only a subset of these individuals were selected by the predictive model for a marketing intervention.
- The marketing intervention (a call) is scripted to be similar for all customers and allows them to apply for the product over the phone or to be sent an electronic application form.
- The call includes a discussion of the clients needs, and introduces this product as a potential solution for quickly acquiring a small loan amount.
- Customers cannot be "pre-approved" and so have to apply for the product, however only customers with a high likelihood of being approved are contacted.

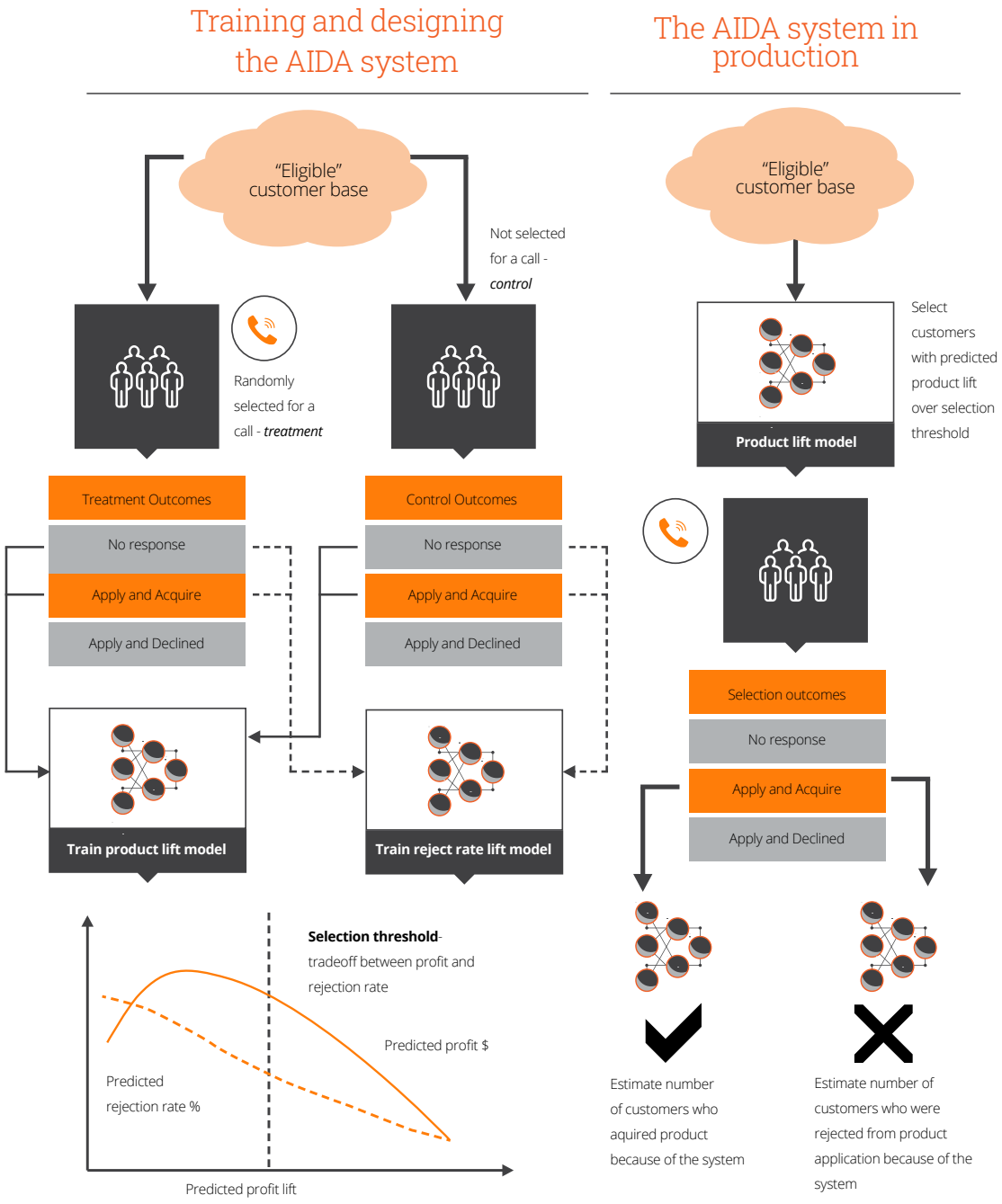


Figure 2.1 - Overview of the AIDA marketing system in training and design, and production phases.

In the training and design phase, a small randomised control trial is conducted to gauge the effectiveness of the marketing intervention and to train the product lift and rejection rate lift models. These models are then used to predict the profit and rejection rate lift for a particular selection of customers based on their predicted product lift score. With these predictions, a selection threshold (or thresholds) is chosen that trades off profit from the system with harm from failed applications. This trained system is then used in production

to select eligible customers (those who aren't ruled out by their age, contact status etc.) from the deployment customer base for a marketing intervention. Using the outcome from these selected customers, and the trained models to predict the outcomes of the selected customers if the marketing system did not exist, we can estimate the effect of the model on profits and rejection rate. This system is also illustrated graphically in Figure 2.1.

A2 *Who are the individuals and groups that are considered to be at-risk of being systematically disadvantaged by the system?*

We identified foreign nationals as being at-risk compared to local customers for this simple example.

[NOTE: In reality many personal attributes would likely be correlated with loan application rejection rates, and this simplification of the simulation and analysis is for expositional purposes.]

A3 *What are potential harms and benefits created by the system's operation that are relevant to the risk of systematically disadvantaging the individuals and groups in A2?*

- We have identified a *benefit* by the customer of acquiring the loan if they had otherwise not known about it and had an unfulfilled need.
- We have identified a *harm* to the customer of having a loan application declined, which will go on their financial record.
- We have identified a (small) *benefit* to the customer from receiving the intervention, which provides them with the opportunity to apply for credit products without waiting in a queue at a branch or on the phone.
- We have identified a *harm* to the customer of having to default on a loan that they would have otherwise not acquired. *[NOTE: This harm has not been analysed in this example in order to make the example more succinct]*

Previous analysis of the institution's loan application records has revealed that non-Singaporean customers are more likely to have a loan application rejected, even after adjusting for income and other relevant factors. The sex and age of the customers show no significant correlation with rejection rates once they have been adjusted for income. Given that the marketing system can bring about a higher rate of applications and hence application rejections, it was decided that care needed to be taken to ensure the marketing system is not amplifying the already disparate rate of loan application rejection on foreign nationals. This is to ensure that the operation of the marketing system does not further harm the creditworthiness of foreign nationals as a group. That is, we have decided that the marketing system must not increase loan application rejection rates of foreign nationals more than local customers. Furthermore, the loan application process will be investigated to uncover the reasons for this disparate rate of rejection.

A4 What are the fairness objectives of the system, with respect to the individuals and groups in A2 and the harms and benefits in A3?

- The marketing system should be constrained so as to not increase disparate harms to different groups in society (for example, increasing average loan application rejection rates).
- The benefit to the customer in receiving a loan they would not otherwise have acquired is exactly proportional to the profit uplift of the system, so this benefit need not be tracked separately.
- By persuading additional customers to apply for loans, the marketing system will in all likelihood increase the harms from the application process and from outcomes of the loans of those people. The profits of the system need to be viewed in light of, and traded off against, these harms.
- Customers who are under 18 years old, are on a "do not disturb" list, or have been contacted within the last three months are automatically excluded from selection.
- Because the benefit from receiving the intervention is small, equality in this benefit (i.e. demographic parity of selection) is monitored rather than enforced.

2.3.2 Part B: examine data and models for unintentional bias

B1 *What errors, biases or properties are present in the data used by the system that may impact the system's fairness?*

The data used for the operation and validation of the system is information related to the marketing and credit history and demographic makeup of the customer base:

- The targets of the system include whether or not a customer applied for a loan product, and whether or not their application was accepted. See Table 2.1 for the exact targets used.
- The covariates of the system contain information about the self-reported income of the customer, how many existing products they have with the bank, their nationality, age and if they have responded to a previous marketing campaign. See Table 2.2.
- The training targets and covariates were from a randomised control trial conducted with the marketing intervention and product.
  - The control group allows for prediction of outcomes "as if the system did not exist"

- The treatment group allows us to predict the outcomes of people who were selected for a marketing intervention, and so helps to simulate what the system would behave like in deployment. Note that everyone in this group is "selected" by the system, unlike in the actual deployment phase.

The covariates (only) are also available for the deployment cohort.

- Sex of the individuals is also available, however it was found not to be a predictive feature in conjunction with age and income. Though it was used to assess the fairness of the system.

Cohort, G	Acquisition target		Rejection target	
Treatment (*t*)	treatment responder acquires product	*P-T*1	treatment responder's application is rejected	*R-T*1
	treatment responder or non-responder does not acquire product	*P-T*0	treatment otherwise	*R-T*0
Control (*c*)	control responder acquires product	*P-C*1	control responder's application is rejected	*R-C*1
	control responder or non-responder does not acquire product	*P-C*0	control otherwise	*R-C*0

Table 2.1: Target conditions used for training the models in the system. There are four possible target values for a model to predict product acquisition, and for a model to predict rejection.

As discussed in [19], these targets yield information about how customers respond to the marketing intervention (call):

- "Persuadable" customers are the primary target of a marketing intervention, they will respond to a call and acquire the product, where they otherwise would have not acquired the product. These individuals are in the *P-T*1 and *P-C*0 target cohorts.
- "Sure-things" are customers who will always acquire the product, therefore should not be selected for a call. They are in the *P-T*1 and *P-C*1 target cohorts.
- Customers who would never acquire the product, even with a call are referred to as "lost causes", and should also not be selected. These customers are in the *P-T*0 and *P-C*0 target cohorts.
- Customers who negatively respond to a call and do not acquire the product as a consequence, where they would have otherwise are referred to as "do not disturbs", and should also not be selected. They are in the *P-T*0 and *P-C*1 cohorts.

An analogous breakdown of customer types also exists for the rejection lift rate model, where the customers of primary concern are those that the system actively persuades to apply for the product, and are rejected. These customers are in the *R-T*1 and *R-C*0 target cohorts.

ID	INCOME	NO_PRODUCTS	DID_RESPOND	AGE	IS_FOREIGN
0	57052	2	0	41	0
1	29838	2	0	27	0
2	31943	1	0	43	0
3	40282	2	0	35	0
4	45470	4	0	38	0
5	50293	1	0	36	0
6	52133	1	0	50	1
7	32778	2	0	39	1
8	26585	0	0	42	0
9	23182	3	0	45	0

Table 2.2: Sample of the possible system covariates. IS_FEMALE was excluded from the prediction models.

The experimental data used to train the models are from randomised trials on representative samples of the eligible customer base. In our FSI, new products are not typically marketed beyond store banners, the institution homepage or customer emails for a number of months after they are first introduced, the resulting "walk in" data serves as the control group. Negatives for this control group are generated from customers who are signed into the website, sent emails or interact with a teller in the store, and who do not apply for the product. Then a random sample of customers with a similar distribution of covariates as the control group are chosen for a marketing intervention (call), this makes up the treatment group. These data are then used to train the models, and estimate the performance of the system on the final deployment cohort.

The main issues with the data that may impact the fair operation of the system are representational

- the treatment group is smaller than the control (60% of the training data was from the control group)
- females and foreign nationals are less represented at 40% and 30% respectively

Also of note is the higher rate of application rejection amongst foreign nationals previously described.

B2 *How are these impacts being mitigated?*

- To account for the disparate control and treatment group sizes, we use method 2B from [19] for both the product and rejection rate lift models. This method adjusts the predicted lift score to account for the different base rates in the cohorts.
- No attempt was made to correct for the lower representation of females in the dataset. Even though the final selection rate was slightly higher for males, the difference was minor.
- Different predicted lift selection thresholds were used to select foreign nationals and locals in an attempt to reduce the disparate effects of the marketing system on application rejection rates between these two groups. More detail is given in answers in Parts C and D.

B3 *How does the system use AIDA models (with, or separately from, business rules and human judgement) to achieve its objectives?*

Two models are used in the marketing system, the first is trained to predict the probability of *loan application rejections* in the control and treatment group, the second is to predict the probability of *a person applying for and acquiring* a loan product in the control and treatment group.

The loan rejection rate predictor is used to estimate the harms from the system's operation, the loan acquisition predictor is used to rank and select customers for interventions, and also to estimate the profit from the system's operation.

The training inputs to the model have already been disclosed in Tables 2.1 (targets) and 2.2 (covariates). The deployment inputs are the same columns as in Table 2.2, and the output of the product lift model is a binary variable indicating whether or not a customer should be called. The output of the rejection rate lift model is a predicted probability of each individual belonging to one of the *R-XX* target classes. These probabilities are subsequently used to estimate the probability the marketing system will cause them to be rejected from a loan application.

Both of these models are *multiclass logistic regression* predictors that minimise *cross entropy loss* as their objective (see the definition of $\mathcal{L}_{\text{cross ent.}}$ below). The targets for these methods are shown in Table 2.1. See method 2B from [19] for more details on the exact losses and models used. They were also tested for balanced accuracy. More complex models (that still used a proper scoring loss) were tested, such as Gaussian processes and feed-forward neural networks, but the predictive power of these models was no better than the simpler multiclass logistic regression algorithm finally selected.

These loss functions, and their associated targets, are only proxies for the objectives and constraints. This is because it is not possible to measure directly if our marketing system causes someone to acquire the loan product, or cause someone's application for the loan

product to be rejected. Measuring these directly would require access to a counterfactual reality where our marketing system does not exist. Using the loss functions (and targets) of method 2B from [19] is the state-of-the-art for constructing models to estimate the outcomes of interest.

- The models use a proper scoring rule for their loss (cross entropy), and so should yield an unbiased estimate of the posterior class probabilities. However, rather than reporting this loss directly, it is standardised against a naive prediction model, a categorical distribution (see the definition of $\mathcal{L}_{\text{categorical}}$, that can only predict the targets based on their proportions in the data. For example, in the case of the rejection prediction model,

$$\mathcal{L}_{\text{cross ent.}}(y,f) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k\in\{\text{R-T1},\ldots,\text{R-C0}\}} \mathbf{1}(y_i = k)\log f_k(X_i),$$

$$\mathcal{L}_{\text{categorical.}}(y) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k\in\{\text{R-T1},\ldots,\text{R-C0}\}} \mathbf{1}(y_i = k)\log \hat{p_k}, \quad \text{where } \hat{p_k} = -\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(y_i = k),$$

$$\mathcal{L}_{\text{std.}}(y,f) = \mathcal{L}_{\text{cross ent.}}(y,f) - \mathcal{L}_{\text{categorical.}}(y).$$

  Here y are the targets, $f_k(x)$, is the predictive model's output of the probability of the target taking value k (from Table 2.1) , x, the input covariates, N the size of the dataset to be evaluated. Also, 1(.) is an indicator function returning 1 is the condition in the brackets is true, otherwise 0. The intuition behind this measure is that it will tell us how much better the prediction model is than a random guess.

There are two primary measures of the performance of the system that are related to the system objectives and constraints;

Uplift profit

the profit solely because of the marketing system (compared to the scenario where the marketing system did not exist).

Uplift rejection rate

the rejection rate solely because of the marketing system (compared to the scenario where the marketing system did not exist).

These have been chosen in light of the presented harm and benefits presented in Section 2.2. Note that the harm is quantified in discrete units of occurrence, rather than an estimate of the *magnitude* of harm caused to a particular individual. Rates of occurrence do not encode the fact that loan application rejections may harm some more than others. It is our eventual goal to be able to estimate how harm varies between applicants or between groups.

Unfortunately uplift measures described above are not directly observable from data and have to be estimated using the loan acquisition and application rejection estimators. These measures can be estimated from the training and testing data as follows,

Product uplift — training/testing:

$$\mathbb{E}[\text{Profit}] = \sum_{i \in S} p_t(\text{Acq}_i=1, \text{App}_i=1 \,|\, \mathbf{X}_i) \times \text{Profit}(\text{Acq}_i=1, S=1) - P_c(\text{Acq}_i=1, \text{App}_i=1 \,|\, \mathbf{X}_i) \times \text{Profit}(\text{Acq}_i=1, S=0)$$

Where $\mathbb{E}(\cdot)$ denotes statistical expectation, $S \in \{0.1\}$ indicates if a customer has been selected for an intervention, and S is the set of customers selected for a marketing intervention.

After selection, there is a sequence of possible stages of outcome (see Section 4.2). To simplify notation the harm and benefit analysis utilises shortened references to possible outcomes (all of which may take values in $\{0.1\}$):

Outcome	Shorthand
Applied for loan	App
Acquired loan	Acq

Also, $P_t(.|\mathbf{x})$ is the probability of the outcome on the treated group, and $P_c(.|\mathbf{x})$ is the probability of the outcome on the control group. Both of these quantities we derive from the multiclass logistic regression acquisition prediction model, f. The profit functions are defined as,

$\text{Profit}(\text{Acq}_i=1, S=1) = \text{Interest Revenue}_i - \text{Treatment Cost}_i$

$\text{Profit}(\text{Acq}_i=1, S=0) = \text{Interest Revenue}_i$

The prediction probabilities obtained from the multiclass classifier have not been adjusted for the disparate size of the control group, $G = c$, and the treatment group, $G = t$, that is,

$P(\text{Acq}_i=1, \text{App}_i=1, G_i=t \,|\, \mathbf{x}_i) \approx f_{\text{P-T1}}(\mathbf{x}_i)$, and

$P(\text{Acq}_i=1, \text{App}_i=1, G_i=c \,|\, \mathbf{x}_i) \approx f_{\text{P-C1}}(\mathbf{x}_i)$.

To approximate the treatment and control distributions we can re-normalise these predictions,

$$P_t(\text{Acq}_i=1, \text{App}_i=1 \,|\, \mathbf{x}_i) = \frac{P(\text{Acq}_i=1, \text{App}_i=1, G_i=t \,|\, \mathbf{x}_i)}{P(G_i=t)} \approx \frac{f_{\text{P-T1}}(\mathbf{x}_i)}{\sum_{i=1}^{N} \mathbf{1}(G_i=t) \Big/ N},$$

and similarly,

$$P_c(\text{Acq}_i=1, \text{App}_i=1 \mid \mathbf{x}_i) = \frac{P(\text{Acq}_i=1, \text{App}_i=1, G_i=c \mid \mathbf{x}_i)}{P(G_i=c)} \approx \frac{f_{\text{P-C1}}(\mathbf{x}_i)}{\sum_{i=1}^{N} \mathbf{1}(G_i=c)/N}.$$

see [19] for more details on uplift estimation models, in particular method 2B was used to inform this approach.

Rejection uplift rate — training/testing:

$$Z(\text{Acq}=0, \text{App}=1) \approx \frac{1}{N} \sum_{i \in S} P_t(\text{Acq}_i=0, \text{App}=1 \mid \mathbf{x}_i) - P_c(\text{Acq}_i=0, \text{App}=1 \mid \mathbf{x}_i)$$

This directly estimates the "harm from a failed application" objective (see Section 4.2). Similarly to the product uplift multiclass classifier, the rejection multiclass classifier has to have its prediction probabilities re-normalised to estimate these treatment and control distributions. That is,

$P(\text{Acq}_i=0, \text{App}_i=1, G_i=t \mid \text{x}_i) \approx f_{\text{R-T1}}(\mathbf{x}_i)$, and

$P(\text{Acq}_i=0, \text{App}_i=1, G_i=c \mid \text{x}_i) \approx f_{\text{R-C1}}(\mathbf{x}_i)$.

Where re-normalisation proceeds in the same manner as the product uplift model,

$$P_t(\text{Acq}_i=0, \text{App}_i=1 \mid \mathbf{x}_i) \quad \frac{P(\text{Acq}_i=0, \text{App}_i=1, G_i=t \mid \mathbf{x}_i)}{P(G_i=t)} \approx \frac{f_{\text{R-T1}}(\mathbf{x}_i)}{\sum_{i=1}^{N} \mathbf{1}(G_i=t)/N},$$

And for the control,

$$P_c(\text{Acq}_i=0, \text{App}_i=1 \mid \mathbf{x}_i) = \frac{P(\text{Acq}_i=0, \text{App}_i=1, G_i=c \mid \mathbf{x}_i)}{P(G_i=c)} \approx \frac{f_{\text{R-C1}}(\mathbf{x}_i)}{\sum_{i=1}^{N} \mathbf{1}(G_i=c)/N},$$

These *measures are partially* observable in a deployed system since the system has made interventions and we have access to the outcomes of these interventions. The counterfactual outcomes "if the marketing system did not exist" are not available, and still have to be estimated from the control data.

Product uplift — deployment:

$$\mathbb{E}[\text{Profit}] \approx \sum_{i \in S} 1(\text{Acq}_i=1 \square \text{App}_i=1 \times \text{Profit}(\text{Acq}_i=1, S_i=1) - P_c(\text{Acq}_i=1, \text{App}_i=1 \mid \mathbf{X}_i) \times \text{Profit}(\text{Acq}_i=1, S_i=0)$$

This is similar to the training data measure, however, the deployment predictor has been replaced with empirical counts of the results of the system's actions, where □ is a logical "and" operator.

Rejection uplift rate — deployment:

$$Z(\text{Acq}=0, \text{App}=1) \approx \frac{1}{N} \sum_{i \in S} 1(\text{Acq}_i=0 \square \text{App}_i=1 - P_c(\text{Acq}_i=0, \text{App}_i=1 \mid \mathbf{x}_i)$$

Similarly it is possible to replace the deployment estimations with empirical counts of observed outcomes from the deployed marketing system.

B4 *What are the performance estimates of the AIDA models in the system and the uncertainties in those estimates?*

For the model performance measures a test set is used, which comprises a subset of individuals involved in the randomised control trial. This dataset was not used for model selection purposes, and it allows for measures such as cross entropy loss and balanced accuracy to be assessed.

Uncertainty for each of the aforementioned measures is calculated using the empirical bootstrap method on the test data as suggested by Document 1 Appendix 1.1. 50 sample replications were used, and 5-95% confidence intervals are reported for each measure.

Measure	Value (mean)	Lower 5%	Upper 95%
Standardised cross entropy loss - rejection model	0.021	0.018	0.024
Standardised cross entropy loss - product model	0.103	0.097	0.110
Balanced accuracy - rejection model	0.250	0.250	0.250
Balanced accuracy - product model	0.343	0.340	0.349

Table 2.3: Marketing system performance measures on the hold-out test set. Standardised cross entropy loss refers to $\mathcal{L}_{\text{std.}}$ in Response B3 above.

The balanced accuracy of the rejection prediction model is very low (no better than a random selection), however its probability predictions (standardised cross entropy loss) are better than a random predictor. In particular, this standardised cross entropy loss for the rejection model is 4.6 times better on foreign nationals compared to the rest of the cohort. This is in line with the previous analysis that foreign national status is predictive of a higher rejection rate.

B5 *What are the quantitative estimates of the system's performance against its business objectives and the uncertainties in those estimates?*

For the profit and rejection uplift measures, the deployment dataset outcomes must be used along with the control outcome estimates. This cohort comprised 10,000 customers, only some of whom (1805) were selected by the system for a marketing intervention.

Uncertainty for each of the aforementioned measures is calculated (where appropriate) using the empirical bootstrap method on the deployment data as suggested in Document 1 Appendix 1.1. 50 sample replications were used, and 5-95% confidence intervals are reported for each measure.

Measure	Value (mean)	Lower 5%	Upper 95%
Number selected for intervention	1805	-	-
Estimated profit because of the marketing system	$87666.88	$79818.84	$95393.18
Estimated uplift in rejection rate because of the marketing system	0.79 %	0.64 %	0.96%
Proportion of cohort who were selected for an intervention (total)	18.05 %	-	-
Proportion of cohort who acquired the product (total)	6.49 %	-	-

Table 2.4: Marketing system performance measures on the deployment set (customer pool of 10,000).

2.3.3 Part C: measure disadvantage

C1 *What are the quantitative estimates of the system's performance against its fairness objectives, assessed over the individuals and groups in A2 and the potential harms and benefits in A3?*

The measures of harms and benefits considered in the marketing system are, in order of priority,

- Benefit to customers from acquiring the loan due to marketing system: proportional to profit
- Harm to customers from a failed application, Z(Acq=0, App=1, A=a), estimated as described in Response B3 on each group, a.
- Benefit of receiving the intervention (demographic parity), P(S=1|A=a), measured directly from the marketing intervention selection system.
- Harm from a default: *not addressed in this example for the sake of brevity*.

It was decided that there was a real danger of the marketing system amplifying an already existing disparate loan application rejection rate on foreign nationals (discovered in the loan application system). This could potentially lead to a worse average financial record for this group, which could in turn further amplify the loan application rejection rate disparities. Benefit of receiving the intervention is very much a secondary goal. It is expected this measure will potentially become worse while disparate harms from failed applications are minimised. The benefit from acquiring the product, if assumed to be constant across the population, is a direct function of product acquisition uplift and therefore need not be

separately computed for this example. The amount of uplift profit from each group will be considered in this analysis.

Measure	Foreign nationals			Locals		
	Mean	*Lower 5%*	*Upper 95%*	*Mean*	*Lower 5%*	*Upper 95%*
Harm from failed application (rejection rate uplift)	0.87 %	0.67 %	1.16 %	0.74 %	0.55 %	0.92 %
Benefit from receiving intervention (proportion selected)	7.2 %	6.3 %	8.1 %	23 %	21 %	24 %
Predicted profit increase from those selected	$7300	$5500	$9600	$81000	$75000	$88000
Lift selection threshold	0.5	-	-	0.4	-	-
Number selected	215	-	-	1606	-	-

Table 2.5: The harm and benefit measures on foreign nationals and locals from the deployed system. Empirical bootstrapping with 50 sample replications was used to estimate the confidence intervals.

We haven't explicitly used any measures of individual fairness that require defining a similarity function. Our chosen fairness objective "harm from a failed application" requires prediction in order to estimate, and is less reliable (higher variance) when viewed at an individual level, and so we only report it aggregated over groups of customers. We do report the "benefit of receiving an intervention" measure with respect to age and income distributions separately in Response D3, which uncovers the distribution of marketing interventions over individuals in the deployment cohort.

Selection of customers for a marketing intervention is based on the predicted product lift from an individual,

$$Z(App_i=1, Acq_i=1 \mid x_i) = P_t(Acq_i=1, App_i=1 \mid \mathbf{x}_i) - P_c(Acq_i=1, App_i=1 \mid \mathbf{x}_i).$$

This is the same as the "benefit of acquiring the product" incidence rate measure. This quantity is predicted for all customers in the deployment cohort, and then all customers over a lift threshold are selected for a marketing intervention. A different threshold is applied to local customers and foreign nationals. These thresholds are chosen based on the predicted profit and the predicted rejection rate increase described in Response B3. Hence, the set of operating parameters that most directly affect the system performance are:

- the product lift threshold on locals (0.4 was chosen),
- the product lift threshold on foreign nationals (0.5 was chosen).

The choice of these parameters and their setting is justified in the next sections.

C2 *What are the achievable tradeoffs between the system's fairness objectives and its other objectives?*

For the rejection from a failed application harm (rejection lift rate) it was easier to quantify the rate of harm on each group separately. This is because the original magnitude of the harm on each group is important in ensuring it is minimised. For instance, we decided that the rejection lift rate should be no higher than 1% on locals and foreign nationals. This also had the additional benefit of simplifying the analysis of the system, and clarifying the subsequent choice of selection thresholds to apply on each group to achieve this objective.

Figure 2.2 shows predicted profit and rejection rate increases for all possible choices of foreign national and local product positive lift selection thresholds on the test data. Note, this overall rejection rate is not a fairness objective, however, it gives a general indication of the harm / benefit tradeoffs in the system. For detailed discussion of fairness objective tradeoffs, see Response C3 (below).

Generally we can see that as the thresholds are lowered (and more customers are selected), profit lift tends to increase, but also so does rejection lift rate. The maximum of the profit surface does not coincide with the maximum of the rejection surface, but the correspondence is still close. This requires the FSI to have to almost directly tradeoff profit for a rejection rate decrease.

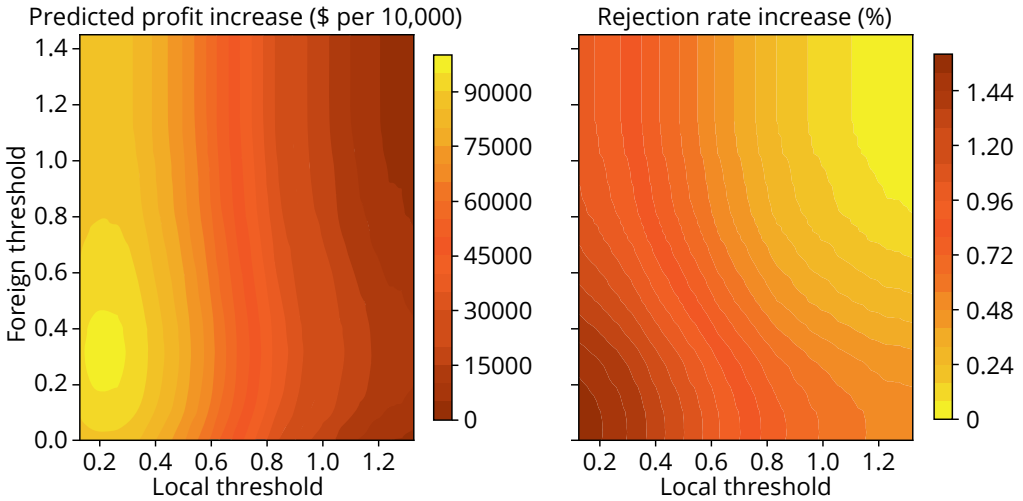


Figure 2.2: Predicted profit lift and rejection lift rate for all choices of product lift thresholds on the test data. Warmer colours are more desirable according to the objectives of the marketing system.

C3 *Why are the fairness outcomes observed in the system preferable to these alternative tradeoffs?*

Upon analysing the prediction results on the test set, we decided that the rejection lift rate should not exceed 1% on average for either local or foreign nationals because of the marketing system (inline with our goals for the system as stated in Part A). Setting this objective then dictates the thresholds chosen, and consequently the number of individuals selected from both groups, and the amount of profit the system will generate. Figure 2.3 visualises the tradeoffs between profit and rejection rate for foreign nationals, and Figure 2.4 for locals. From these, a conservative threshold of 0.5 was chosen for foreign nationals and 0.4 for locals. Unfortunately it is not possible to obtain profit from the system without incurring a cost in an increase in loan application rejections. We have justified our choice of thresholds by setting the maximum level of rejection rate lift created by the system at 1%, and then maximising profit within this constraint.

As can be seen from the outcomes on the deployment data, these rejections rates were actually overestimated on the test data, with foreign nationals having a predicted increased rejection rate of 0.873% and locals 0.734% in deployment. This can be explained by the rejection lift prediction model being more predictive on foreign nationals as indicated by the test data (Table 2.6):

Measure (test data)	Value (mean)	5% lower	95% upper
Standardised cross entropy of cohort	0.021	0.018	0.024
Std. cross entropy foreign / std. cross entropy local	4.74	1.309	6.538

Table 2.6: Model performance (standardised cross entropy loss) of the rejection lift rate model on the test data. The model is more predictive on the foreign national cohort than on locals.

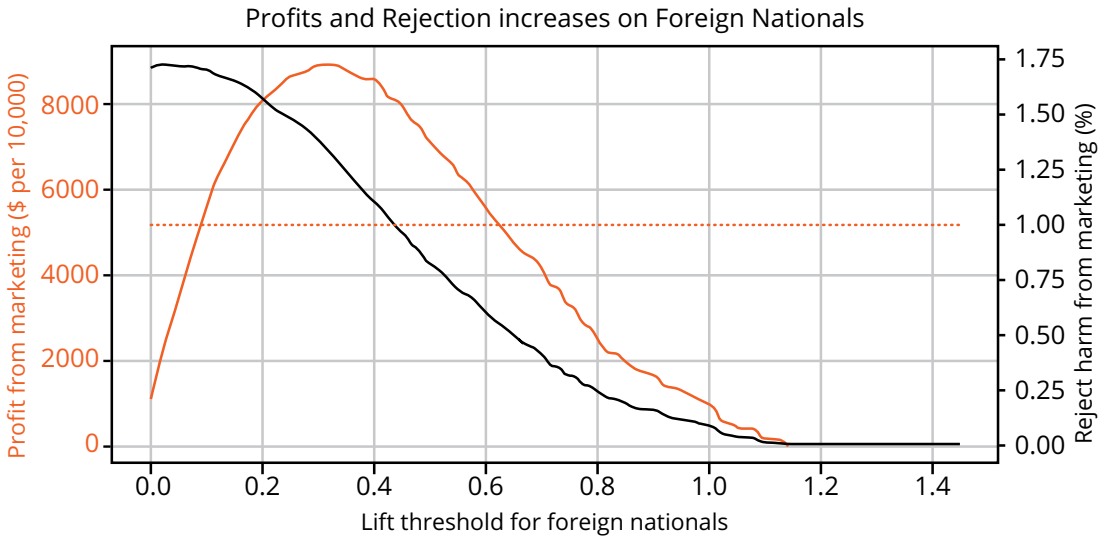


Figure 2.3: Lift threshold for foreign nationals, and predicted profit and rejection rate increase at the threshold level on the test data. A conservative level of 0.5 was chosen for foreign nationals to keep the rejection rate lift below 1%.

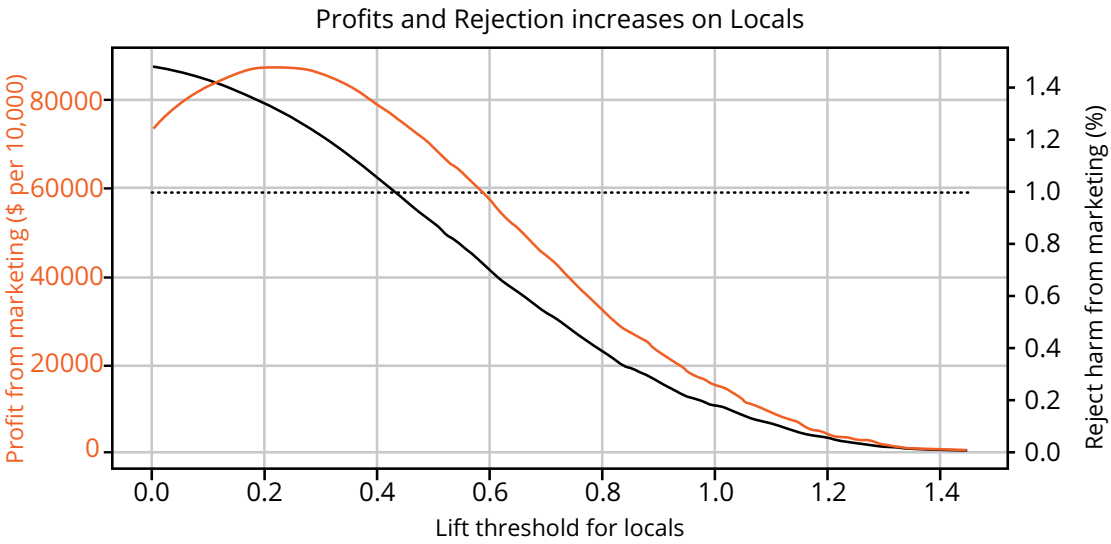


Figure 2.4: Lift threshold for locals, and predicted profit and rejection rate increase at the threshold level on the test data. A level of 0.4 was chosen for locals to keep the rejection rate at around 1%.

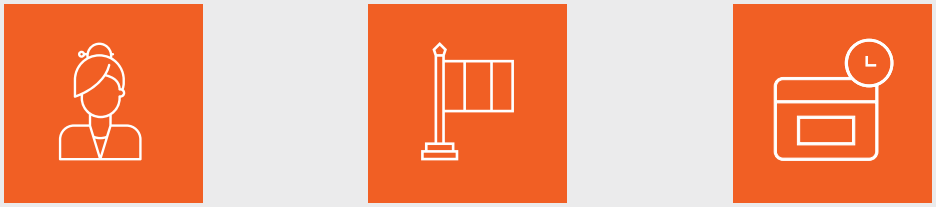
The rejection lift rate estimator will be refined before the next marketing campaign in order to better predict rejection lift rates, and better inform threshold levels for both cohorts.

Equalising rates of the "benefit of receiving an intervention" between foreign nationals and locals (which was a fairness objective) was not attempted owing to the small magnitude of this benefit. The rates remain disparate, as can be seen in Table 2.5. This is because foreign nationals are predicted on average to have a lower product acquisition

lift than locals. Furthermore increasing the number of foreign nationals being selected for a marketing intervention is predicted to directly increase the rate of harm from failed applications in the group. To address this disparity in the future, it is first necessary to understand the reasons for the higher rate of loan rejection amongst foreign nationals.

2.3.4 Part D: justify the use of personal attributes

D1 *What personal attributes are used as part of the operation or assessment of the system?*



Sex [IS_FEMALE]

Nationality status [IS_FOREIGN]

Age [AGE]

D2 *How did the process of identifying personal attributes take into account ethical objectives of the system, and the people identified as being at risk of disadvantage?*

As previously discussed, analysis of the FSI's historical data identified foreign nationals as being the most at risk of harm from the marketing system because of their historically higher loan rejection rates compared to other groups. It was anticipated that unconstrained operation of the system may have led to even higher loan application rejections for foreign nationals.

We also engaged an external customer representative group to discuss what attributes they considered private, and how comfortable they were in having them used in a marketing system. There was most consensus around sex being a personal attribute and fewest people were comfortable with its use in a predictive system. Discrimination law within other countries that the company operates in was also considered when identifying personal attributes.

D3 *For every personal attribute and potential proxy for a personal attribute, why is its inclusion justified given the system objectives, the data, and the quantified performance and fairness measures?*

We will justify the inclusion or exclusion of personal features by testing the models' predictive performance with and without including these features in the covariates. Table 2.7 summarises system performance when excluding the attributes IS_FEMALE, IS_FOREIGN and AGE from the product and rejection models. Note that the performance measures in Tables

2.7 and 2.8 can only be evaluated on the test data with experimental outcomes. The fairness measure "harm from a failed application" does not change on this data since it is used to select the lift thresholds for selection. This fairness measure cannot be measured on the deployment dataset since it would require re-deployment of the system on the same cohort, which is impossible. The benefit of selection can be assessed on deployment data however if we assume the same threshold criteria from above applies to the system (no more than 1% predicted rejection lift rate).

Excluding	Measure	Value	Lower 5%	Upper 95%
None	Std. Cross Entropy - Product	0.103	0.097	0.111
	Std. Cross Entropy - Rejection	0.020	0.016	0.023
IS_FEMALE (model in production)	Std. Cross Entropy - Product	0.103	0.097	0.110
	Std. Cross Entropy - Rejection	0.020	0.017	0.023
IS_FOREIGN	Std. Cross Entropy - Product	0.100	0.095	0.108
	Std. Cross Entropy - Rejection	0.011	0.010	0.014
AGE	Std. Cross Entropy - Product	0.084	0.078	0.089
	Std. Cross Entropy - Rejection	0.017	0.013	0.020

Table 2.7: Model performance measures when excluding attributes from the covariates on the test data. A higher value is better. Std. Cross Entropy refers to the standardised cross entropy loss model measure from Response B3.

Excluding	Measure	Foreign	Local	Female	Male
None	Benefit of receiving (%)	7.1	22.8	17.4	18.6
IS_FEMALE	Benefit of receiving (%)	7.2	22.9	16.9	19.6
IS_FOREIGN*	Benefit of receiving (%)	4.4	11.9	9.1	10.0
AGE+	Benefit of receiving (%)	0.0	0.0	0.0	0.0

** This required setting a lift threshold of 0.6 to keep the rejection lift rate below 1%, and the estimated value of rejection lift rate for foreigners was severely biased downwards.*
+ the product lift estimator had such poor performance that only negative profit could be predicted.

Table 2.8: Fairness measures when excluding attributes from the covariates on the deployment data.

We also examined the cumulative *deployment* distribution of customers selected (i.e. cumulative benefit of receiving an intervention) versus the cumulative distributions of age and income in Figures 2.5 and 2.3.6. These Probability-Probability (P-P) plots can be interpreted in a similar fashion to Lorenz curves, where a diagonal line indicates an equal distribution of selection — this is a measure of demographic parity for continuous attributes. We have also included on the plots the average product lift rates, $Z(\text{App}=1, \text{Acq}=1 \mid A=a)$, which is our selection criterion, for each percentile of age and income.

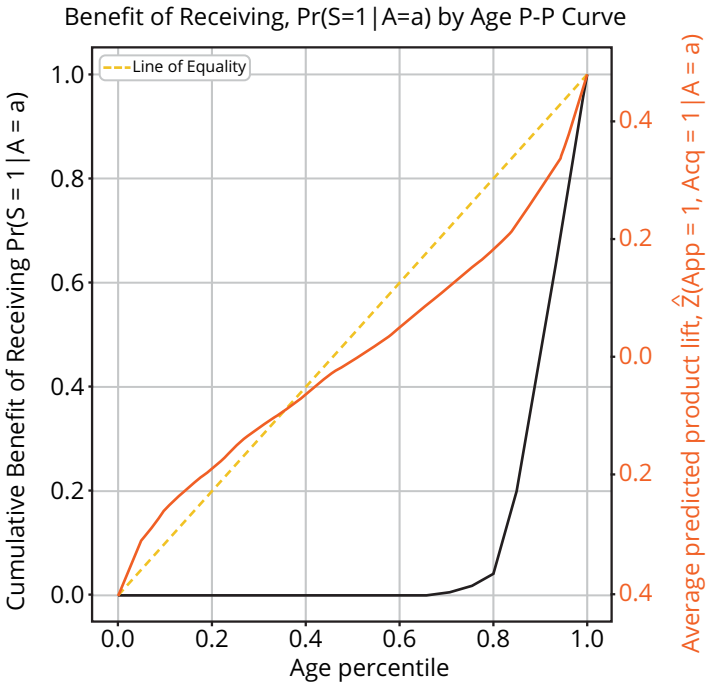


Figure 2.5: P-P plot of benefit of receiving and intervention by age. Also plotted is the average predicted product lift per age percentile, which is our selection criterion.

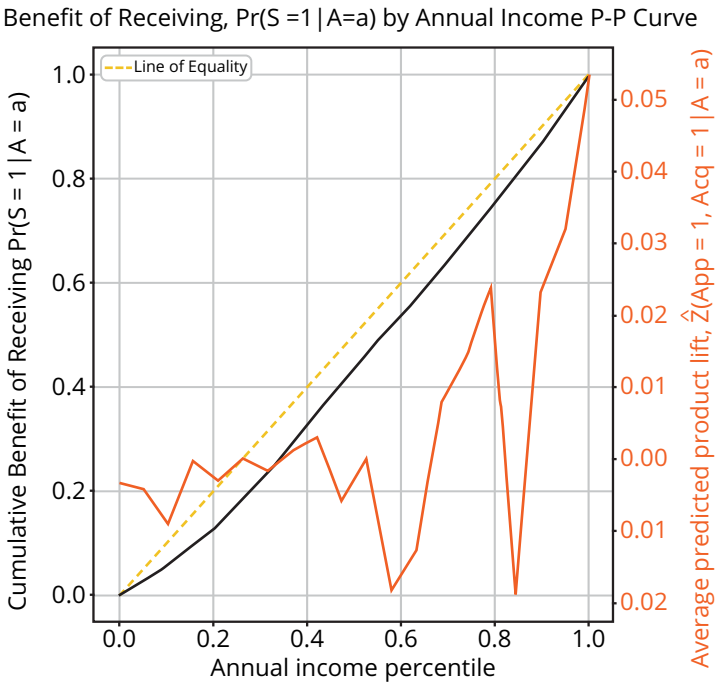


Figure 2.6: P-P plot of benefit of receiving and intervention by income. Also plotted is the average predicted product lift per income percentile, which is our selection criterion.

From these figures we can see our system has selected an unequal distribution of people for intervention by age. We justify this difference however by observing that the predicted product lift is a strongly increasing function of age, with many young people actually having negative lift, where selecting them would have made them not acquire the product when they otherwise would have. From this we can infer that the marketing system is actually distributing interventions to people, as categorised by age, who would respond (and potentially benefit) from them. The distribution of selection by income is much closer to the line of equality, and similarly predicted product lift is a much weaker function of this variable, except for the higher income levels.

To improve the way the intervention is distributed across age, we plan to modify the system's intervention to be more appealing to young people (increasing the lift in this cohort).

These results indicate that excluding the AGE and IS_FOREIGN covariates from the model detrimentally affects the model performance and the outcomes of the marketing system, hence they were included. Excluding the IS_FEMALE covariate did not significantly affect the predictive performance of the system, but excluding it did lead to slightly fewer females being selected than the models using full set of covariates. However, the internal model governance of our FSI determined that, due to legal risk across many countries of operation, the policy of our FSI is to omit sex unless it is critical to the system's performance (which we judged not to be the case).

Figure 2.7 shows a correlation matrix between all of the attributes and features in the AIDA system (not all of them are used in the models). INCOME and NO_PRODUCTS are both correlated with IS_FEMALE, and INCOME is correlated with IS_FOREIGN. Hence INCOME and NO_PRODUCTS could be viewed as potential proxy features for personal attributes.

However, they are critical for the predictive performance of the marketing system: if they are excluded, then all predictive power of the machine learning models is lost and the system cannot function to achieve any of its business for fairness objectives.

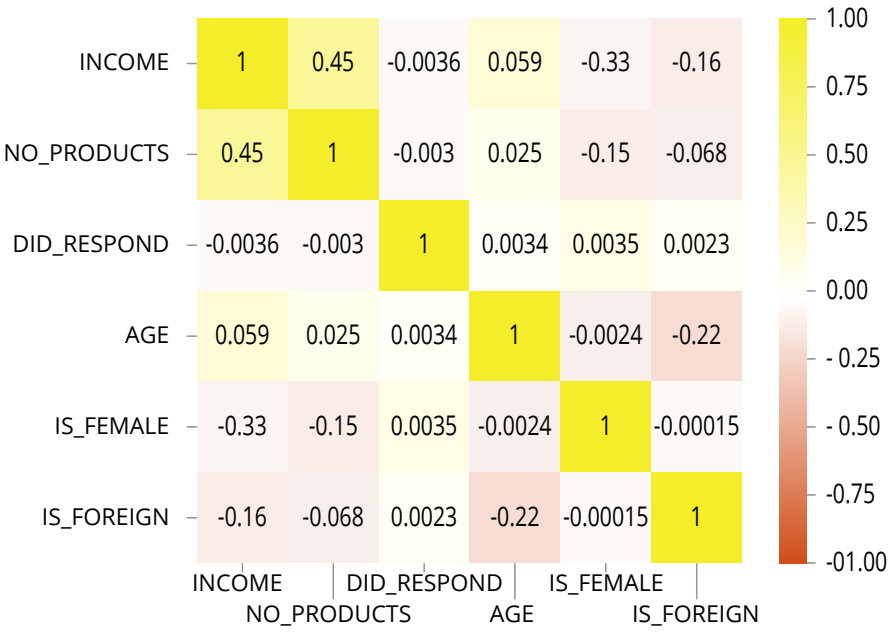
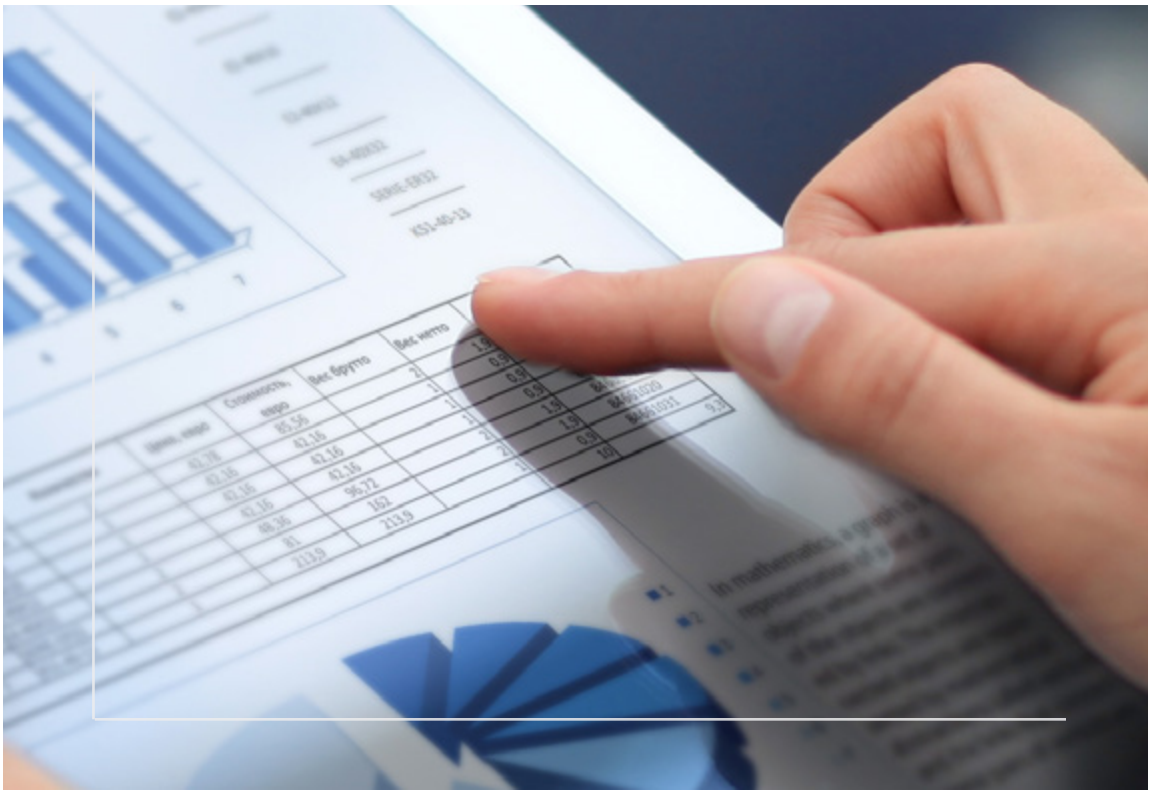


Figure 2.7: Correlation matrix of all features and personal attributes in the AIDA system.



2.3.5 Part E: examine system monitoring and review

E1 *How is the system's monitoring and review regime designed to detect abnormal operation and unintended harms to individuals or groups?*

Because of the causal nature of the marketing system, detecting changes in its performance is quite challenging. For instance, directly detecting degradation of the prediction models for product and rejection rate lift (using cross entropy loss) would require *re-running* a randomised control trial to obtain more targets, *R-XX* and *P-XX* from Table 2.1. This is because we do not obtain representative samples of these targets from the system's operation. Re-running this trial would potentially be expensive and may also increase loan application rejection rates.

However, proxy measures do exist that are monitored and regularly reviewed, primarily:

- The distribution of actual "treatment responders" conditioned on the predicted lift of the selected cohort for the product and rejection rate lift models respectively, i.e. $P(Y = 1 \mid Z(x) > z*)$ where $z*$ is the product lift selection threshold(s) and Y is the outcome (acquired or rejected respectively). These distributions are regularly reviewed and checked for consistency against the distributions of responders conditioned on corresponding lift scores in the training/test data. If these vary substantially, this may be indicative of a prior probability shift or concept shift [25]. This is also explicitly checked for foreign nationals and locals for any discrepancies.

- The distribution of covariates between the training data, $P_{train}(x)$, and the production data, $P_{query}(x)$, is regularly reviewed using the discriminative (classification) approach in [4], though no sample reweighting or retraining is implemented. If a classifier can discriminate between the training and current production covariates with a balanced accuracy greater than 0.6 accounting for uncertainty estimates, this is considered abnormal. This threshold has been chosen to be conservative based on previously deployed systems.

These properties are appropriate to monitor as they are good indicators that the inductive assumptions used to build the product and rejection lift models have been violated. Namely, they will indicate if the conditional relationships between the covariates and targets has shifted (point 1) or if the coviates distributions are different between the training and production datasets (point 2).

E2 *How does the system's monitoring and review regime ensure that the system's impacts are aligned with its fairness and other objectives (A1 and A4)?*

There are a number of monitoring and review processes and stages:

- Before a campaign is initiated, the selection pool is analysed for any covariate shift using the discriminatory method (performance measure 2). This analysis is carried out by the data scientists responsible for the ongoing operation of the system. If a significant shift is detected, then the campaign is not initiated until further analysis is done.
- During the campaign, every week the distributions of response for each model with respect to the product lift score are analysed (performance measure 1) is analysed by the data scientists responsible for the system.
- If no major issues are found with covariate or concept, the profitability of the system and the rejection lift rate is estimated weekly and reported to senior management responsible for the system.

Furthermore, there is constant communication between the team responsible for the operation of this system, and the team responsible for the loan application process. This is to ensure that if the loan application process changes, the rejection lift model is reviewed and/or re-trained.

All versions of the models used in production, including the data used to train them, are versioned with a unique hash. Each query, prediction and outcome from a model is associated with the hash of the model that consumed or produced it. All calls from the call centre are recorded and kept for a period of one year. Using this versioning system, the decisions made by the system can be completely re-produced and then reviewed in light of the system's fairness and other objectives.

E3 *What are the mechanisms for mitigating unintended harms to individuals or groups arising from the system's operation?*

Depending on the severity of the changes detected or of the issues under review, the marketing campaign may be halted. A decision will then have to be made as to whether the system's operation should be permanently discontinued, or if another trial should be conducted to re-train the models. Given the finite size of the customer base, it is expected that eventually many of the "persuadable" customers will be contacted, and the system will begin to exhibit diminishing returns with respect to the profit objective. Furthermore, channels of communication are kept open to those who may be impacted by the system:

- The customers are first given an opportunity to rate their experience of the call they receive, or to opt out from future marketing interventions. They are also a separate channel for complaints they are directed to if they so choose.
- A customer relations team also exists to manage customers complaints about their loan products, including having their application rejected.
- Although not directly related to the marketing system, customers can also request information as to why their loan application was rejected; and the loan application team has a model interpretability capability built into their models that they can use to help the customer understand why they were rejected.

All of this information is available in an anonymised manner to the designers of the marketing system.

2.4 Synthetic lower risk case study

The following is a running example of a hypothetical, simulated, AIDA direct marketing system used for marketing unsecured loans. Please note that this running example:

- *is evaluated at a low level of detail to illustrate the Methodology for a lower risk system*
- *is an example assessment of a system determined by a fictional FSI to be lower risk, not guidance for FSIs on the actual risk associated with this example (for more details on the risk-based approach of the Methodology see Document 1 Section 2)*
- *is intended to be a simple illustration of how to use the Methodology*
- *does not represent any AIDA systems in place at any of the Consortium members*
- *should not be taken as guidance for any context- or value-sensitive decision such as choices of fairness objectives, measures, or personal attributes*
- *is not intended to constrain the scope of the Methodology: other uses may have different interventions, products, objectives, and use of AIDA systems*
- *uses simulated data that is not intended to depict realistic statistical relationships or performance measures*

The terms "we", "us" or "our" in the running example refer to the functional author of the assessment and not to members of the Consortium as elsewhere in this document.

A (fictional) FSI is rolling out a targeted intervention to its customer base that has loan products. The purpose of the intervention is to make the customer aware that they can contact the FSI in the case they are experiencing financial hardship and are finding it difficult to maintain their loan repayments. The aim of this system is to make customers aware of the financial hardship services the FSI offers at an early stage of their financial hardship: ideally before they miss too many successive loan repayments and risk defaulting.

2.4.1 Part A: describe system objectives and context

A1 *What are the business objectives of the system and how is AIDA used to achieve these objectives?*

- The business objectives of the hardship outreach system is to inform as many of the customers as possible who have loans with the FSI of the FSI's financial hardship services.
- The system has a fixed budget for calls to customers, so attempts to contact those customers most likely to be experiencing financial difficulties at an early stage of their financial hardship.
- A hybrid business rule/machine learning system alerts FSI customer service employees of customers who are predicted to be experiencing financial hardship. This is based on their repayment history, and historical bank account data and demographics from previous customers who have experienced financial hardship.
- Customers that are predicted to be at risk of financial hardship are contacted directly by the bank's customer service employees.
- Because of the lack of materiality and possibility for negative consequences of the system's operation, it has been designated as low risk by the risk management process (and therefore suitable for a summary-level assessment as presented).

A2 *Who are the individuals and groups that are considered to be at-risk of being systematically disadvantaged by the system?*

reminder: this is fictitious data from a fictitious FSI and the bias described below is invented for illustrative purposes only.

Based on internal FSI data, groups at higher risk of financial hardship for the relevant loan products tend to be older, with a small bias towards females, so it is important to make sure the system effectively captures these individuals if they are not already availing themselves of the financial hardship support.

A3 *What are potential harms and benefits created by the system's operation that are relevant to the risk of systematically disadvantaging the individuals and groups in A2?*

The main potential harm would be through missed opportunities for an early discussion with

the financial hardship team. Note that a certain number of missed loan repayments from a customer will automatically prompt action from the financial hardship team.

Also, whether targeted by the system or not, these services are available to all customers. For example, a banner is displayed to all customers who log into their internet banking portal and have emailed statements from the FSI with information about how to contact the FSI to discuss any financial hardship they are experiencing. They are presented with a phone number, an email address, and an online form they can fill out - both the email and online form will result in a call to the customer.

A4 *What are the fairness objectives of the system, with respect to the individuals and groups in A2 and the harms and benefits in A3?*

The hardship targeting system should not be less effective at targeting individuals in financial hardship within the high risk groups compared to the overall customer cohort. This is equivalent to ensuring the system does not exhibit a lower recall on at risk groups.

2.4.2 Part B: examine data and models for unintentional bias

B1 *What errors, biases or properties are present in the data used by the system that may impact the system's fairness?*

Financial hardship is infrequent in the customer base, so the sample of “positive” examples is small. This means that assessing the model’s performance is difficult, and that it will also be difficult to achieve high precision (low fraction of false positives).

B2 *How are these impacts being mitigated?*

Errors in targeting that may be caused by under-representation are being mitigated by conducting broad outreach (a banner displayed in internet banking) to notify individuals of the FSI’s financial hardship services.

B3 *How does the system use AIDA models (with, or separately from, business rules and human judgement) to achieve its objectives?*

- A predictive model (gradient boosted tree classification) is used to predict whether or not an individual is experiencing financial hardship based on previous customers who have come into contact with the bank's financial hardship services. This system uses loan account transactions and repayment data, as well as other associated bank account transaction data, and demographic information.
- The system ranks individuals in order of those at risk, and these individuals are then queued for a call by the FSI's customer services team.
- An existing business rule system automatically prioritises customers who have missed more than a certain number of loan repayments.

B4 *What are the performance estimates of the AIDA models in the system?*

Recall is the primary performance indicator of the hardship targeting system. This is currently estimated to be approximately **0.77** (0.72 @ 5%, 0.84 @ 95%) on the deployment data. False negatives are estimated by the outreach approaches and by capturing subsequent missed payments.

B5 *What are the quantitative estimates of the system's performance against its business objectives?*

See above (model performance directly measures system performance as defined).

2.4.3 Part C: measure disadvantage

C1 *What are the quantitative estimates of the system's performance against its fairness objectives, assessed over the individuals and groups in A3 and the potential harms and benefits in A4?*

- Recall on females - **0.78** (0.73 @ 5%, 0.85 @ 95%)
- Recall on >55 year-olds - **0.83** (0.79 @ 5%, 0.88 @ 95%)
- Both of these are above the recall rate of the system measured over all of the deployment customers, and so is acceptable according to the fairness objectives for this system.

C2 *What are the achievable tradeoffs between the system's fairness objectives and its other objectives?*

Perfect statistical parity could be enforced for this system (which, given a fixed budget for calls, would likely lower recall).

C3 *Why are the fairness outcomes observed in the system preferable to these alternative tradeoffs?*

This operating point maximises the recall of the system given a finite budget for calls, and satisfies its fairness objectives. The marginal benefit of improving fairness here is small compared to ensuring as many individuals as possible receive assistance with potential financial hardship before it has a substantial negative impact on them.

2.4.4 Part D: justify the use of personal attributes

D1 *What personal attributes are used as part of the operation or assessment of the system?*

Age, sex, and customer financial details (e.g. debt-to-income) are used by the system, though only financial details are used as features in the predictive model.

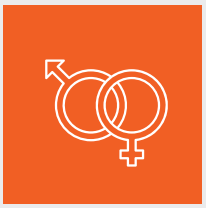
D2 *How did the process of identifying personal attributes take into account ethical objectives of the system, and the people identified as being at risk of disadvantage?*

The above attributes are all considered "personal" by internal FSI policy, which is set by the Bank's ethics committee. These attributes include those defining groups identified as at-risk of disadvantage in Response A2

D3 *For every personal attribute and potential proxy for a personal attribute, why is its inclusion justified given the system objectives, the data, and the quantified performance and fairness measures?*

age

not used by the predictive model but its use is required for the purpose of measuring fairness objectives



sex

not used by the predictive model but its use is required for the purpose of measuring fairness objectives

financial details

causally related to prediction target and system objective, and vital to the operation of the system

2.4.5 Part E: examine system monitoring and review

E1 *How is the system's monitoring and review regime designed to detect abnormal operation and unintended harms to individuals or groups?*

The performance variables are monitored weekly, with changes of >10% causing automatic escalation procedures for data scientists to investigate.

E2 *How does the system's monitoring and review regime ensure that the system's impacts are aligned with its fairness and other objectives (A1 and A4)?*

The overall performance of the FSI's financial hardship management services are reviewed annually. Part of this review involves assessing the effectiveness of the targeting system relative to other interventions.

E3 *What are the mechanisms for mitigating unintended harms to individuals or groups arising from the system's operation?*

The main objective of this system is early harm detection and prevention; maintaining or improving the system's performance will minimise harms to individuals or groups. If the predictive model is underperforming on an at-risk group (i.e. displays lower recall), then more data for this group may be captured in the future to improve model performance.

2.5 HSBC reflections on applying the Methodology

2.5.1 Use case

Evaluating fairness in AIDA models used for marketing the bank's products to fulfill customer's borrowing needs.

2.5.2 Context

HSBC proactively contacts existing credit card customers to discuss solutions for customers' borrowing needs. A selection system consisting of event triggers, business rules and machine learning (ML) models is used to prioritise leads for proactive contact.

Based on past interaction data and other signals, the selection system tries to infer customers' need for credit and the propensity for them to subscribe to the bank's lending solutions. Safeguards like credit and contact exclusions, regular model validation and ongoing performance reviews are integral to the system's development and use.

The FEAT Fairness Assessment Methodology and accompanying customer marketing case studies in this document has been used to check for unintended bias in the selection process with respect to specific protected attributes.

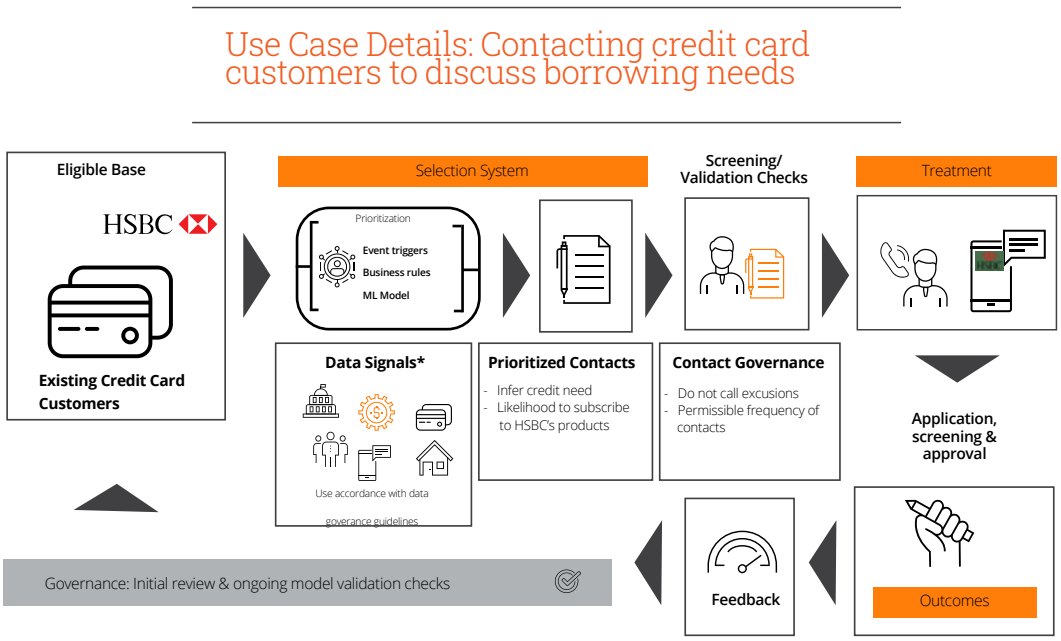


Figure 2.8: HSBC Customer marketing selection system

2.5.3 Key components

1. **AIDA selection models** which prioritise customers to be proactively contacted for a borrowing needs conversation. Specific to a borrowing needs conversation, existing credit card customers form the eligible base subject to prevalent regulatory and credit risk exclusion criteria. Three distinct types of AIDA models are used: event triggers, business rules and machine learning models. Based on recent customer behaviour and past engagement history, these models try to infer the customer's need for credit and the likelihood for them to subscribe to the bank's solutions. Use of data signals for building these three AIDA models is strictly governed by internal guidelines. Attributes that do not meet the permissible guidelines are discarded even though they may bring good discriminatory power. A defined set of criteria helps prioritise and establish hierarchy among AIDA model selections which are actioned subject to outbound direct marketing capacity (call, sms, email, etc.)
2. **Treatment** includes a proactive direct marketing outreach. In this use case, only outbound telephone conversations about borrowing needs and related personal loan solutions are considered for fairness evaluation. Customers who have applied but have not been proactively contacted (walk-in customers) are included in the study for evaluation purposes.
3. **Outcomes** relate to "approve" or "decline" decisions on a customer's loan application as per the bank's policies during the decision period. A customer conversation may result in take up of other credit products, e.g. installment plan or credit card. However such outcomes are excluded to maintain comparison rigour.
4. **Personal attributes** are crucial to the fairness evaluation exercise. The AIDA models' behaviour and the strength of any systematic bias is expressed in relation to these attributes.
    A. HSBC does not allow use of select attributes, referred to as "personal" in this exercise, for the purpose of AIDA modeling unless there are exceptional circumstances. Inclusion of any personal attribute requires the AIDA model to undergo enhanced governance: justification, independent review and a multi-member approval process.
    B. For the purpose of this exercise, fairness across two personal attributes — gender and nationality — are evaluated, although these are explicitly not used for AIDA model training or development. Data on several other personal attributes are not collected by HSBC.
5. **Exclusions:** Product or credit solution design, features including pricing, treatment scripts/messaging and credit risk criteria are out of scope for this fairness evaluation of the customer selection system. However, the combined impact of all these dimensions is included as fairness is evaluated on final outcome decisions (i.e. the "approve" or "decline" decision on loan applications

2.5.4 Fairness evaluation approach

- The selection system is evaluated as a "single" AIDA model. Evaluation is model agnostic in the sense that we focus on general diagnostics that enable systematic characterisation of treatments with respect to group fairness.
- Outcomes ("approve" or "decline" decisions) are not fully attributable to the customer selection system as other factors have an influence, for example, credit risk policies, and customer choice to not subscribe, etc. However, evaluating fairness of outcomes with respect to bank-initiated contacts helps provide a structured health check of any unintended systemic biases specific to the AIDA-based customer selection population.
- Proactive customer contact ("treatment") may lead to borrowing needs being fulfilled by appropriate credit products viz. installment plans, personal loans or credit cards. For the purpose of this study, however, only specific outcomes — successful personal loan applications — are considered to maintain comparison rigour. We call it the business impact in terms of customer needs fulfilled.
- Guidelines for evaluation:
  - Keep it **real**: use case led development
  - **Outcomes** focused: be customer centric
  - Aligned to HSBC **values**: structured health check for bias with respect to personal attributes
  - Is it **consequential**: focus on material harms and benefits for customers
  - Develop a **methodology** not a specific solution: model agnostic. Include AIDA not just AI or DA
  - Should be **scalable**: useful for system and system components
- While the use of personal attributes is restricted for AIDA modeling, this exercise goes a step forward. Evaluating fairness of outcomes helps to provide a structured health check around the degree of unintended systemic biases that might permeate AIDA based customer selection systems.
- **Note that in the following section, the answers to questions in the Methodology Parts A-E have been summarised for brevity**. Information relating to some questions is not included either because it is not relevant for the use case or because it may be proprietary.

2.5.5 Evaluation results & interpretation

Confusion matrix for overall selection system

Outcome \ treatment	No treatment (No proactive contact)	Treatment offered (Selected for proactive contact)
Not applied	31044 [TN]	12024 [FP]
Applied and approved	36 [FN]	102 [TP]

FEAT Principle F1

Individuals or groups of individuals are not systematically disadvantaged through AIDA-driven decisions, unless these decisions can be justified

Investigation highlight

Product accessibility — For comparable risk profiles, do proactively selected customers receive advantageous terms vis-à-vis walk-in customers? If so, these should be justified.

Result

Customers or prospects have access to the same product terms across multiple channels (online or manned) irrespective of inbound or outbound contact. In other words, for comparable risk profiles, there is no difference in evaluation of personal loan applications sourced via customer initiated or bank initiated contact.

Investigation highlight

Adverse selection and long term harm — Is the performance of pro-actively selected customers better or worse vis-à-vis walk-in customers over 12 months?

Result

Among approved loan customers, at end of 12 months-on- book,

- 95% of approved customers from the proactively contacted group were current and on-book versus 89% of walk-in approved customers.
- Walk-in group had 3X more attrition as compared to the system selected customer group.

FEAT Principle F2

Use of personal attributes as input factors for AIDA-driven decisions is justified

Investigation highlight

Use of personal attributes — Are personal attributes used as input factors for customer selection? And if so, is their use justified?

Result

No personal attributes are used by this selection system.

FEAT Principle F3

Data and models used for AIDA-driven decisions are regularly reviewed and validated for accuracy and relevance, and to minimise unintentional bias

Investigation highlight

Extent of bias — What is the degree of bias with respect to personal attributes (Gender, Nationality) in the customer selection system using AIDA?

Result

We used balanced accuracy and recall to evaluate outcome fairness, and similar "p% rule" for the system fairness evaluation.

Analysis for Gender:

	Mean	5%	95%
Balanced Acc. G1/G2	1.046	0.951	1.128
Recall G1/G2	0.988	0.830	1.143

- Balanced accuracy for treatment equality evaluation: 74.8% (G1) and 71.5% (G2) for the observation period.
- Recall measures what proportion of approved customer subgroup was proactively selected for borrowing needs conversation (treatment): 73.3% (G1) and 74.2% (G2). In other words, the selection system missed 26.7% (G1) and 25.8% (G2) customers for proactive calling. Profiling of these customers can help identify new data signals as inputs to strengthen selection model performance.
- Overall, disparity among treatment of gender subgroups is not significant.

Analysis for Nationality

	Mean	5%	95%
Balanced Acc. N2/N1	0.916	0.824	1.024
Recall N2/N1	1.011	0.867	1.176

- Balanced accuracy for treatment equality evaluation: 68.2% (N2), 74.5% (N1).
- Recall measures what proportion of approved customer subgroup was proactively selected for borrowing needs conversation (treatment): 74.5% (N2), 73.6% (N1). In other words, the selection system missed 25.5% (N2) and 26.4% (N1) customers for proactive calling. Profiling of these customers can help identify new data signals as inputs to strengthen selection model performance.
- Overall, disparity among treatment of nationality subgroups is not significant.

Investigation highlight

Mitigating action — Does the selection system need to be revised to minimise unintentional bias?

Result

Degree of bias is not significant. Continue periodic monitoring and review.

FEAT Principle F4

AIDA-driven decisions are regularly reviewed so that models behave as designed and intended

Investigation highlight

Selection system performance — Is the customer selection system performing as intended?

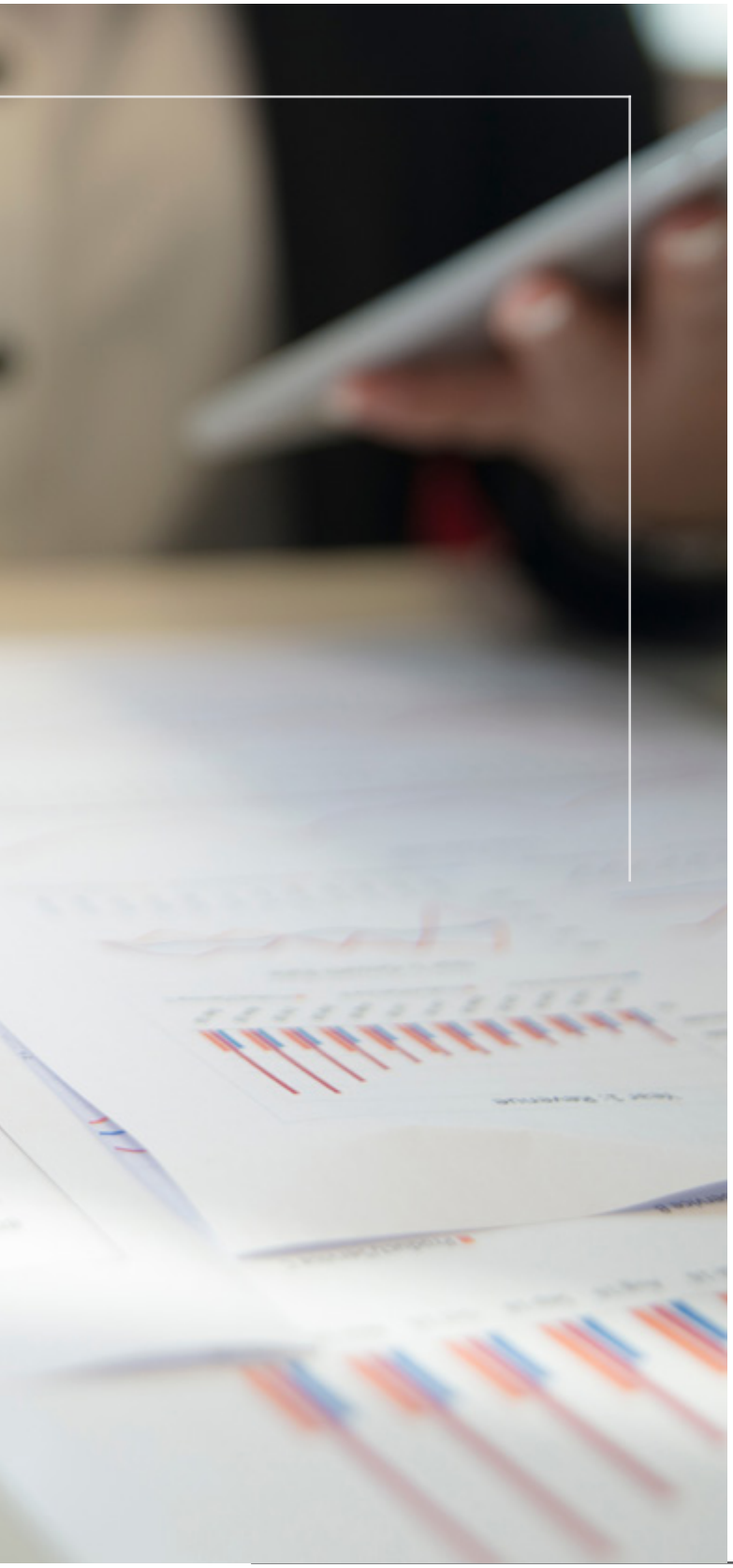
Result

We used balanced accuracy, recall and precision to evaluate outcome fairness, and similar "p% rule" for the system performance evaluation.

	Mean	5%	95%
Balanced Accuracy	0.730	0.704	0.759
Recall	0.739	0.674	0.807
Precision (outcome specific)	0.008	0.007	0.010

- Balanced accuracy, the mean of true positive and true negative rates, is at 73%.
- Recall suggests about 70+% of actual positives (approved customer applications) were identified correctly.
- Precision (0.8%) is about what proportion of proactively contacted customer population was approved. While useful, the measure does not cover the full picture as the conversation may result in a different customer need being identified and fulfilled. There is an economic tradeoff here. Although the incremental cost of outreach is relatively low and the bank wants to reach out to as many customers as possible to assist with borrowing needs, economic viability lens remains an important filter.

03 Credit Scoring



3.1 Introduction

In this case study we focus on a narrow area of the credit decisioning process, a part of the overall lending cycle in a FSI. Specifically, within the credit decisioning process, credit scoring is the focus area for analysis, as shown below.

Lending Function Overview in Banking

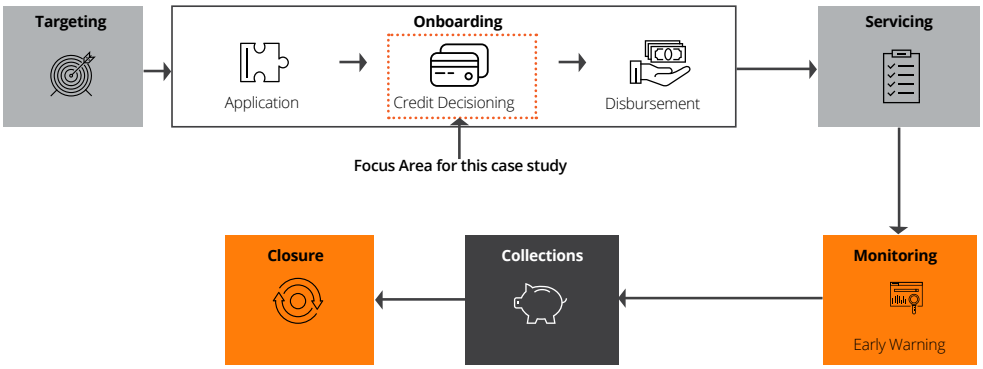


Figure 3.1: A high level overview of the end to end Lending function in banks.

A credit score is a quantification of an individual's likelihood to repay a loan. Increasingly, credit scores around the world are used to predict more than credit risk: they are used to determine access to housing, utilities, insurance premiums, and even employment or social standing [32]. A high score can open doors for consumers, possibly bringing them out of poverty or lower-income status. A low score can very quickly do the opposite and takes years, usually a decade, to rebuild [29]. Since a credit score can affect an individual's opportunities throughout their life, it is ever more important to get it right.

Traditional credit scoring models rely heavily on external credit bureau information and previous payment behaviour to predict the future likelihood to pay for consumers. Further traditional data inputs include socio-economic factors for the applicant, public records, and in case of existing customers payment histories and credit utilisation within the FSI. Credit scores do not predict individual defaults: they aim at positioning individuals amongst pools of people that are expected to exhibit a measurable default rate.

The importance of credit bureau information for traditional credit scoring models can create a major obstacle for access to financing. FSIs typically limit loan approvals to consumer segments with credit scores beyond their own acceptance thresholds based on their respective risk appetite. Certain consumer segments have no credit file and consequently no credit bureau score ("no-file"), as they have simply not made use of loans and/or have no payment/bank history in the relevant jurisdiction (i.e. students or recent immigrants). Other consumers have a credit file, but the information it contains is insufficient to derive a credit bureau score, either because they have not used credit in recent history

(e.g. older people, pensioners), or have made very limited use of credit ("thin-file"). Both segments, as well as consumer-segments with a credit bureau score below the acceptance criteria of banks, are likely to have restricted access to credit. These consumer groups are specifically vulnerable to high-priced credit alternatives (e.g. payday loans) and / or severely limited in their access to credit as a means of personal or professional growth. Whilst credit bureaus themselves are pursuing opportunities to improve their market coverage, banks have been pursuing similar endeavours.

Credit scores are combined with business rules (such as eligibility criteria and lending policies) to produce credit approval/denial decisions. Credit approval processes in the consumer banking space have evolved towards straight-through-processing over recent years and are largely processed automatically, with some manual interventions for marginal cases. Although we follow the literature in referring to credit scoring below, more precisely the AIDA system of interest is the one that produces credit approvals, as it is the approve-or-deny decisions (not the scores themselves) that produce benefits and harms.

As a testbed use case for the FEAT Principles, credit scoring is of particular interest to FSIs because it emcompases a generally data-intensive practice with the potential to access a large number of "underbanked" consumers. These consumers are underbanked primarily because they have not generated data that are traditionally used in developing credit scores, such as banking history or loan repayments. In the past, given little information, FSIs might have passed on these consumers for whom it would be difficult to assess credit risk. With the advent of new data and modeling techniques, FSIs are reconsidering how they assess credit risk for these consumers, possibly increasing the lending market. Incorporating non-traditional data, modeling techniques, and consumers is not without its own set of risks: the Methodology presented in Document 1 represents a first step to address them.

Resources for assessing AIDA credit scoring systems

Building on the considerations presented in the Document 1 Section 3, the next section provides additional considerations specific to credit scoring use cases.

Following the considerations, Section 3.3 presents a case study of a FEAT fairness assessment conducted on credit scoring (loan approval) system. This case study is designed to illustrate the application of the Methodology and provide practitioners conducting assessments with concrete examples. The system is considered "higher risk" due to the significant impact that credit systems can have on individuals and groups. It is analysed at a high level of detail.

Finally, Section 3.4 presents reflections from UOB, an FSI that applied the Methodology on one of their credit scoring models. The aim of these reflections is to help practitioners identify some of the practical challenges FSIs may face conducting assessments, and suggest approaches to overcome them.

3.2 Methodology considerations for credit scoring

This section provides considerations specific to credit scoring use cases for each part of the Methodology.

3.2.1 Part A: describe system objectives and context

Many uses of credit scores

While a FSI might primarily be interested in an individual's credit score so as to determine that individual's likelihood to repay a loan, there are other uses for credit scores. Credit scores are sometimes used to determine access to housing, utilities, insurance premiums, and even employment or social standing [32] even though what a credit scores measures (a person's "likelihood to repay") may be very different from the variable of interest in other domains (such as the person's "likelihood to be a good renter"). For "underbanked" or "unbanked" individuals, the group in focus for this use case, a credit score can be harder to determine since fewer of the traditional data sources, such as loan repayment history, are available. When coupled with the fact that credit scores can be used as proxies for assessments beyond credit risk, the underbanked population quickly becomes one of the most at-risk groups for unfair credit scoring. Being a good credit risk can be synonymous with opportunity. Conversely, a low score can very quickly eliminate access to opportunities and can take years to turn around.

These reflections may help practitioners consider the positive and negative impacts of a credit scoring system:

- What else is the credit score being used for within the FSI? Does the FSI share its credit score or related insights with external parties?
- Are there different types of impacts the system could have on an individual, such as financial, reputational, social or emotional?
- What are the customers' views about the benefits or risks or the system, and are there any potential harms to vulnerable or at-risk customers that may have been missed by any consultation processes undertaken?
- What steps were taken to ensure this list does not omit important entries (for example, through customer and expert consultation)?

Including human judgments within the system boundaries

Credit scoring models and other automated decision tools can limit the potential for credit applicants to be treated differently on an unjustified basis by reducing the amount of discretion in credit decisions. The specific model is likely embedded within a process that contains numerous other rules (e.g. eligibility criteria) as well as potential interventions from human actors which affect the FSI's decisions and therefore the outcomes (harms and benefits). In the U.S., credit processes that do not meet the standards for being "empirically

derived, demonstrably and statistically sound" are considered to be "judgmental" systems, and are afforded less favourable presumptions when faced with discrimination claims [9]. A process view of the typical role of credit risk modelling within the credit AIDA system is below:

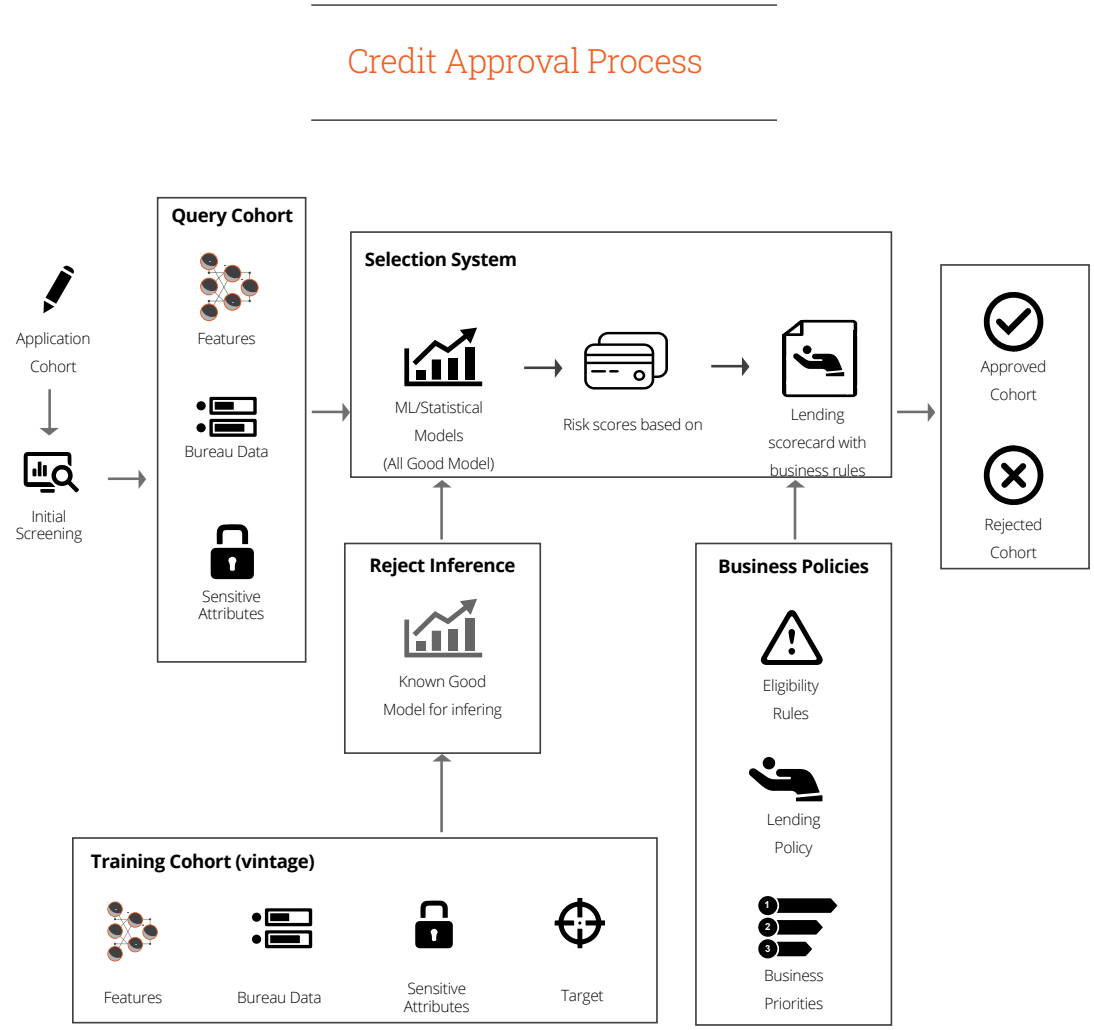


Figure 3.2: A process view of the typical role of credit risk modelling within the credit AIDA system.

While the "empirically derived, demonstrably and statistically sound" requirement is derived from a U.S. law (the Equal Credit Opportunity Act — ECOA) that does not have an equivalent in Singapore, it is still important to consider the broader system in which the statistical model is embedded. These reflections may help practitioners develop their responses to the questions in this section:

- Will initial screening, eligibility rules, and lending policies also be assessed for fairness in this exercise?
- If the outputs of AIDA systems are used by people who make final decisions related to credit, how are the human aspects of the overall system being audited for fairness?
- Are marginal approvals and marginal denials compared to assess the impact of discretion?

3.2.2 Part B: examine data and models for unintended bias

Reject inference

In credit scoring, an issue of sample bias is well-studied and well-known, hence it warrants its own specific discussion. When a model is built on a sample of individuals that is not representative of the individuals who apply for credit, there is sample bias [31]. In this use case, the data skew in favour of previously-accepted applicants because FSIs are unable to observe cases of previously-rejected applicants: FSIs are unable to observe the outcome (target variable), being whether or not the applicant was able to repay their debt. This information on outcome is required in the training set of any supervised predictive model since the goal is to predict which applicants are good credit risks. Thus, sample bias stems from estimating the default probabilities for all future credit applicants using a model trained on a skewed sample, containing only the previously accepted applicants.

Reject inference techniques are designed to minimise the effects of sample bias in model training, such as population drainage or biased estimates, by also taking account of data in rejected applications [31]. It is common practice to use reject inference techniques to impute the target variable (the hypothetical loan outcome) for rejected cases, allowing some portion of the reject cohort to be used when building subsequent models. However, this imputation introduces additional uncertainty. The unobserved target variables (those for the reject cohort) are typically imputed from information about the accepted cohort, for which the outcomes are known. As mentioned by [14], the reject cohort has been so designated precisely because it differs in a non-trivial way from the accepted cohort. Thus, systematic errors (biases) in the imputation are likely. Furthermore, the quality of the imputation is very difficult to evaluate because no ground truth exists. It is impossible to definitively know the outcome of the counterfactual: what would have happened if the applicant had been accepted?

Sample bias and reject inference pose a problem not only for measuring the performance

of the model but also for assessing fairness, as many fairness metrics rely on the target variables for the rejected cohort. If the rejected targets are unavailable, these metrics cannot be computed. If the rejected targets have been imputed, then the imputation process and the errors it introduces can add considerable uncertainty to the fairness metrics which rely on them. Depending on the reject inference technique, this uncertainty may be difficult to quantify. The affected fairness metrics should be interpreted with this in mind.

These reflections may help practitioners develop their responses to the questions in this section:

- If the data from previously rejected applicants are not used to train the model, how might the sample bias affect model performance (e.g. drainage, bias)?
- If a reject inference technique has been used, what uncertainty does it introduce?
- Can the uncertainty introduced by reject reference be quantified?
- If reject inference techniques have not been used, how are the issues (such as sample bias) being addressed? Why is this approach superior and what are the justifications?

Performance measures

Choosing a performance measure that aligns well with the business and fairness objectives is an important part of effectively evaluating the system. The choice of which measure is appropriate will depend on the particular system and the FSI's objectives. For a credit scoring system, the underlying model used for assessing the creditworthiness of the applicant produces risk score bands and banks use cut-off thresholds to generate a binary outcome. Credit scoring is well-established and the typical measures used to judge the effectiveness of credit scoring models include, but are not limited to: Gini Coefficient or Accuracy Ratios (AR), Kolmogorov-Smirnov (KS) statistics, Receiver Operating Characteristic Curve (ROC), Pietra Index, and confusion matrices. It is not enough to simply measure performance, since there must be some regard for the impact on fairness. We discuss this further in Section 3.3.4 where we compare the influence of features on accuracy versus fairness objectives, as part of the justification for using personal attributes.

3.2.3 Part C: measure disadvantage bias

Definition of harms and benefits

To understand whether a system's decisions systematically disadvantage individuals or groups (FEAT Fairness Principle F1) it is necessary to understand the potential harms and benefits the system causes. Fairness in the operation of the system can then be assessed by examining how the system distributes these harms and benefits. One approach is to consider the harms and benefits of a credit product in terms of the outcomes it creates for individuals and groups, compared to the baseline of that product not being made available to anyone. Similarly, one way to justify the inclusion of personal attributes in modelling (FEAT Fairness Principle F2), is to show that such inclusion creates more benefits, fewer harms, or distributes these more equally.

Every real system will have unique harms and benefits. These will depend on the product or service, the audience, and the timing, location and context amongst many other factors. Undertaking careful consultation with customers and impacted individuals and groups will help FSIs understand the potential harms and benefits of their particular system. If a specific credit product was not made available by the FSI, no individuals or group could benefit from access to the loan it provides. Moreover, no individual or group could suffer from the social and financial consequences of defaulting on that loan. Thus, one can think of a credit product as producing a benefit for those that gain access, and a harm to those that default.

Through this lens, the denial of credit to an individual or group can be thought of as a harm stemming from a lack of access to a benefit. Similarly, not defaulting on a loan can be thought of as a benefit stemming from an avoidance of harm. Thus, when exploring the counterfactual where individuals in the reject cohort are instead approved, those that would have resolved their loans are harmed from a lack of access to credit, while those that would have defaulted on their loans are harmed from the lack of access to credit but also benefit from having avoided default.

Quantifying harms and benefits

Quantifying, or even just comparing, these harms and benefits requires careful thought, as they may vary in magnitude between individuals or groups. For example, a small loan may be of great value to an individual in dire need of cash, but of little value to a wealthy individual with plenty of liquidity. Similarly, an entrepreneur may view defaulting on a loan as a very acceptable risk, while a more conservative individual may view it as a very negative outcome, perhaps because they or their community view it as a source of shame. However, most of the algorithmic bias literature treats the magnitude of the harms and benefits as invariant between individuals, since not doing so complicates the analysis considerably. To fully realise an analysis of the harms and benefits of a system, conducting careful consultation with customers and domain experts is likely required.

It is also unclear whether the benefit of gaining access to credit can be put in the same "units" as the harm stemming from a default to make meaningful comparisons. For instance: if group *A* has one extra default and one extra successful loan compared to another group *B*, should this be considered to balance out so that groups *A* and *B* have equal outcomes? If it is deemed they can be meaningfully compared, there is the further question of how they weigh in comparison to one another. To continue the previous example, are the groups' outcomes still equal if the loan and the default are for different amounts? And for an individual, is the harm of default worse than the benefit of credit?

It is possible for different stakeholders and different FSIs to have different opinions about these questions. When considering the harms and benefits it may be helpful to overlay them on the system's confusion matrix. A simple example of this is illustrated in the table below:

Outcome / decision	**Approve**	**Reject** *(hypothetical outcomes)*
Resolves ("good" loan)	True Positive (TP) Benefit: access to credit	False Negative (FN) (Lack of benefit)
Defaults ("bad" loan)	False Positive (FP) Benefit: access to credit Harm: the social and financial consequences of default	True Negative (TN) (Lack of benefit and lack of harm)

It should also be noted that the loan outcome (resolves or defaults) is not an inherent attribute of the applicant, but rather an outcome of a stochastic process that is only partially controlled by the applicant. Many random factors such as macro and microeconomics conditions, and unforeseeable personal circumstances will impact a loan outcome. Even the most fiscally responsible individuals may encounter insurmountable external events that lead them to default on a loan. This should be kept in mind when interpreting the confusion matrix.

Choosing among fairness criteria

Independence, separation and sufficiency are broad theoretical fairness criteria. In the algorithmic bias literature, many measures of fairness have been proposed which aim to uphold whichever criteria the authors believe to be most relevant.

This is done either directly, or with some added variation or relaxation. Each theoretical criteria makes either implicit or explicit assumptions about the harms and benefits of the system. They share a common assumption that the outcome is always either positive (such as getting into a school) or negative (such as being denied bail). However as discussed above, in the case of credit scoring, receiving a loan may be considered positive or negative depending on whether the applicant is likely to repay it, and possibly also other subjective factors.

These criteria vary in their implicit reliance on reject inference. As discussed in Section 3.3.2 above, in the context of credit scoring, one of the key considerations is the reliance on reject inference to impute the outcomes (target variables) for the reject cohort. Here we present a set of popular group fairness measures that are relevant to the credit scoring use case. The table below discusses their "pros and cons" in the context of harms and benefits as well as their reliance on reject inference. Importantly, it is impossible to build a system that meaningfully satisfies all of these fairness measures. Therefore it is important to choose one or more measures. Furthermore, this choice should be made before the analysis is conducted, so that the system's performance is evaluated against the fairness metrics that are most relevant, rather than justifying ex post the choice of measures that show the least unfairness.

A note on notation: In the tables below we use *A* to refer to the personal attribute(s) that define group membership. We use *Y* to refer to the target variable, where *Y=1* indicates that a loan is resolved, and *Y=0* indicates that a loan ends in default. We use *R* to indicate the lending decision, where *R=1* indicates that an applicant is approved, and *R=0* indicates that an applicant is rejected.

Fairness measure	Criteria for fairness	Probabilistic view	Depends on reject inference	Discussion
Demographic parity *(Independence)*	Equal approval rates between groups (equivalently, equal rejection rates)	*P(R)* independent of *A*	No	**Pros:** This criterion is both simple and independent of reject inference. It captures the benefits of gaining access to credit. **Cons:** In its basic form, it does not in any way account for differences in base default rates between groups, thus it may only allow for solutions with considerably lower utility to the FSI. Moreover, by only considering approvals, it lumps true positives and false positives together, which ignores the harms of defaulting.
Equal opportunity *(relaxed separation)*	Equal true positive rates between groups (equivalently, equal false negative rates)	*P(R)* independent of *A* given *Y=1*	Yes	**Pros:** By requiring the true positive rates (and by consequence false negative rates) to be equal between groups, it ensures that an equal fraction, in each group, of the applicants who would repay their loans will receive loans, and thus benefit. It also tends to allow for solutions with greater utility to the FSI, because it accounts for differences in base rates. **Cons:** It is dependent on reject inference, and it does not capture all the harms and benefits (those from FPs and TNs).
False positive rate balance *(relaxed separation)*	Equal false positive rates between groups (equivalently equal true negative rates)	*P(R)* independent of *A* given *Y=0*	Yes	**Pros**: By requiring the false positive rates (and by consequence true negative rates) to be equal between groups, it ensures an equal fraction, in each group, of the applicants who would default their loans be denied loans, and thus avoid harm. Like equal opportunity, it also tends to allow for solutions with greater utility to the FSI, because it accounts for differences in the base rates between groups. **Cons:** It also depends on reject inference, and it also does not capture all the harms and benefits (those from TPs and FNs).
Equalised odds or average odds *(separation)*	Equal true positive rates between groups AND	*P(R)* independent of *A* given *Y*	Yes	Equalised odds is really two criteria, requiring both equal opportunity and false positive rate balance. **Pros:** It upholds the criteria of both equal opportunity and false positive rate balance and allows for the consideration of base default rates.

Fairness measure	Criteria for fairness	Probabilistic view	Depends on reject inference	Discussion
	Equal false positive rates between groups (equivalently, equal false negative rates AND equal true negative rates)			**Cons:** It is stricter than either equal opportunity or false positive rate parity alone, and thus may allow for lesser utility for the FSI. Reject inference is required. Post-processing a model to uphold equalised odds may require randomisation. Quantifying deviance from the equalised odds criteria as a scalar quantity requires that the harms and benefits from the TPs and FNs be comparable to those from the FPs and TNs (i.e. the units must be matched).
Positive Predictive parity *(relaxed sufficiency)*	Equal positive predictive value (precision) between groups (equivalently equal false discovery rates)	*P(Y)* independent of *A* given *R=1*	No	**Pros:** This criterion is appealing because it does not rely on reject inference (imputed target variables). **Cons:** On its own, it does do much to measure the distribution of harms and benefits. It simply compares the fraction of true positives to the fraction of approvals in each group.
False omission rate balance (*relaxed sufficiency)*	Equal false omission rates between groups (equivalently equal negative predictive value)	*P(Y)* independent of *A* given *R=0*	Yes	**Cons:** This criterion relies entirely on reject inference and does not align well with measuring the distribution of harms and benefits. It simply compares the fraction of false negatives to the fraction of rejections.
calibration by group *(sufficiency)*	Equal positive predictive value (precision) between groups AND equal false omission rates between groups (equivalently equal false discovery rates AND equal negative predictive value)	*P(Y)* independent of *A* given *R*	Yes	Calibration by group is two criteria, requiring both predictive parity and false omission rate balance. **Pros:** These criteria ensure that the model is behaving similarly for each group. Effectively it ensures that a recommendation to accept or reject an applicant carries the same predictive value in each group. **Cons:** It relies on reject inference and what it measures is somewhat orthogonal to the distribution of harms and benefits in this use case.

To help in understanding these metrics, and to illustrate how they are computed in practice, we present two tables of sample calculations below. The first summarises a simple credit scoring model, while the second shows how the metrics can be computed from the outcomes. We exemplify the analysis by using binary gender as a way to create two groups.

System Variables	Equation	Group a (A=Men)	Group b (A=Women)
These are taken as the input variables for the subsequent analysis			
M: number of applicants	(observed) variable	14000	8000
PPV: Positive Predictive Value (model precision)	(observed) variable**	95%	94%
L: Lending (acceptance) rate	(observed) variable**	83%	84%
** *Interdependent functions of the model's predictive performance and the risk threshold chosen*			
Observed Outcomes *These counts and rates are directly observed, no reject inference is required.*			
Counts			
Number of accepted applicants: R=1	M x L	11620	6720
Number of rejected applicants: R=0	M x (1 - L)	2380	1280
TP: True Positives: R=1, Y=1	M x L x PPV	11039	6317
FP: False Positives: R=1, Y=0	M x L x (1 - PPV)	581	403
Rates ~ Probability			
PPV: Positive Predictive Value: P(Y=1|R=1)	TP / (TP + FP)	95.0%	94.0%
FDR: False Discovery Rate: P(Y=0|R=1)	FP / (TP + FP)	5.0%	6.0%

System Variables	Equation	Group a (A=Men)	Group b (A=Women)
L: Lending (acceptance) rate: P(R=1)	(TP + FP) / M	83.0%	84.0%
Group default rate: P(Y=0, R=1)	FP / M = L x FDR	4.2%	5.0%
Imputed Outcomes *These counts and rates require some form of reject inference. Here they are obtained by assuming the base default rate, B, under the hypothetical setting of 100% loan acceptance.*			
B: Base default rate (in the group): P(Y=0)	(imputed) variable	10%	8%
Counts			
P: Number of positives (would resolve): Y=1	M x (1 - B) = TP + FN	12600	7360
N: Number of negatives (would default): Y=0	M x B = TN + FP	1400	640
TN: True Negatives: R=0, Y=0	N - FP	819	237
FN: False Negatives: R=0, Y=1	P - TP	1561	1043
Rates ~ Probability			
TPR: True Positive Rate: P(R=1|Y=1)	TP / P	87.6%	85.8%
FPR: False Positive Rate: P(R=1|Y=0)	FP / N	41.5%	63.0%
TNR: True Negative Rate: P(R=0|Y=0)	TN / N	58.5%	37.0%
FNR: False Negative Rate: P(R=0|Y=1)	FN / P	12.4%	14.2%
FOR: False Omission Rate: P(Y=1|R=0)	FN / (TN + FN)	65.6%	81.5%

With the above tabulated AIDA system, we can compute the aforementioned metrics as shown in the table below. We denote value X computed on group a as X_a and on group b as X_b. We use the symbol ⊥ to mean statistical independence.

Fairness Metric	Criteria	Eq. for deviance (as a rate difference)	Value
Demographic Parity	$P(R) \perp A$	$L_a - L_b$	-1.0%
Equal Opportunity	$P(R \mid Y=1) \perp A$	$TPR_a - TPR_b$	1.8%
False Positive Rate Balance	$P(R \mid Y=0) \perp A$	$FPR_a - FPR_b$	-21.5%
Average Odds	$P(R \mid Y) \perp A$	$((TPR_a - TPR_b) + (FPR_a - FPR_b)) / 2$	-9.9%
Positive Predictive Parity	$P(Y \mid R=1) \perp A$	$PPV_a - PPV_b$	1.0%
False Omission Rate Balance	$P(Y \mid R=0) \perp A$	$FOR_a - FOR_b$	-15.9%
(Average) Calibration by Group	$P(Y \mid R) \perp A$	$((PPV_a - PPV_b) + (FOR_a - FOR_b)) / 2$	-7.5%

With the metrics presented above in mind, we highlight two approaches that could be taken to choose fairness criteria for credit scoring.

Separated equalised odds: If reject inference is carried out in a way that reliably imputes the outcomes (target variables) for the reject cohort, then monitoring both equal opportunity and false positive rate balance (separately) is a good choice.

Equal opportunity focuses on the fraction of applicants who would repay their loan that receive loans, capturing the benefits of the true positives (TPs) and the harms (from denial of benefits) of the false negatives (FNs), which arguably produce the clearest benefits and harms. This also puts the focus on the applicants that the FSI are targeting for the credit product, and incentivises the FSI to learn how to identify "good" applicants equally well in all groups.

Assessing false positive rate balance will consider the harms and benefits from false positives (FPs). It can be given weight in the fairness analysis according to a decision of how the benefits of gaining access to credit compare to the harms of default. If these are viewed as equal (and opposite) the false positive rate balance can be ignored. If the harms are viewed as being greater than the benefits, it can be monitored more closely to ensure that one group is not suffering from an abundance of defaults.

Demographic parity and positive predictive parity: If reject inference is not available or not reliable, then monitoring demographic parity (or some base default rate adjusted variant) along with positive predictive parity is a good choice.

Demographic parity exactly aligns with the benefits of gaining access to credit (acceptance rates), while positive predictive parity ensures that an equal fraction of those accepted in

each group resolve their loans, thus capturing the harms. If this approach severely limits utility for the FSI (because base default rates vary greatly between groups), rather than requiring exact demographic parity, the FSI may opt for a base default rate adjusted variant. For example, rather than requiring that an equal fraction of group *a* and group *b* are accepted, they may allow the fraction to deviate from equality by some justifiable amount.

These reflections may help practitioners develop their responses to the questions in this section:

- Can reject inference be trusted to reliably impute outcomes for the reject cohort equally well for all groups? If not, has the chosen fairness criteria taken this into account?
- Do customers consider the benefits of gaining access to credit as equal (but opposite) from the harms of defaulting? Do the communities at risk of systematic disadvantage agree? If not, then is false positive rate balance also being monitored?
- How does the choice of fairness criteria incentivise the FSI (or not) to learn about different groups?

Volume versus magnitude of unfairness

All of the fairness measures above are traditionally calculated by looking at the relative incidence of outcomes (such as false positive and false negatives) between groups. While this makes sense in contexts where the outcomes are essentially unitary (such as a hiring or not hiring an applicant), fairness analysis in credit scoring can be more complex because the amount of loans and the loan terms may also vary. For instance, groups *a* and *b* might have identical rates of false negatives (incorrect declines), but most of the errors for group *a* could be for high-value loans, and most of the errors for group *b* could be for low-value loans. As a result, the total value of credit incorrectly denied to group *b* would be much higher than for group *a*.

The table below shows this type of result for the Open credit dataset example referenced in the sample answers:

Group	True Positives	False Positives	TPR	FPR
Male	No: 8546 Total: USD $66,907,470.30 Avg: USD $8,156.46	No: 448 Total: USD $3,294,388.81 Avg: USD $7,825.15	0.45	0.21
Female	No: 22550 Total: USD $179,061,164.94 Avg: USD $8,119.58	No: 878 Total: USD $7,151,642.30 Avg: USD $8,433.54	0.60	0.31

This table suggests that while men are disadvantaged in terms of volume, women may be slightly disadvantaged in terms of value (because the average value of a loan incorrectly provided to men is slightly higher). An actual example of type of finding from another jurisdiction is a study that found lenders in the U.S. "charged otherwise-equivalent Latinx/African-American borrowers 7.9 (3.6) bps higher rates for purchase (refinance) mortgages, costing $765M yearly" [3]. This type of price discrimination would not have been unocvered by looking only at the rate of approvals.

A full analysis should consider incorporating the cost of false positives and false negatives to weight the optimal tradeoffs between them, both for business and fairness-related objectives. In the credit scoring context, because credit is sometimes a first banking product, the cost of a false negative could include the total Lifetime value of the new customer, including other future products (discounted by the likelihood of later cross-selling). The cost of a false positive could take into account the FSI's effectiveness at collections. These may introduce additional fairness considerations. For instance, if analysis reveals that an FSI is more effective at collecting from one group, then the expected loss from defaults by applicants from that group would go down, which in turn might lead to increased lending to that group. This could either improve or worsen fairness metrics.

These reflections may help practitioners develop their responses to the questions in this section:

- Is fairness being measured in terms of incidence (volume) or value or some other quantity? Why? How does this choice relate to the overall business and fairness objectives?
- Are the same groups disadvantaged when applying an incidence lens versus a value lens?
- If there is disadvantage, have experts or representatives from the community been consulted on which lens is most appropriate?

Feedback and long term impacts on fairness

It is also important to consider what constitutes harms and benefits at the group level in terms of a community rather than just aggregates of individuals. People in some communities may distrust financial institutions and be less likely to even apply for loans or other financial products. This may be because they have been repeatedly rejected in the past. The undersupply of credit to these underbanked people would not even show up in an analysis of harms and benefits that looks at outcomes after an application is submitted (because it would not consider the individuals that did not apply).

If an individual from a community (group) that has historically been systematically undersupplied with loans is given a loan, even if that person defaults, the community as a whole may benefit. This may be especially true if the group has lacked essential credit for its local businesses or to help its members learn new skills. The historic undersupply of loans may have several underlying causes, for example: a history of discriminatory lending policies,

FSIs having struggled to reliably model risk for members from this group (possibly because of a lack of data), or because community-specific economic conditions have led to increased default rates. In this setting, one strategy to mitigate unfairness may be to use different lending risk thresholds for different groups. Lowering the threshold for a group will lead to more loans for individuals in that group, but unless an extra effort is made to ensure that these loans are provided to individuals with a high expected ability to pay, the result of lower thresholds may principally be more defaults. In the short term, the additional credit could be helpful, but in the long term, a poor execution of this strategy could impact the community negatively. It may lead to unmanaged debt and higher credit rejection rates over time. This feedback is illustrated in the chart below:

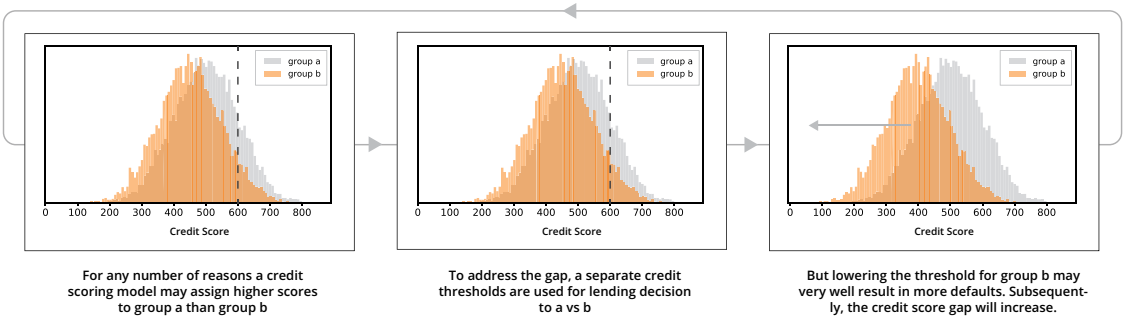


Figure 3.3: Possible long term impact of a simple fairness remediation technique

These reflections may help practitioners develop their responses to the questions in this section:

- Can the definition of harms and benefits capture upstream issues (such as a lack of applicants from some group)? (Additional questions to self-assess potential pre-application screening can be found here [13])
- Can the definition of harms and benefits capture downstream issues (such as the dynamics of credit thresholds)?

Uncertainty in model performance

It is not a common practice in credit scoring to measure uncertainty in model performance metrics such as AR or KS statistics. In consumer portfolios, there are typically a large number of data points available, and a sufficient number of Bads (instances of default). Hence point estimates of performance tend to be robust. For other non-retail portfolios, this may not be the case due to the low number of Bads. For such models, in the event that these measures may not be sufficiently robust, alternative assessments (e.g. comparing model score against other benchmark ratings) may be used to to justify the separation power, rather than attempting to measure uncertainty around AR and KS statistics. The benchmark ratings could refer to external rating agencies' ratings, or could be in the form of a "blind" rating, by getting credit approvers to rate the customers without seeing the model's ratings.

3.2.4 Part D: justify the use of personal attributes

No credit scoring specific considerations.

3.2.5 Part E: examine system monitoring and review

Override analysis

An override occurs whenever the model output has been ignored or amended, and varies significantly by the nature of the portfolio or business. Overrides or exceptions can create fair lending risk by causing similarly qualified applicants to be treated differently. To the extent that judgmental score overrides are allowed within an FSI, it is important that there are clear guidelines regarding the allowable reasons for overrides, documentation of the reasons for granting an override, and monitoring of the volume or frequency of exceptions remains within an acceptable range. This applies to both overrides "approve"and "decline" decisions. Unsecured retail lending models should generate a low number of model overrides, while significant model overrides are more the norm for secured retail or wholesale lending models. Since overrides are part of the overall credit process system, it is important that they be monitored as part of ensuring the system continues to work as designed and intended.

These reflections may help practitioners develop their responses to the questions about override analysis:

- Is the rate of overrides monitored? How is this monitoring and review integrated with model-level monitoring and review to provide an effective check on the overall system?
- Can the documentation of override rationale be queried systematically and analytically? If not, what controls are in place to assure that there is not systematic unfairness being introduced through override discretion?

Model performance monitoring with Population Stability Index (PSI)

For banks, loans are not only assets — as they are income producing — but also liabilities when customers default and do not repay their debt. In many jurisdictions, these liabilities are measured by procedures in regulations such as the Basel Accord [5] for capital and the International Financial Reporting Standards (IFRS 9) for provisioning [17]. Capital is required in case of a severe economic downturn, while provisions reflect losses expected in current economic conditions. Stability and performance (i.e. prediction accuracy) are extremely important as they provide information about the quality of the scoring models. As such, they should be tracked and analysed at least on a monthly basis by banks, regardless of the validation exercise.

Population stability refers to whether the characteristics of the portfolio (especially the distribution of explanatory variables) is changing over time. When this distribution changes (low population stability) there is more concern over whether the model is currently fit-for-

purpose since the data used to develop the model differs from the data the model is being applied to. Applying the model to these new types of customers might involve extrapolation and hence lower confidence in model outputs.

There are other characteristics of a model that requires monitoring to ensure the model is fit-for-purpose. These include calibration (whether the model is unbiased), discrimination (whether the model correctly orders the loans from best to worst) and fairness measure described above. While these measures are important, they require known outcomes. For example, a probability of default model predicting defaults in a one year window must evaluate loans at least one year old to determine calibration and discrimination. Therefore, conclusions from these measures are at least one year out of date compared to the current portfolio.

Population stability is important as it requires no lag; it can be measured with the current portfolio since the outcome is not required. Therefore, it is important to monitor population stability to gain insights concerning whether the current portfolio (rather than the portfolio one year ago) is fit-for-purpose.

The PSI is closely related to well-established entropy measures, and essentially is a symmetric measure of the difference between two statistical distributions [18]. PSI is used to either monitor overall population score stability ("System stability report") or, as a likely follow-up, the stability of individual explanatory variables ("Characteristic analysis report") in credit risk modelling scorecards for the banking industry [30]. The same formulation has appeared in the statistical literature as the "*J* divergence" [21].

The formula for the PSI assumes there are K mutually exclusive categories, numbered 1 to K, with:

$$\text{PSI} = \sum_{i=1}^{K} (O_i - E_i) \times \ln\left(\frac{O_i}{E_i}\right)$$

where O_i is the observed relative frequency of accounts in category i at review; E_i is the relative frequency of accounts in category i at development (the review relative frequency is expected to be similar to the development relative frequency); i is the category, taking values from 1 to K; and $ln()$ is the natural logarithm.

These reflections may help practitioners develop their responses to the questions in this section:

- Are there leading (not just lagging) indicators of the model's performance from the current loan portfolio?
- Are fit-for-purpose assessments using techniques that are appropriate for machine learning based models (if these are components of the AIDA system)?

3.3 Open credit dataset case study

This section is a case study assessment conducted on an open dataset. **Note that the sample answers are illustrative and do not represent the actual AIDA systems in place at any of the Consortium members.** To provide illustrations of potential answers, the example uses an open data set and built simplified credit scoring models. The findings of systematic disadvantage are illustrative only of this toy model and dataset, not the operations of any consortium members. For the same reason, the sample should not be considered a complete answer. For instance, the case study identifies, for the sake of illustration, only two personal attributes (GENDER and MARITAL STATUS). **This choice should not be considered guidance that these attributes are a necessary or sufficient scope for group fairness analysis.** Similarly, the case study illustrates individual fairness based on risk score as a measure of similarity, whereas other measures could be considered. **The terms "we", "us", or "our" below refer to the hypothetical author of this assessment and not the members of the Consortium as elsewhere in this document.**

The code to run some this analysis can be found in the following GitHub repo https://github.com/veritas-project/phase1/

3.3.1 Part A: describe system objectives and context

A1 What are the business objectives of the system and how is AIDA used to achieve these objectives?

The business objective of this credit approval AIDA system is to provide unsecured loans to all eligible customers who have a sufficiently low credit default risk based on the risk appetite, policies, and business strategy of the FSI loan provider (hereafter referred to as "we", "our", or "us").

This model is focused on serving "underbanked" individuals who have little to no credit history (also known as "thin-file" or "no-files" since their credit history file is limited). These types of individuals represent new potential customers for us, and are a priority growth segment to promote financial inclusion.

As a regulated financial institution we are subject to capital and liquidity requirements that may constrain our lending activities as losses on loans may ultimately affect our capital.

Neither capital and liquidity requirements, nor fair dealing and consumer banking practices create constraints that are specific to credit scoring. To the degree that the underlying principles are applicable to any consumer banking product, we consider them as such for this document.

The Guidelines on Fair Dealing [23] and the Code of Consumer Banking Practice [1] are the two primary documents informing interactions with consumers. The Fair Dealing Guidelines set out five fair dealing outcomes and leave the decision on their application to each FSI, per their business model and customer base. The Code of Consumer Banking was developed by the Association of Banks in Singapore (ABS) and is premised on five principles for consumer engagement: Accountability, Fairness, Privacy, Reliability and Transparency.

Internally, there are Credit Scorecards that codify the policy, eligibility, business priorities along with the credit risk to help make a credit approval decision. When an applicant is denied, reasons for this can include, but are not limited to, one of the bank's policies or business objectives not being met, or the individual's credit score being lower than the bank's cut-off threshold.

AIDA models are used to predict the likelihood of default for a customer. Default is defined as being more than 60 days late in payment within 18 months of the loan being disbursed.

The statistical prediction is combined with business rules (such as eligibility criteria or lending policies) to determine whether the application should be approved or denied. For instance, an applicant might have a very low risk of default but nonetheless be declined because they are ineligible based on their citizenship. These rules and policies are encoded in the form of scorecards that are applied automatically unless the risk score is near a cut-off. For marginal cases, the application may be reviewed manually. The overall AIDA system allows us to meet several objectives, including growing revenue within our risk appetite and whilst maintaining compliance.

We consider the initial screen, risk score, automated scorecards, and human decisions and overrides to all be part of the overall AIDA system.

[NOTE: for brevity, the answers that follow focus mainly on the risk scoring model but a full response would include all the portions of the AIDA system as well as their interaction].

A2 Who are the individuals and groups that are considered to be at-risk of being systematically disadvantaged by the system?

We consider women and unmarried individuals to be at risk of systematic disadvantage.

We examine this by making comparisons based on the attribute GENDER (which can take on the value MALE or FEMALE), and the attribute MARITAL STATUS (which can take on the value MARRIED or UNMARRIED).

Individuals are generally "underbanked" because they have had less access to the traditional credit system. In developing economies, women are less likely than men to have an account at a formal FSI and less likely to have borrowed formally [Demirgüç-Kunt2017]. As a result, we have included GENDER as a target for fairness analysis as women might be at risk of systemic disadvantage.

Married people can share income and wealth, so marital status can have a direct impact on an individual's ability to repay a loan. However there would not appear to be a clear justification why some differences in marital status (e.g. a married person making an individual application versus a joint application) should matter for individuals who are otherwise similar. As a result, we have included MARITAL STATUS as a target for fairness analysis.

In the answers below we refer to the attributes that define these chosen groups as "personal attributes". Personal attributes could have included other choices such as race, place of birth, or religion. This information is not in our dataset, so it is not possible to perform analyses to test for differences based on these attributes. This is not a claim that no systematic disadvantage exists for these groups.

A3 What are potential harms and benefits created by the system's operation that are relevant to the risk of systematically disadvantaging the individuals and groups in A2?

We consider that true positives and true negatives are benefits and false positives and false negatives are harms. This is illustrated in the confusion matrix below.

	Will repay	Will not repay
Approved	**TRUE POSITIVE** Credit expansion for customer and revenue for the bank	**FALSE POSITIVE** Write-down/off for bank and lowered credit score for customer over tme
Denied	**FALSE NEGATIVE** Missed opportunity to productively use credit for customer and missed revenue for bank	**TRUE NEGATIVE** Bank avoids write-off, customer avoid credit issues, but does not get to deploy the credit card feels rejected

Figure 3.4: Harms and benefits of a credit allocation system.

The system is successful when it correctly approves or denies a credit applicant, limiting the harms and maximising the benefits for both the consumer and our organisation.

There are two types of correct decisions that the system can make:

- *True positive* — an applicant is approved, they subsequently paid back the loan (or at least did not default within the 18 month period following its issuance).
- *True negative* — an applicant who would not have paid back the loan is denied. This is a counterfactual outcome that is inferred through proxy.

There are two types of error that the system can make:

- *False positive*— an applicant is approved, and they subsequently defaulted on the loan.

- *False negative* — an applicant who would have paid back the loan is denied. Similar to the true negative, this is a counterfactual outcome that is inferred through proxy.

In general, correct decisions are benefits and incorrect decisions harms. For instance, a true positive is beneficial for us, since we gain a new customer, and for the applicant, since they gain access to credit products. A false negative is harmful for us because it represents lost revenue, and for the applicant because they miss out on the benefits provided by access to credit. However, at a high-level, we assume that in the underbanked customer segment we are targeting with this credit product, having more access to credit is (in aggregate for a group) a benefit that outweighs other harms.

We make two other simplifying assumptions:

- *Harms are independent of applicant type:* false positives (i.e. defaults) are assumed to be equally harmful to all individuals. We recognise that this may be less true for SELF-EMPLOYED applicants, as entrepreneurs generally have a higher risk tolerance, and may be using the loan to invest in a business. They might perceive receiving a loan they are likely to default on to be a benefit, even if it negatively affects their credit. Similarly if the loan is taken out to pay for medical bills or other essential services, an applicant who receives a false positive might consider it a benefit.

- *Benefits can be determined after 18 months,* at which point the loan is labeled as resolved. We recognise that defaults occurring thereafter may still cause harm to the recipient.

We believe this is appropriate as this is the first FEAT fairness assessment of our system. We will consider ways to incorporate these complexities into future assessments.

A4 What are the fairness objectives of the system, with respect to the individuals and groups in A2 and the harms and benefits in A3?

In addition to our business objectives, we have two fairness objectives:

- To operate at the efficient frontier of tradeoffs so that there are no harms (or lost benefits) that could have been avoided without incurring some other harms or lost benefits. For instance, if the model can be made more accurate for virtually all applicants, this would enable fewer errors to be made without requiring tradeoffs.

- To ensure that at the operating point chosen on that frontier, particular groups and individuals are not systematically disadvantaged by the distribution of benefits and harms. More specifically, we recognise that the systematic undersupply of credit can be especially harmful to underbanked communities. We are therefore most concerned with ensuring that our AIDA system supplies *lower risk* individuals with loans at similar rates regardless of their GENDER or MARITAL STATUS. This is discussed in more detail in Part C.

3.3.2 Part B: examine data and models for unintended bias

B1 What errors, biases or properties are present in the data used by the system that may impact the system's fairness?

We use an open dataset to train and test our AIDA system. A complete list of the features used in the model can be found in the answer to B3 below.

Representation bias

The dataset was collected specifically to target the underbanked but no information is available about the upstream system that generated the population. Therefore, it is likely there are unknown representation biases inherited from the upstream system.

The risk of representation bias also depends on both absolute and relative amounts of training data. On a relative basis, less than 50 percent imbalance between classes is generally considered a relatively low level of imbalance. There are approximately twice as many FEMALE applicants as MALE and a large total number of each, and the base rate of the population is roughly a 50/50 split, so women are not underrepresented in the dataset.

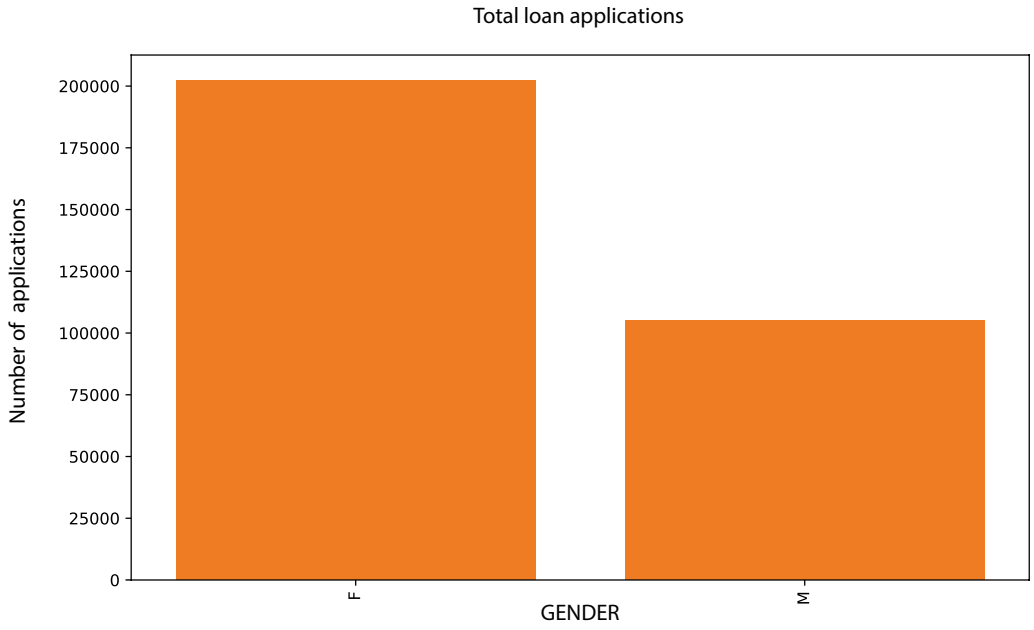


Figure 3.5: Distribution of GENDER in the dataset.

There are more MARRIED than UNMARRIED applicants, but the binned total are similar and the total numbers in each unbinned category is in the thousands, so the risk of underrepresentation in the data is low.

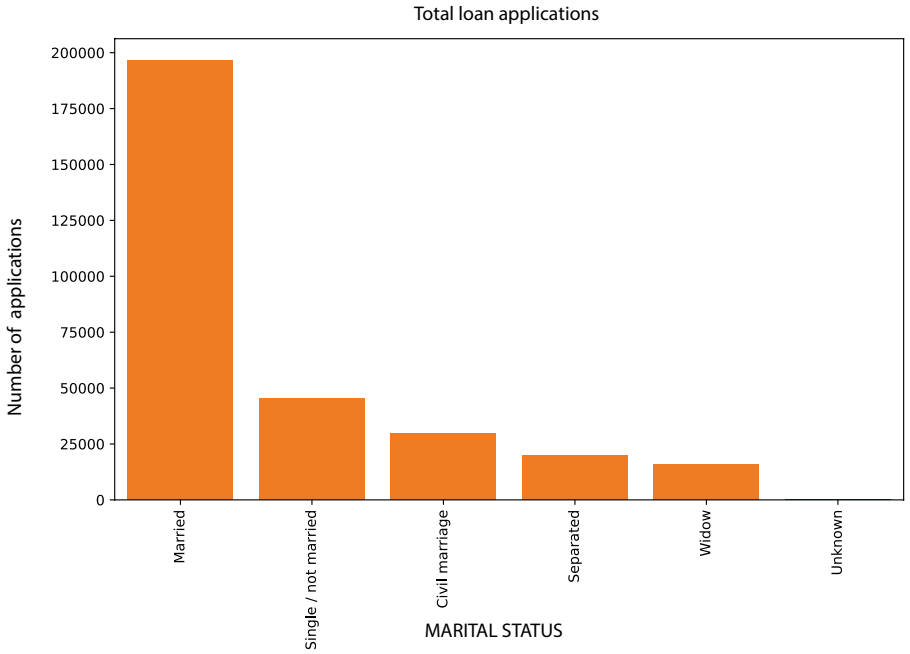


Figure 3.6: Distribution of MARITAL STATUS in the dataset.

Measurement bias

The Open credit dataset contains the target labels (default / resolve) for every datapoint. For the sake of this case study, we assume that we performed reject inference to infill missing labels, and discuss this below.

Ground truth labels of default/resolve are available for customers that were approved for a loan. There are no ground truth labels for customers that were declined (i.e. it is a hypothetical question whether a customer who was rejected would have actually defaulted). In the absence of true labels, we perform reject inference using a secondary model to supply inferred labels. This secondary model is trained on the Known population only (people for whom the true labels are known) and called the Known Good Bad model (KGB). The KGB model assigns reject records a score based on a credit scoring model.

If the predictor in the KGB credit scoring model is biased toward any groups of individual differences, then these biases will be inherited by the labels in the training data. This is a form of potential measurement bias. In addition KGB introduces a covariate shift issue. The input features are distributed differently in the accepted cohorts than in the rejected cohort. These differences are precisely what allows the model to have predictive power. Thus, the KGB model is trained on an input distribution that differs meaningfully from what it predicts on.

KGB is standard industry practice. Attempting to directly collect ground truth labels

would have its own methodological challenges. For instance, if we contacted customers who had been declined and surveyed them about whether they had received loans and had paid back those loans or defaulted, it might be only customers that did pay back their loans to other FSIs that would respond. This self-selection would re-introduce a new measurement bias.

Other potential sources of bias

The dataset indicates approximately an 8% default rate. This is a considerable class imbalance, and needs to be addressed when fitting a predictive model.

Some people do not identify as male or female but rather a non-binary gender. The application form and data scheme forces this binary choice which might create additional forms of bias. This has not yet been addressed.

The variable GENDER, which was self-reported on the application, is used to assign individuals to either the group MALE or FEMALE.

MARITAL STATUS was reduced from 6 categories to either the group MARRIED or UNMARRIED only. Response above illustrates the binning logic, which was done to simplify the analysis

B2 How are these impacts being mitigated?

Representation bias

No action was taken to address representation differences among groups.

Measurement bias

No action taken

Other potential sources of bias

We upsampled the data during training and corrected the class imbalance using the synthetic minority oversampling technique (SMOTe) [8].

B3 How does the system use AIDA models (with, or separately from, business rules and human judgement) to achieve its objectives?

A credit scoring model is used to predict the likelihood of default for a customer. Default is defined as being more than 60 days late in payment within 18 months of the loan being disbursed. The model outputs the risk of default as a scalar value between 0 and 1. Zero

indicates the highest risk of default, while 1 indicates the lowest risk of default. Applicants with a sufficiently low risk of default (above a chosen threshold) are, for the sake of this case study, considered to be approved.

More specifically, we trained a logistic regression model to predict the likelihood of default for each application based on a mixture of bureau, application, position and cash balances, and previous application features. Logistic regression has several advantages in a credit scoring context. The loss function is convex which facilitates reproducibility. The individual features each contribute in an interpretable, monotonic way to the model output, and the decision boundary is a hyperplane (linear) in the feature space, all of which make the model decisions easier to explain.

The model outputs the likelihood that the loan will be resolved as a probability value between 0 and 1. This is interpreted as a risk score. Applicants with a risk score above a certain threshold are approved, otherwise the application is declined. For the sake of this case study, we set the threshold to maximise balanced accuracy, but in the case of an actual lending scenario this threshold would be chosen more carefully based on the historic bad rates and the risk appetite of the bank. In general, model output probabilities are not necessarily well-calibrated. Calibration is addressed subsequently in B4.

The models were trained on an open dataset that had been preprocessed. During training, instances of default were upsampled using the synthetic minority oversampling technique (SMOTe) [Chawla2002]. This dataset strives to broaden financial inclusion for the unbanked population. The dataset includes attributes about the applicant and the loan contained in the current application as well as past information. Applicant attributes from the application include:

- GENDER
- MARITAL STATUS
- NUMBER OF CHILDREN
- AGE
- EDUCATION
- INCOME
- INCOME TYPE
- OCCUPATION
- LOAN TYPE
- EMPLOYMENT PERIOD
- DWELLING TYPE
- FRAUD INDICATORS
- AFFLUENCE INDICATORS

Past information includes:

- All client's previous credits provided by other FSIs that were reported to Credit Bureau
- Monthly balances of previous credits in Credit Bureau
- Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with the bank
- Monthly balance snapshots of previous credit cards that the applicant has with the bank.
- All previous applications for bank loans of clients who have loans in our sample.
- Repayment history for the previously disbursed credits related to the loans in our sample
- Behavioural data

The final model was trained on a subset of the above features, plus engineered features:

APPLICATION FEATURES

- GENDER
- AGE
- MARITAL STATUS
- NUMBER OF CHILDREN
- AGE
- EDUCATION
- INCOME (INCOME / AGE ratio, Group Income by Education, Gender, Age, Income Type)
- INCOME TYPE
- OCCUPATION
- LOAN TYPE
- EMPLOYMENT PERIOD

BUREAU FEATURES

- Number of loan enquiries
- Bureau scores and its engineered features

PAST LOANS

- Number of ACTIVE/CLOSED Loans
- Amount of those loans
- Repayment behaviour on those loans

CREDIT CARD TRANSACTIONS

- Repayment behaviour on credit card transactions

As a pre-processing step we binned categorical attributes:

- EMPLOYMENT PERIOD was binned into: <= 7yrs, > 7yrs
- AGE was binned into: <=25 yrs, 26-64 yrs, >= 65 yrs
- EDUCATION was binned into: Lower (Incomplete & Lower secondary) vs Higher (Academic, Higher & Secondary)
- MARITAL STATUS was binned into: Married (including civil); Unmarried (including single, separated, widowed)
- NUMBER OF CHILDREN was binned into: 0, 1-2, 3-6, >7

Other preprocessing steps include:

- Feature scaling using StandardScaler
- Converting categorical variables to numeric representations
- The original dataset categorised 1 as the “default” label and 0 as the “resolve” label. This ground truth was flipped for the purpose of this analysis

B4 What are the performance estimates of the AIDA models in the system and the uncertainties in those estimates?

There is only one model in the AIDA system. It is a logistic regression model trained to predict the likelihood of default. We fit the parameters of the model via maximum likelihood:

$$\sum_{i}^{n} [y_i \log(p_i) + (1-y_i) \log(1-p_i)],$$

where $p_i = \sigma(X_i^T w + c)$ and $\sigma(x) = 1/(1+e^{-x})$

The loss function applied on the logistic regression model can be fully expressed by expanding the above and including the applied l_2 regularisation term [28]. The $l2$ regularisation term added is a standard way to prevent overfitting and improve generalisation performance.

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1)$$

Typically credit scoring datasets have a strong class imbalance. To take this into account, we used *balanced accuracy* as a performance measure. Balanced accuracy is calculated as the average of true positive rate (TPR) and true negative rates: (TPR + TNR)/2. A true positive is an approved loan that was repaid. A true negative is a denied loan that would have defaulted. We also calculate the ROC curve plotting sensitivity (false negatives) against specificity (false positives). We use the area under the ROC curve (AUC) as a summary statistic of the ROC curve as another performance measure.

All performance measures are computed on a test set which is held out from the model during training and hyperparameter sweeps. The overall balanced accuracy for the logistic regression model is 0.64, while the AUC is 0.70. Additional rates are detailed in the table below. Plus-minus interval indicates two standard deviations as computed via bootstrap resampling of the test set.

Performance measure (rate)	Value	Meaning
Balanced accuracy	0.64 +/- 0.004	The average of: • the proportion of applicants approved who did (or hypothetically would) repay their loans, and • the proportion of applicants declined who did (or hypothetically would) default [The arithmetic mean of the *true positive rate* and *true negative rate* (see below)]
True positive rate	0.54 +/- 0.007	The proportion of applicants approved who did (or hypothetically would) repay their loans
True negative rate	0.74 +/- 0.012	The proportion of applicants declined who did (or hypothetically would) default
False positive rate	0.26 +/- 0.012	The proportion of applicants approved who did (or hypothetically would) default
False negative rate	0.46 +/- 0.007	The proportion of applicants declined who did (or hypothetically would) repay their loans
Positive predictive value	0.96 +/- 0.002	The proportion of repaid loans out of all approved loans
Positive rate	0.51 +/- 0.003	The percentage of all applicants getting approved loans

The test set confusion matrix for the model is tabulated below.

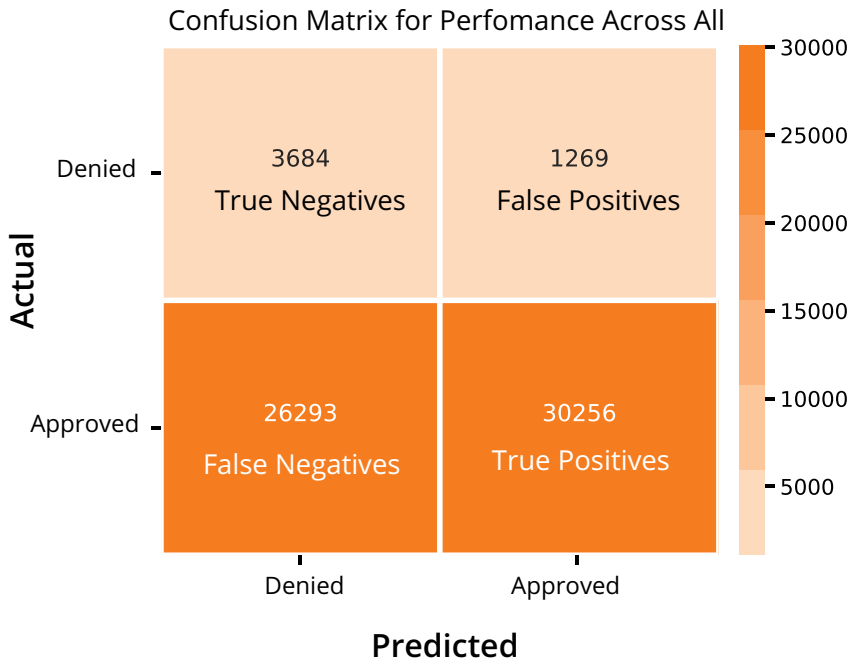


Figure 3.7: Confusion matrix on the test set.

The ROC curve for the logistic regression model on the test set is plotted below.

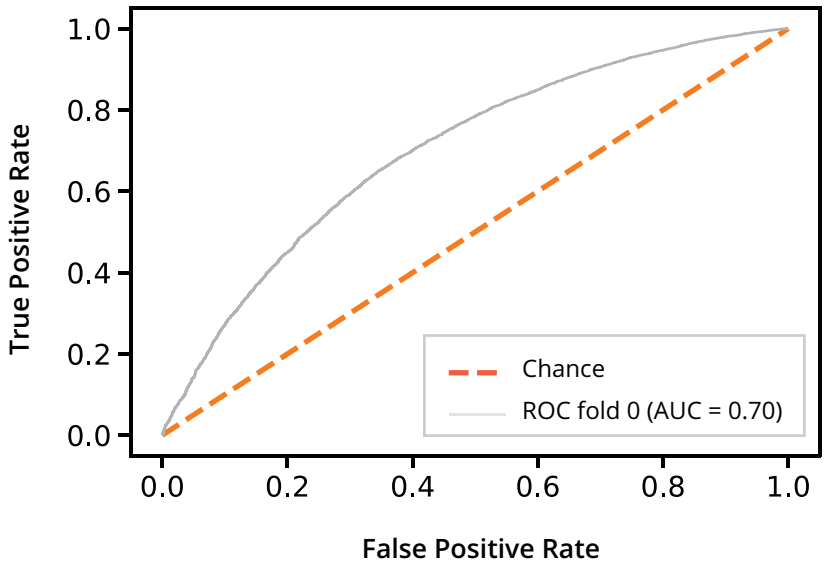


Figure 3.8: ROC curve of model performance on the test set.

We created a 80/20 split between train and test sets. This test set was fully held out during model development and contained 61K data points. Modelling choices, such as which features to include, tuning of the $l2$ regularisation parameter, and tuning the prediction threshold, were made using k-fold cross validation (k=10) on the training set.

We also want the credit scoring model to be well-calibrated, since it aims to quantify risk. Generally, logistic regression models produce well calibrated predictions [28]. However, instances of default were upsampled during training, thus the base default rate, P(y=0), does not match between the train and test set. As a result, the model output probabilities are skewed, and cannot be interpreted as a true probability. In our credit decisioning system only thresholded predictions will be used, thus no post-hoc calibration was carried out. Threshold selection was done via cross-validation, on non-upsampled validation sets.

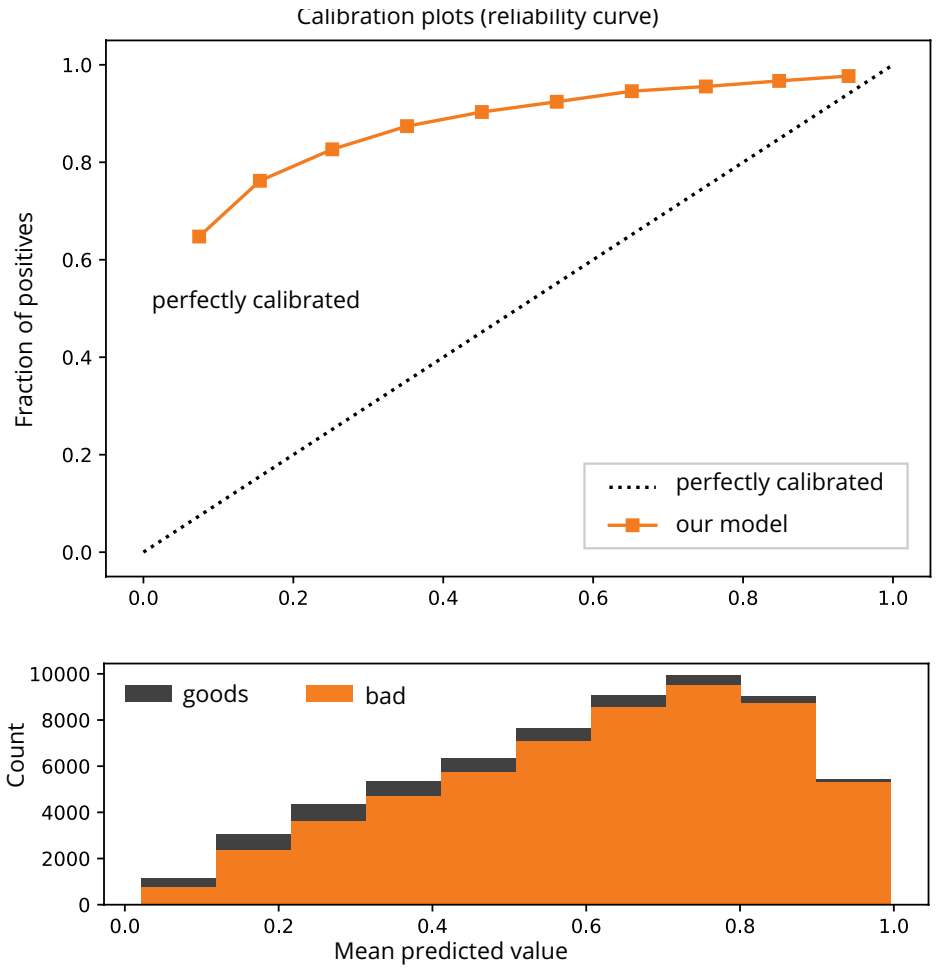


Figure 3.9: Calibration curves of model performance on the test set.

B5 What are the quantitative estimates of the system's performance against its business objectives and the uncertainties in those estimates?

Overall system performance is judged based on our business objectives, such as serving additional customers while maintaining reasonable default rates. These should correlate closely with our model's principal performance metrics of balanced accuracy and AUC.

True positives (successful loans) generate revenue for the FSI and credit for the loan applicant, and true negatives (correct declines) reduce loan losses for the FSI and may preserve the credit rating of applicants by avoiding later defaults. Therefore, the business objective is served by taking both into account.

False negatives (incorrect declines) deny credit to applicants. As some groups may already face challenges with accessing credit, it is important to take the false negative rate into account to align with the social objective of increasing access to credit, and the social constraint of avoiding increasing disadvantage.

The value to the FSI of a successful loan depends on both the interest charges and increased potential for cross-selling other products, while the cost of a defaulted loan depends on collections effectiveness and default timing. This AIDA system aims to maximise the gross profit from the difference.

For the sake of this case study, we do not presume the value of a successful loan, or the expected cost to the FSI of a default. This information was not available for the proxy dataset. Instead we use the model's primary performance indicator (balanced accuracy) to quantify the performance against our business objective (gross profit maximisation). Implicitly, this assumes that the mean cost from a default is 11.4 times greater than the mean gain from a successful loan, since the ratio of good to bad loans in the Open credit dataset is 11.4. See figure below for details.

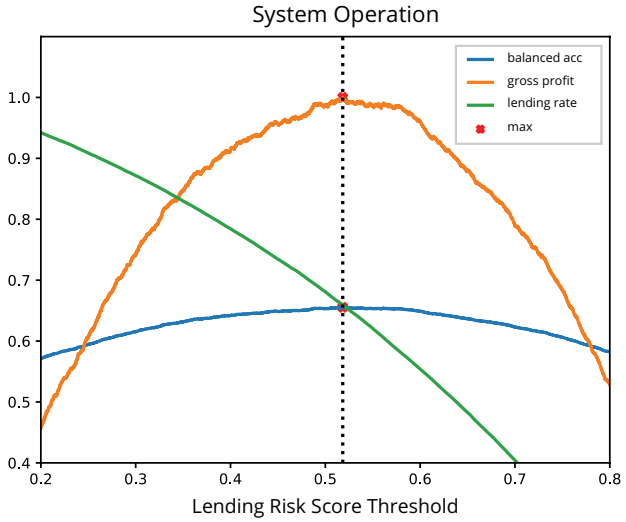


Figure 3.10: Relation between lending rate, balanced accuracy, and gross profit for the model on the test set.

Importantly, in production, we will not have access to the number of true negative or false negatives, as these correspond to rejected applicants. The "ground truth" label does not exist for the loans that never get issued, nor does it exist for the accepted applicants until 18 months after the loan is issued. (Recall we define a loan to be resolved if it does not lead to default within the first 18 months.) Thus balanced accuracy cannot be computed. Instead, we must tune our model's performance on our training data, then monitor its performance in production using other indicators.

It should be noted that the chosen measures capture the system's performance for applicants, but not for non-applicants. The initial screening (which is part of the overall AIDA system) could affect the ratio of applicants and non-applicants among individuals and groups, and this would not be fully captured by the chosen performance measures.

3.3.3 Part C: measure disadvantage

C1 What are the quantitative estimates of the system's performance against its fairness objectives and the uncertainties in those estimates, assessed over the individuals and groups in A2 and the potential harms and benefits in A3?

Disadvantage is defined as a significant difference in the rates of occurrence of harms and benefits between groups. We considered the following group fairness measures:

Fairness measure	Interpretation in credit scoring context
Demographic parity	Applicants have an equal likelihood (across groups) of having their loans approved
Equal opportunity	An equal fraction of loans are approved (across groups) to the applicants who do (or hypothetically would) pay back their loan
False positive rate balance	An equal fraction of loans are approved (across groups) to the applicants who default (or hypothetically would default) on their loan
Equalised odds	Ensuring equal opportunity and false positive rate balance
Positive predictive parity	An equal fraction of loans are paid back (across groups) by the applicants who have their loan approved
False omission rate balance	An equal fraction of loans would have hypothetically have been paid back (across groups) by the applicants who were rejected
Calibration by group	Ensuring positive predictive parity and false omission rate balance

We selected *equal opportunity* as the criteria for fairness. That is, we consider "unfairness" to mean that one group has a lower fraction of loans approved to applicants who do (or hypothetically would) repay the loan. This best fits our fairness objectives (see Response A4).

For future fairness assessments we will consider whether positive predictive rate parity should also be included.

For GENDER the 6 fairness metrics along with confidence scores are below:

FAIRNESS METRICS

	Demographic Parity	Equal Opportunity	False Positive Rate Balance	Predictive Parity	Average Odds	Calibration
GENDER	-0.15	-0.14	0.11	-0.01	0.12	-0.04

CONFIDENCE INTERVALS

	RangeL	ower	Upper
Demographic Parity @ 95%	**+/- 0.008**	**-0.157**	**0.141**
Equal Opportunity @ 95%	**+/- 0.008**	**-0.151**	**-0.135**
False Positive Rate Balance @ 95%	**+/- 0.007**	**-0.113**	**-0.099**
Predictive Parity @ 95%	**+/- 0.003**	**-0.116**	**-0.009**
Average Odds @ 95%	**+/- 0.003**	**-0.127**	**-0.122**
Calibration @ 95%	**+/- 0.002**	**-0.044**	**-0.041**

In our analysis on this dataset the *equal opportunity* score for GENDER is **-0.14**. An equal opportunity Score of **-0.14** means that if there were 100 MALE and 100 FEMALE applicants who would have paid back their loan, **14** more FEMALE applicants were actually approved than MALE applicants (when looking only at those 200 applicants).

For MARITAL STATUS the 6 fairness metrics along with confidence scores are below:

FAIRNESS METRICS

	Demographic Parity	Equal Opportunity	False Positive Rate Balance	Predictive Parity	Average Odds	Calibration
MARITAL_STATUS	-0.18	0.18	-0.15	0.01	-0.17	0.05

CONFIDENCE INTERVALS

	Range	Lower	Upper
Demographic Parity @ 95%+	**/- 0.009**	**-0.187**	**-0.170**
Equal Opportunity @ 95%	**+/- 0.008**	**-0.192**	**-0.175**
False Positive Rate Balance @ 95%	**+/- 0.008**	**-0.156**	**-0.140**
Predictive Parity @ 95%	**+/- 0.004**	**-0.008**	**-0.016**
Average Odds @ 95%	**+/- 0.003**	**-0.169**	**-0.163**
Calibration @ 95%	**+/- 0.002**	**-0.051**	**-0.054**

In our analysis on this dataset the *Equal Opportunity* score for MARITAL STATUS is **-0.18**. An *equal opportunity* score of -0.18 means that if there were 100 MARRIED and 100 UNMARRIED applicants who would have paid back their loan, 18 more UNMARRIED applicants were approved than MARRIED applicants (when looking only at those 200 applicants).

In addition to these group fairness considerations, we aim to be fair to individual applicants. Final credit decisions are made based on the application of scorecards (which implement various business rules and policies) to the results of the credit risk scoring algorithm. We consider the risk score output to be the measure of individual similarity. That is, we interpret individual fairness to mean that applicants with similar risk scores receive similar approve/deny decisions.

We measure this by identifying the volume of applicants who have nearly identical credit scores who receive different decisions due to manual review.

In order to increase the objectivity of the system and reduce the scope for judgmental discretion, we aim to apply straight through processing (STP) to as large a volume of applications as possible. In cases of STP, the output risk scores are sufficiently high or low that the scorecard can be applied automatically to produce a decision with no human intervention. In some cases the risk score may be marginal (near the cut-off) which requires manual review.

We provide extensive documentation and training in order to make the human review process as consistent as possible. Nonetheless it is possible that different reviewers might come to different conclusions. We measure the overall volume of applicants with similar scores who receive different decisions to track whether additional training or documentation is needed.

C2 What are the achievable tradeoffs between the system's fairness objectives and its other objectives?

There are numerous design choices that have been made with the objective of meeting our business and regulatory objectives. Many of these choices, such as our choice to use a logistic regression model, cannot be feasibly changed at this time, and thus are considered unachievable and out of scope.

The threshold for loan approval is a key operating parameter that affects the fairness metrics. It has been set to maximise our performance metric (balanced accuracy), but could feasibly be adjusted. We report on its effects on MALE and FEMALE applicants.

With the same threshold for all applicants, there is a slight bias against MALES in terms of equal opportunity. We investigate the effect of varying this threshold on both the model's balanced accuracy and the equal opportunity metric. We find that for the given data distribution, reducing the threshold for loan approvals would improve fairness slightly, but also lower our model's balanced accuracy.

We could also use separate thresholds for MALE and FEMALE applicants to more strongly impact the equal opportunity metric. We conduct a grid search for threshold settings that bring equal opportunity to within +/-0.0001 of neutrality while otherwise maximising balanced accuracy.

Once the analysis has been run, we can visualize the fairness-performance tradeoffs of operating the model at various lending risk threshold settings (Figure 3.11). Operating the model at a higher lending risk threshold equates to a lower loan approval rate. The x-axis shows a range of possible lending risk thresholds for men, while the y-axis shows a range of possible lending risk thresholds for women.

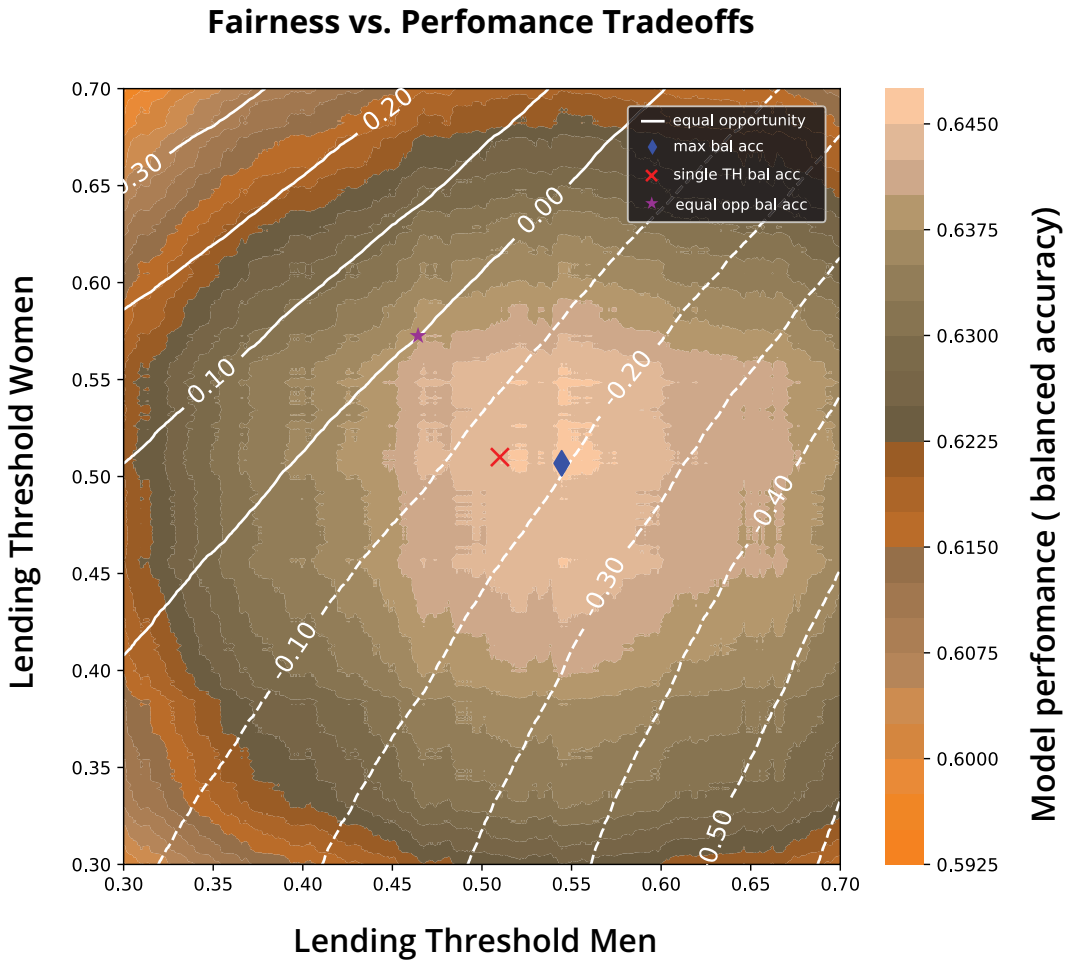


Figure 3.11: Tradeoffs between model performance and group fairness for different lending thresholds (THs) by GENDER.

The heatmap indicates the model's expected performance (balanced accuracy) when operated at each pair of risk thresholds. Recall, in our AIDA credit approval system, balanced accuracy is directly proportional to the expected gross profit.

The white contour lines indicate the equal opportunity group fairness metric with respect to gender. Equal opportunity measures the difference in the true positive rates between two groups of individuals, in this case men and women. It is computed as TPR_men - TPR_women, thus it is optimal when equal to zero (0). The true positive rate corresponds to the probability that an applicant who would hypothetically repay their loan is accepted by the model.

We plot three points of interest. The blue diamond maximizes the unconstrained model performance. The red X maximizes model performance while keeping the same lending risk threshold for both men and women. The purple star maximizes the model performance while ensuring optimal gender fairness as measured via equal opportunity.

We find that if the threshold for recommending a loan approval was adjusted lower: 0.47 for MALE and higher: 0.57 for FEMALE, this would produce fairness according to our equal opportunity measure.

In this configuration, the model's balanced accuracy drops slightly. However, there are other effects that must be considered. This type of intervention could increase systemic disadvantage for other groups that have not been prioritised. For example, the table below shows the change in fairness metrics for EDUCATION as a result of changing the (now split) thresholds to optimise for GENDER.

Before:

	FAIRNESS METRICS					
	Demographic Parity	**Equal Opportunity**	**False Positive Rate BalanceP**	**redictive Parity**	**Average Odds**	**Calibration**
EDUCATION	-0.37	-0.37	-0.46	0.03	-0.41	0.11

After:

	FAIRNESS METRICS					
	Demographic Parity	**Equal Opportunity**	**False Positive Rate BalanceP**	**redictive Parity**	**Average Odds**	**Calibration**
EDUCATION	-0.40	-0.40	-0.45	0.03	-0.42	0.09

C3 Why are the fairness outcomes observed in the system preferable to these alternative tradeoffs?

Based on the analysis in C2, for only a small drop in utility (balanced accuracy), the AIDA system could be operated with a split lending threshold to create less systematic disadvantage across MALE and FEMALE applicants according to our chosen metric of *equal opportunity*.

At this time, we will continue to operate with a single threshold maximising balanced accuracy. However, a more in-depth study should be carried out to determine whether the identified split threshold can be used in an effort to better meet our fairness objectives. This study would consider the additional financial risks from operating at a more relaxed lending rate, as well as the repercussions on other groups of applicants. It would also investigate whether using a split threshold, which treats applicants differently based on their GENDER (albeit with the aim of improving fairness), is compliant with our internal ethics policies.

3.3.4 Part D: justify the use of personal attributes

D1 What personal attributes are used as part of the operation or assessment of the system?

The following variables were considered to be personal attributes:

- GENDER
- AGE
- MARITAL STATUS
- NUMBER OF CHILDREN

D2 How did the process of identifying personal attributes take into account ethical objectives of the system, and the people identified as being at risk of disadvantage?

GENDER and MARITAL STATUS were considered personal attributes as explained in Response A2.

AGE is often correlated with income and wealth as individuals progress in their careers, and younger people are most likely to be unbanked or underbanked [Parker2016]. However age discrimination is an issue in other settings (such as employment) and is protected for credit scoring in jurisdictions such the U.S., and U.K. In Singapore, even though there is no legislative protection for age, if other factors (such as income) are controlled for, given the short 18 month loan term considered, there would not be a clear justification why AGE is an individual difference that should matter — that is, it would appear unfair if a 40-year old was approved and a 60-year old with identical characteristics was denied. This is why AGE was considered personal.

Similarly, each additional child adds expenses to a household, so the number of children that a person has can have a direct impact on the residual money they have available to repay a loan. Nevertheless, discrimination against parents is protected in some other settings (for example in the U.S., rental application cannot be denied on the basis of having children), and our case study brand is “family-friendly”. This is why NUMBER OF CHILDREN was considered to be a personal attribute.

D3 For every personal attribute and potential proxy for a personal attribute, why is its inclusion justified given the system objectives, the data, and the quantified performance and fairness measures?

Each personal attribute's impact on systemic disadvantage, as measured by equal opportunity, based on the personal attributes GENDER and MARITAL STATUS was calculated. We computed the impact using a Leave-One-Covariate-Out (LOCO) [Lei2018] approach on the logistic regression model. In this approach, we train a new model by dropping each feature, one at a time, to see the impact on the fairness metric.

The results for GENDER are tabulated below. We report the difference between the baseline model and LOCO result (baseline - LOCO).

Personal attribute	Impact on systematic disadvantage based on MARITAL STATUS	Impact on model accuracy based on LOCO Analysis
GENDER	-0.0001	0.0000
MARITAL STATUS	+0.1590	0.0051
AGE	-0.0018	0.0007
NUMBER OF CHILDREN	-0.0090	0.0056

Other key attributes from the model are included below to provide a relative scale for personal attributes

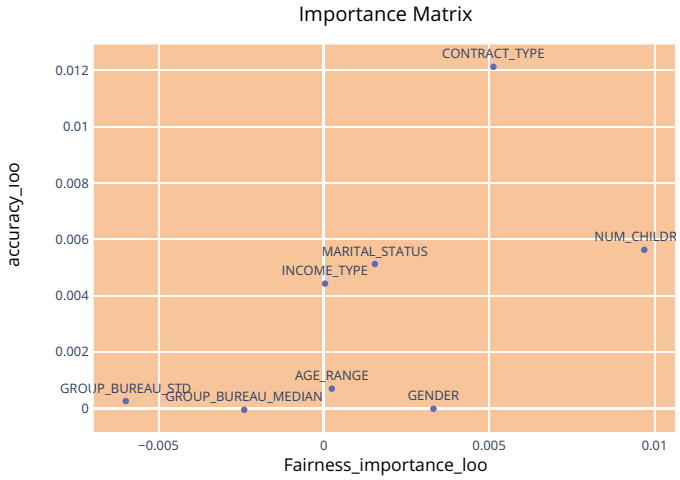


Figure 3.12: Feature importance vs fairness importance (as measured by the equal opportunity for GENDER).

GENDER appears to be an outlier: it has relatively high impact on systematic disadvantage as measured by equal opportunity for GENDER, and relatively low impact on overall model accuracy (and thus business objectives). We would recommend to the AIDA System Owner that it be considered for removal. Another option is to remove NUMBER OF CHILDREN. This removal would reduce the model accuracy by 0.006. which translates into following values:

Additional Bad loans (Increase in FPR) : $ 1,845,484

Additional Missed Opportunity (Increase in FNR) : $ 1,684,714

The next step is to evaluate, possibly with input from the AIDA System Assessor, whether the cost is worth the benefit.

The results for MARITAL STATUS are below:

Personal attribute	Impact on systematic disadvantage based on MARITAL STATUS	Impact on model accuracy based on LOCO Analysis
GENDER	-0.0001	0.0000
MARITAL STATUS	+0.1590	0.0051
AGE	-0.0018	0.0007
NUMBER OF CHILDREN	-0.0090	0.0056

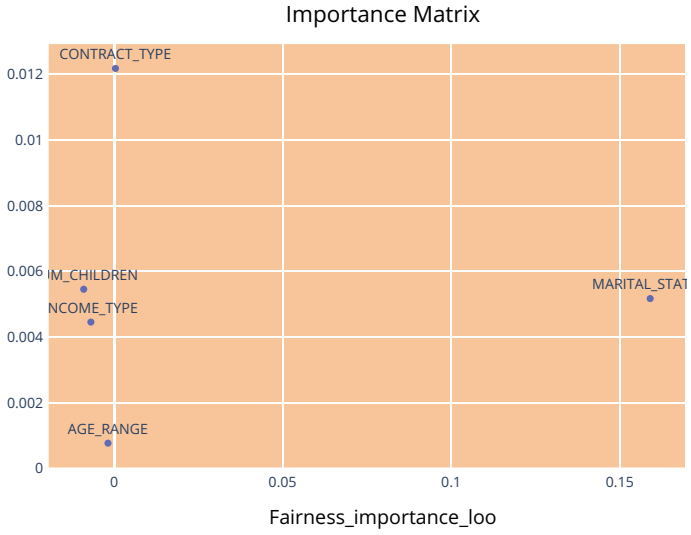


Figure 3.13: Feature importance vs fairness importance (as measured by equal opportunity for

MARITAL STATUS).

In case of fairness for MARITAL_STATUS, we can find the financial impact of dropping INCOME TYPE as a result of the accuracy drop of 0.0045 as shown below.

Additional Bad loans (Increase in FPR) : $USD 2,375,339

Additional Missed Opportunity (Increase in FNR) : $USD 2,185,011

The next step is to evaluate, possibly with input from the AIDA System Assessor, whether the cost is worth the benefit.

The following were considered non-personal attributes:

EDUCATION

INCOME

OCCUPATION

BUREAU_1

BUREAU_2

BUREAU_3

In theory, some non-personal attributes such as education or income might act as proxies for personal attributes such as GENDER (for instance, if more women were self-employed). We did not find strong correlations between any of the personal and non-personal attributes except for the bureau attributes BUREAU_1 which is associated with GENDER and AGE.

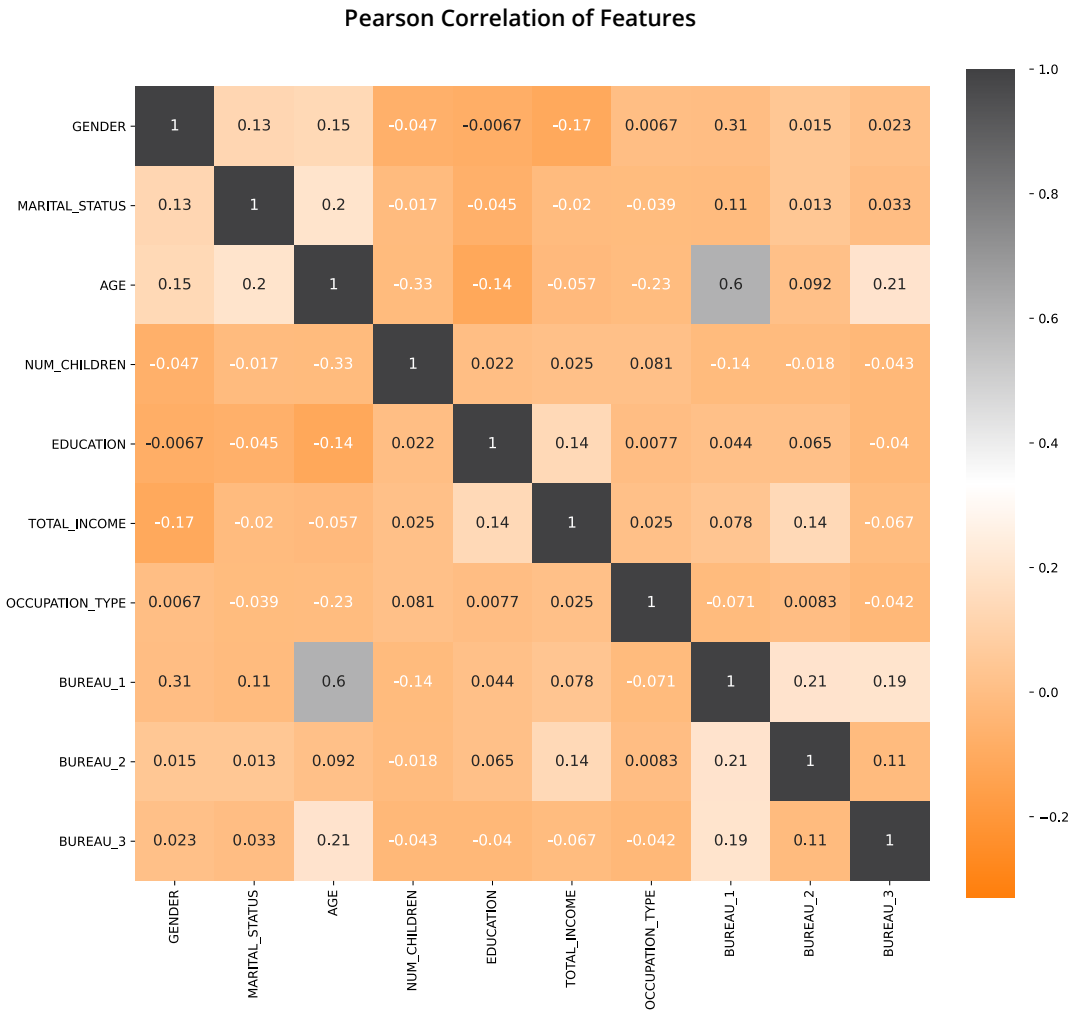


Figure 3.14: Correlation plot between model features.

As mentioned in Response D4, BUREAU_1 is correlated with GENDER. It is also less important to the model's performance than BUREAU_2 and BUREAU_3, which are also bureau-derived scores.

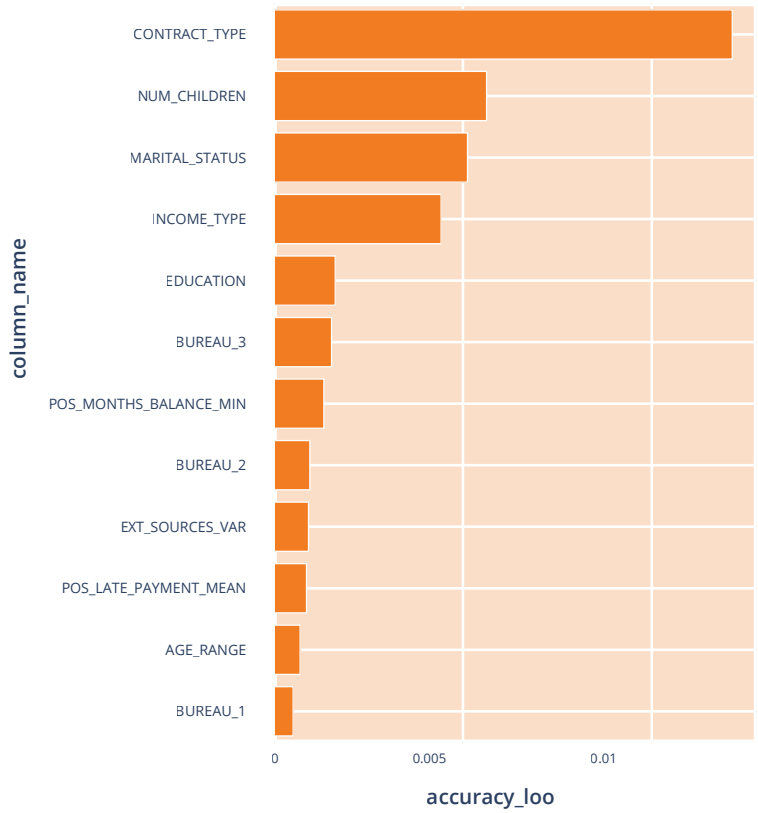


Figure 3.15: Feature importance results of leave-one-out analysis (top 12 features), as measured by the feature's impact on balanced accuracy during our leave-one-out analysis.

BUREAU_1, BUREAU_2 and BUREAU_3 are all very similar variables, but BUREAU_1 is correlated with GENDER. Removing BUREAU_1 from the model reduces the accuracy by only 0.0005. Since the model's performance is nearly the same when only including BUREAU_2 and BUREAU_3, then BUREAU_1 can be considered for removal.

3.3.5 Part E: examine system monitoring and review

Since this illustrative model is not in production, discussion of system monitoring and review is omitted. However, please refer to Section 3.2.5 for credit specific monitoring and review considerations, and Document 1 Appendix 1.4 for general considerations.

E1 How is the system's monitoring and review regime designed to detect abnormal operation and unintended harms to individuals or groups?

E2 How does the system's monitoring and review regime ensure that the system's impacts are aligned with its fairness and other objectives (A1 and A4)?

[The below answers are general and hypothetical because the case-study model was not connected to a real production system with monitoring and review processes or technology]

Unfortunately for credit scoring models, where the outcomes are only known several months in future, we can't put real time monitoring of fairness metrics. Hence practically, we can only calculate these fairness metrics while in model development/validation and then later review them periodically on historic data as and when outcomes become known.

Unlike other fairness metrics which rely on knowing outcomes, demographic parity simply compares the lending rate between groups. While it is not our primary fairness metric, we would monitor its value in production to ensure the system isn't deviating too far from its expected behaviour. This would be used in combination with the techniques for drift detection listed in the general considerations section.

E3 What are the mechanisms for mitigating unintended harms to individuals or groups arising from the system's operation?

As discussed in Response E1, the main analysis will take place during development and validation. When outcomes of loans become available which characterize the model's behaviour in production, they are analyzed and might lead to a decision to update the model. Hence, these learnings are cycled into the development of future models.

3.4 UOB reflections on applying the Methodology

3.4.1 Introduction

At UOB, we are committed to cultivate deep customer relationships and to turn our customers into strong advocates. Customers' opinions are shaped based on their impression and manner in which we treat them, understand their needs, address their concerns and care for them in good and bad times. We ensure that through our approach and consistent practice, our customers will feel that they can count on us, grow with us, recommend us to their family and friends and experience our culture of trust that helps us stand out in the industry.

As we become more digital-focused, it is also increasingly important that we manage our customer's data ethically to ensure our strong commitment to our customers. At UOB, we have set up a multi-disciplinary data ethics task force to develop a governing framework, policies and processes that ensure the Bank uses data in a responsible and ethical manner. We are sharing the lessons that we have learnt along the way on the application of data ethics in real-world situations with others in the banking and finance industry through the Veritas Consortium. We believe that doing right by customers through the ethical use of data is the responsible and sustainable way to do business.

In this phase of Veritas project, UOB partnered with Element AI to apply the Methodology to real credit scoring systems and business processes. Credit scoring is an important business process in UOB for which we ensure that the credit assessment of new credit applications is robust and effective. This helps us understand our customers better so that we can recommend better products and services to them. In this section of the document, we share our experiences and findings of the project.

3.4.2 Learning from applications of the Methodology

While applying the FEAT Fairness Assessment Methodology on our credit scoring model, we have made several key findings and observations. We will share the details of our findings in this section of the document.

Part A: objectives and context

Objectives and constraints

The credit scoring model analysed for the Veritas project is the Credit Card & CashPlus Application Scorecard model which analyses retail customers applying for their first credit card or unsecured loan with UOB. The objective of the model is to optimise the analysis of the credit profile of new customers so that we understand the credit profile of our customers better and guide better credit decisions.

The model focuses on the "Thin" Bureau segment of customers who have no credit bureau history, or very short bureau history of less than 6 months, and banks are typically limited to only demographic information on application form to analyse the credit scores of this segment of customers. It is also called the "New-To-Bureau" (NTB) model to reflect the nature of the segment of customers.

At UOB, we have developed several bespoke Application and Behaviour scorecard models for credit scoring for a comprehensive range of business objectives. We have specifically chosen the NTB model for the Veritas project as we believe that a model that depends heavily on demographic information will suit the purpose of the Veritas project in quantifying measures on personal attributes for the FEAT Fairness Principles.

Harms and benefits discussion

We are aligned with the credit scoring case study that true positives and true negatives are generally considered as benefits, and false positives and false negatives are generally considered as harms. This is illustrated in the confusion matrix below.

	Will repay	**Will not repay**
Approved	**TRUE POSITIVE** Credit expansion for customer and revenue for the bank	**FALSE POSITIVE** Write-down/off for bank and lowered credit score for customer over tme
Denied	**FALSE NEGATIVE** Missed opportunity to productively use credit for customer and missed revenue for bank	**TRUE NEGATIVE** Bank avoids write-off, customer avoid credit issues, but does not get to deploy the credit card feels rejected

Figure 3.16: Harms and benefits for UOB's credit allocation system.

However, there are many complex situations in real-world that cannot be generalised by the simple confusion matrix and will require much more consideration for such situations. At UOB, we are focused on establishing a positive outcome for our customers, and we have set up a thorough credit approving process in which credit scoring is one of the many important inputs to the process for the consideration of applications with varying degrees of complexity.

For the purpose of this project, we will focus on the general cases as defined by the mathematical measures of systemic disadvantage in Section 3.3.3.

Protected groups and individuals

As the segment of customers covered by the NTB model do not have sufficient credit bureau history, the NTB model depends heavily on the personal attributes provided through the application form. Examples of demographic details include age, gender and residential address. The details provided were analysed comprehensively to establish the importance and relevance in the model so as to achieve the positive outcomes for all our customers.

Part B: data and models (accuracy and bias)

Dataset

The development process of the NTB model includes model training, validation and monitoring, and we have utilised 3 years of historical data for model development to ensure that the model is robust.

Model

The NTB model was developed with the industry-standard classification machine learning algorithm.

Fairness measures

We applied the fairness measures, as described in the Methodology and accompanying credit scoring case study, on the NTB model for evaluation, and we observed that the full set of fairness measures should not be applied blindly without much consideration. Instead, we should carefully select a relevant subset of fairness measures for each attribute of a model on a case by case basis.

For example, the “equal opportunity” and “false positive rate balance” metrics are individual components of the “equalised odds” metric, and we should select either set depending on the nature of the analysis. We also observed that the “equal opportunity” and “predictive parity” metrics are more relevant than the “demographic parity” metric in a number of attributes where base default rates between groups are very different. This is also noted in Section 3.3.3.

Part C: systemic disadvantage

Findings from the fairness analysis

The overall findings from the application of the various fairness measures on the NTB model were in line with expectations, and the various aspects of distribution for the different applicant groups were captured by the measures. From the initial assessment of the various key fairness measures of the NTB model, we did not observe any systemic disadvantage for any group of customers as the fairness measures are close to the neutral point.

Often during the application of the fairness measures, our analysis led to examining the magnitude of benefits and harms that vary between groups of customers and we looked at rates of incidence between relevant cohorts of applicants. For example, instead of examining rates of loan approvals for all applicants, we wanted to quantify more specifically, the rates of loan approvals for defaulting applicants and the rate of approvals for non-defaulting (repaying) applicants. Naturally, our attention tended to focus more on the equal opportunity or predictive parity measures as opposed to demographic parity.

Generally, our analysis brought to light the tradeoffs between potential benefits and harms of an AIDA system and areas where justifications are warranted. In the case of the NTB model, most of the metrics calculated had acceptable level of values and required minimal justification with a few border line cases as demonstrated below:

Example 1

The credit card and CashPlus application form requests for “employment type” information where applicants are allowed to specify whether they are self-employed or receive a salaried income. The fairness metrics were run on the groups of salaried and self-employed applicants. The attention of our analysis was drawn to the “equal opportunity” metric which indicated that salaried applicants had slightly more loans approved out of all qualified applicants (people who have repaid). However, the metric’s confidence interval range was fairly large due to the low volume of applicants belonging to the self-employed group in the validation dataset. As a result, we noted that the “equal opportunity” analysis results for the “employment type” attribute is currently inconclusive, and will review the analysis again after we have collected sufficient data to have a meaningful analysis.

Example 2

The credit card and CashPlus application form also requests for gender information - either male or female. Calculating the “demographic parity” metric on the gender groups, examines the rate at which the NTB model predicts an advantageous outcome for one group compared to another. The demographic parity metric showed that female applicants had a slightly higher chance of getting a loan as compared to male applicants.

However, the "demographic parity" metric fails to account for whether the proportion of approved loans are actually granted to qualified applicants, who are applicants will repay the loan. And thus the resulting metric could just be a reflection on the accuracy of the NTB model. In fact, this was this case. We examined the rates in which each gender actually repaid their approved loans and found that 95% of all female applicants repaid compared to 85% of all male applicants that repaid. This is referred to as a difference in the base rates of qualified applicants.

The "equal opportunity" and "predictive parity" metrics, which take into account the distribution of correct and incorrect loan approvals, showed a much smaller disparity in approval rates between qualified male and female applicants. Thus, by examining the circumstances under which harms and benefits occur, we were able to justify the gender disparity according to the demographic parity measure.

Part D: justifying personal attributes

According to FEAT Fairness Principles, the use of personal attributes as input factors for AIDA-driven decisions should be justified. Inline with these principles, we are utilising the provided references in the credit scoring case study to conduct an analysis of all the personal attributes that were used as input features to the NTB model. The approach of the analysis is as follows:

1. We performed the permutation approach as suggested in the credit scoring case study to find the fairness feature importance for the personal attributes in regards to the protected attributes and a select fairness metrics. The features higher on this list contribute more towards the bias measured by the fairness metrics.
2. We then overlaid this analysis with the feature importance for all the features in the model.

This analysis gives us the ability to classify the features by their importance based on accuracy and fairness. The analysis is still in progress, and our initial findings did not reveal any features that have a high impact on systematic disadvantage and low impact on model accuracy

Part E: monitoring and review

We have set up a multi-disciplinary data ethics task force to formulate the data ethics governance model within UOB, encompassing the best practices and tools for FEAT evaluation.

Conclusion

At UOB, we are committed to put our customers and their financial goals first, and the commitment includes managing our customers' data ethically. Through the Veritas project, we are able to contribute to the development of the "Fairness" measures, and the overall findings from the use of the measures helps us verify the fairness of our AIDA model through better understanding of the tradeoffs specified by the measures. We will continue to enhance our governance mode and processes through the incorporation of best practices and tools such as the "Fairness" measures by our data ethics task force.

Acknowledgements

We acknowledge the following people for their contribution on this project.

Role	Name	Organization
Project Sponsor	Sopnendu Mohanty	Monetary Authority of Singapore
Project Director	Li Xuchun	Monetary Authority of Singapore
Project Lead	Zhang Qiang	Monetary Authority of Singapore
Project Manager	Hardeep Arora	Element AI
Project Manager	Heok Kee Oon	EY
Project Manager	Oliver Cheung	IAG Firemark Labs
Project Manager	Dave Vijay	HSBC
Project Manager	Amos Ong	HSBC
Project Manager	Ng Kok Keong	UOB
Subject Matter Advisor	Prof. Robert Williamson	Fellow of the Australian Academy of Science
Subject Matter Advisor	Prof. Richard Zemel	University of Toronto
Lead Subject Matter Expert	Marc Etienne Brunet	Element AI
Lead Subject Matter Expert	Grace Abuhamad	Element AI
Lead Subject Matter Expert	Robin Schlarb	EY
Lead Subject Matter Expert	Lachlan McCalman	Gradient Institute
Subject Matter Expert	Daniel Steinberg	Gradient Institute

Role	Name	Organization
Subject Matter Expert	Hardeep Arora	Element AI
Subject Matter Expert	Richard Zuroff	Element AI
Subject Matter Expert	Marie-Claude Ferland	Element AI
Subject Matter Expert	Heok Kee Oon	EY
Subject Matter Expert	Vikas Deep Sharma	EY
Subject Matter Expert	Neena Antal	EY
Subject Matter Expert	Jason Tuo	EY
Subject Matter Expert	Tiberio Caetano	Gradient Institute
Subject Matter Expert	Gunjan Bhatt	HSBC
Subject Matter Expert	Dexter Ang	HSBC
Subject Matter Expert	Marjorie Chan Bona	HSBC
Subject Matter Expert	Huah Cheng Jiann	UOB
Data Scientist	Summit Bhalla	Element AI
Data Scientist	Eniola Alese	Element AI
Data Scientist	Angus Leigh	Element AI
Data Scientist	Andrey Kostenko	IAG Firemark Labs
Data Scientist	Oxana Samko	HSBC
Data Scientist	Jun Xu	HSBC
Data Scientist	Robert Cheah Tong Ngee	UOB

We would also like to express our appreciation to Amy Shi-Nash (HSBC), Bill Simpson-Young (Gradient Institute), Chris Lim (EY), Johnson Poh (UOB), Nicolas Chapados (Element AI) and Richard Lowe (UOB) who have supported and contributed to this project.

Bibliography

1-10

[1] Association of Banks in Singapore. (2017). Code of Consumer Banking Practice. Retrieved from *https://www.abs.org.sg/docs/library/cocbp_nov2017.pdf*.

[2] Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. Retrieved from *https://fairmlbook.org/*.

[3] Bartlett, R., Morse, A., Stanton, R., Wallace, N., Adelino, M., Das, S., … participants at Berkeley, seminar U. (2019). Consumer-Lending Discrimination in the FinTech Era.

[4] Bickel, S., Brückner, M., Scheffer, T., 2009. Discriminative learning under covariate shift. Journal of Machine Learning Research 10, 2137–2155.

[5] BIS. (2006). Basel Committee on Banking Supervision. 2006. Retrieved from *https://www.bis.org/list/bcbs/spp_12/from_01012006/index.htm*.

[6] Box, G. E. P., Hunter, J. S., & Hunter, W. G. (2005). Statistics for Experimenters: Design, Innovation, and Discovery.

[7] Sharp, B. (2013). Marketing: theory, evidence, practice. Retrieved from *https://catalogue.nla.gov.au/Record/5982946.*

[8] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2011). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. *https://doi.org/10.1613/jair.953*.

[9] Charles River Associates. (2014). Evaluating the fair lending risk of credit scoring models. Retrieved from *http://www.crai.ca/sites/default/files/publications/FE-Insights-Fair-lending-risk-credit-scoring-models-0214.pdf*.

[10] D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., & Halpern, Y. (2020). Fairness is not static: Deeper understanding of long term fairness via simulation studies. FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 525–534. *https://doi.org/10.1145/3351095.3372878*.

11-20

[11] Demirgüç-Kunt, A., Klapper, L., Singer, D., Ansar, S. and Hess, J. (2018). The Global Findex Database 2017: Measuring Financial Inclusion and the Fintech Revolution. The World Bank. *https://globalfindex.worldbank.org/*.

[12] Ferrier A. (2014). The Advertising Effect: How to change behavior. Oxford University Press.

[13] Federal Financial Institutions Examination Council's (FFIEC). (2009). Interagency Fair Lending Examination Procedures. Retrieved from *https://www.ffiec.gov/PDF/fairappx.pdf*.

[14] Hand, D.J. and Henley, W.E. (1994). Can reject inference ever work?. IMA Journal of Mathematics Applied in Business and Industry, 5 (1), 45-55.

[15] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 3323–3331. Retrieved from *https://arxiv.org/abs/1610.02413v1*.

[16] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning Data Mining, Inference, and Prediction. Springer Series in Statistics.

[17] International Accounting Standards Board. (2014). IFRS 9 Financial Instruments. Retrieved from: *https://www.ifrs.org/issued-standards/list-of-standards/ifrs-9-financial-instruments/*.

[18] Karakoulas, G. (2004). Empirical validation of retail credit-scoring models. The RMA Journal, 87(1), 56-60.

[19] Kane, K., Lo, V. S. Y., & Zheng, J. (2014). Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. Journal of Marketing Analytics, 2(4), 218–238. *https://doi.org/10.1057/jma.2014.18*.

[20] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-Free Predictive Inference for Regression. Journal of the American Statistical Association, 113(523), 1094–1111. *https://doi.org/10.1080/01621459.2017.1307116*.

[21] Lin, J. (1991). Divergence Measures Based on the Shannon Entropy. IEEE Transactions on Information Theory, 37(1), 145–151. *https://doi.org/10.1109/18.61115*.

[22] Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). Delayed Impact of Fair Machine Learning. Retrieved from *https://arxiv.org/abs/1803.04383*.

[23] Monetary Authority of Singapore. (2013). Guidelines on Fair Dealing. *https://www.mas.gov.sg/regulation/guidelines/guidelines-on-fair-dealing---board-and-senior-management-responsibilities-for-delivering-fair-dealing-outcomes-to-customers*.

[24] McCarthy, J. E. (1964). Basic Marketing. A Managerial Approach. Homewood, IL: Irwin.

[25] Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. Pattern Recognition, 45(1), 521–530. *https://doi.org/10.1016/j.patcog.2011.06.019*.

[26] Morgan, S. L., & Winship, C. (2014). Counterfactuals and causal inference: Methods and principles for social research. In Counterfactuals and Causal Inference: Methods and Principles for Social Research. *https://doi.org/10.1017/CBO9781107587991*.

[27] Parker, GG., Van Alstyne, MW., and Choudary, SP. (2016). Platform Revolution. 1st ed. New York, NY: W. W. Norton.

[28] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. In Journal of Machine Learning Research (Vol. 12). Retrieved from *http://scikit-learn.sourceforge.net*..

21-32

[29] Olegario, R. (2016). The Engine of Enterprise: Credit in America. Cambridge, MA: Harvard University Press.

[30] Siddiqi, N. (2005) Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. Wiley.

[31] Verstraeten, G., & Van Den Poel, D. (2004). The Impact of Sample Bias on Consumer Credit Scoring Performance and Profitability.

[32] White House Office of Science and Technology Policy. (2014). Big Data: Seizing Opportunities, Preserving Values. Retrieved from *https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf*.

LEGAL NOTICE

This report is prepared and issued by the MAS, HSBC, UOB, EY, Element AI, Gradient Institute, and IAG Firemark Labs.