Validate AI

Predicting through a crisis

Second white paper

Predicting through a Crisis

Authors

Dr A. Bellotti, School of Computer Science, University of Nottingham Ningbo China
Professor D.J. Hand, Department of Mathematics, Imperial College, London
S. Khan, Chief Data Officer's team, Her Majesty's Revenue and Customs, Manchester

Corresponding author: Anthony Bellotti, Anthony-Graham.Bellotti@nottingham.edu.cn

Contents

1. Overview

The COVID 19 pandemic has thrust the world into an unprecedented crisis and the immediate priority for governments has been the preservation of life and sustaining basic standards of living and security for citizens. As the crisis unfolds and in the post-crisis era, numerous issues will need to be addressed, including importantly how existing AI systems should be re-calibrated to cope with the new environment in which they are to function. Fundamentally we need to ask the questions "Is the AI system still relevant and trusted for the purpose it was developed?" and "How should the AI system be modified or modify itself so that it continues to be relevant and trusted?"

This paper highlights options for evaluating and addressing AI system changes, facilitating answers to those questions resulting from the pandemic and indeed other dramatic changes. It is a sequel to an earlier White Paper available from the website of the inaugural *Validate AI* conference, 2019 (https://validateai.org/2019-conference).

The Validate AI initiative focusses on issues of AI system trustworthiness and maintenance, promoting the discussion of topics such as population drift and model sensitivity that historically have been considered more in the academic sector and less so in practitioner communities. On the other hand, some sectors have taken the lead in other areas, such as stress testing. Banking is one example where financial regulators have adopted stress testing to test banks' resilience to economic downturn. Thus, we draw on both academic research and industry practice in considering AI validation methodology. In the case of the current pandemic, problems of trustworthiness in AI systems that predict human behaviour and outcomes are likely to occur. Examples include AI systems to predict risk relating to patient health, banking fraud or tax evasion. The risks may well be less evident in AI systems predicting automatic/mechanistic outcomes such as autonomous driving, assembly line failure etc.

There is now an urgent need to provide practical solutions to address pandemic related AI system issues that can be tested and deployed quickly, as well as testing more novel complex methods that can be adopted later. This paper is intended to serve as a starting point, and we invite experienced academics and practitioners in the field to help develop a robust set of principles to maintain AI system relevance in a time of global crisis. Solutions need to consider changing data feeds impacting the relevance of existing model features and weights and to explore appropriate re-alignment, re-build, or replacement options. We intend to develop a flow chart to enable easier identification and adoption of the most appropriate techniques to address different AI system misalignment issues.

> We invite experts in the field as well as other interested parties to email Dr Anthony Bellotti in the first instance to progress topics highlighted in this paper for your organisation or as a part of the wider group discussion.

This paper is organized as follows. In Section 2 we provide a detailed overview of the problem and then we address tackling the problem in three parts:-

- Monitoring (Section 3)
- Testing for Sensitivity (Section 4)
- Remedial actions (Section 5)

2. Problem Definition

If we are predicting through a crisis or social/economic regime shift, the central problem is that our **predictive algorithm P** is making predictions based on historical data from one distribution, call it H, for future observations following a different distribution, call it F, that reflects some major change or response to events. The central question we pose is:

How reliable will P be as a predictor under the distributional change?

As an example, we may have a credit risk model that was built on training data over a period of time when unemployment rate was 2 to 4%. What would happen if the unemployment rate increases to 7% or even 25%? (There are some economic projections that 25% is a plausible scenario.) We can usefully distinguish between theory-driven (iconic, mechanistic) and data-driven (empirical) models. The former are based on some kind of understanding of the mechanisms underlying the process (e.g. disease progression models in medicine, econometric models in finance, human behaviour models in retail). The latter are based purely on observed relationships in empirical data. This means that the former are at risk from mis-specified theory, while the latter are at risk if the circumstances and the relevant distributions change. This latter is, of course, exactly the state we find ourselves in now, and is what prompted this paper. Vulnerability to changing distributions means that AI models have a fundamental brittleness. In some situations (e.g. the financial sector) proposals have been made to combine the two types of model to yield less vulnerable approaches (e.g. Hand et al, 2008). For discussion of the role and specification of models in statistical analysis, also see Box and Hunter 1965, Lehmann 1990 and Cox 1990.

The issue here is one of **robustness**. How robust is the performance of the predictive algorithm to changes in the underlying distribution of data? To answer this question, we need to monitor and measure robustness[1]. If we determine that the AI system is not robust, we need to take some remedial action. Sections 3 and 4 of this article address these two activities respectively. The financial industry is familiar with **model risk** and considering the robustness of a model over time. In particular, strategies involving scenario and stress testing that we list below are well-known, along with other model management strategies (Black et al. 2018). Nevertheless, model risk for AI systems and managing models through a crisis is not well understood. As Noel Quinn, Chief Executive of HSBC said, "The one thing that I have learned in 34 years in banking is that credit risk models and scorecards will work really, really well except when there is a significant abnormal shock." [White and Cruise, 2020].

Some further considerations are given below.

[1] The word "robustness" is also used in a somewhat different way by statisticians to refer to how robust a model is to deviations from modelling assumptions. Alternatively, the word "sensitivity" is used in statistics.

Notation used in this article.

P = Predictive algorithm;
y = Label to be predicted by P;
x = Vector of features used by P to predict y;
H = Distribution of historic data (known);
F = Distribution of future data (unknown);
F^* = Estimate of distribution F.

2.1 Is it really a distributional change problem?

We discuss the problem in terms of moving from distribution H to F, to emphasize that the change could be sudden. More generally we could talk about a distribution parameterized by time, $D(x;t)$ say. Then H and F are snapshots of D at different times. On the other hand, it may also be useful to deal with the marginal M of D across all time (covering the periods for H and F). Dealing with the marginal M is a useful device for applying many statistical methods. In particular, it makes sense to build a statistical model on M and think of crisis events in terms of extreme values from M, such as Value-at-Risk that is used in finance (e.g. see Bellotti and Crook 2013). This approach is just a different way to express distributional change over time.

2.2 How much change do we expect?

While we may suppose we understand distribution H, at least from the empirical distribution of historical data, we may have very limited understanding of F. If we suppose we can know nothing, then no forecasts can be made. However, in reality, we might suppose that F would have some resemblance to H. There are some characteristics we suppose would remain fairly steady, e.g. proportion of males/females in a population, whilst others we may imagine changing quite dramatically, e.g. unemployment rate. This sort of assumption is the first step in shifting from a purely data-driven model to a hybrid data/theory-driven model. Then, even though we may expect dramatic distributional changes, these changes could perhaps be estimated through economic models or through extrapolation from changes in distribution H over time. These expectations of what F will be like can be encapsulated in an expected distribution F^* (at least theoretically). We can then measure the discrepancy between H and F^* using Kullback-Liebler Divergence or *population stability index* (see e.g. Wu and Olson 2010) or some other measure of distributional difference. However, we cannot quantify how far our expectation F^* will deviate from F so this will only be indicative of the difference between H and F itself.

Of course, distributional change per se does not necessarily mean an impact on predictive performance, and mis-specified models can be as effective as appropriately specified models in some situations. This is nicely illustrated by the fact that Naïve Bayes models (also known as *Idiot's Bayes* models) often yield excellent classification performance, even though they give biased estimates of class membership probabilities (see Hand and Yu, 2001). See also Qian (2020).

On the optimistic side, there is evidence that predictive models can be resilient in many ways. The flat-maximum effect suggests that a broad range of modelling parameters based on fits to changing distributions often differ very little in terms of model fit and will lead to similar predictive outcomes (Overstreet et al. 1992).

2.3 Types of change that will affect the predictor

A central question is what type of distributional changes will affect how P performs. There may be dramatic changes in distribution which have limited impact on predictive performance as well as small changes which have dramatic effect.

Suppose, for example, that P's job is to predict some feature y from a vector of features $\mathbf{x}$.

We can think of **marginal distributional change** in $\mathbf{x}$ or y. These changes will affect P since values of $\mathbf{x}$ and y that are under-represented in H will lead to less reliable predictions for those values. If those values become more frequent in F, then we will see P's overall performance decline overall. Additionally, if P is constructed on a model, the more mis-specified the model is, the more risk of error on under-represented values: another aspect of model risk. But can we quantify and forecast this decline in performance in some way? We will describe augmented testing below as one way to answer this.

We can also consider **conditional distributional changes**. In particular, $y|\mathbf{x}$. Such changes are much more problematic for P, since it means a change in the relationship that P is modelling. The hope here is that we can anticipate some of the changes in relationships between predictors and predictand, either through an understanding of the application area or perhaps through trends in H (e.g. using time-varying coefficients in models of H; see Section 5.6).

For a valuable discussion of these different types of perspectives, in the context of medical diagnostics, see Dawid (1976).

Another type of change is when we stop receiving relevant data. An example of this can be found during the COVID pandemic, where certain items needed for calculating the RPI can no longer be collected. For example, surveys to collect prices are not being conducted (so straightforward missing values) and in other cases, such as restaurant meal prices, the things have ceased to exist - you can't get a restaurant meal, so should it be included in the basket at all for RPI calculation? This is a particularly problematic type of change that can impact a predictor and can be for a wide variety of reasons, not just changes in human behaviour, ranging from instrument failure to changing definitions. The challenges arising from such *dark data* are described in Hand (2020), along with methods for tackling those challenges.

2.4 Related statistical and machine learning topics

The problem of robustness to changing distribution is related to the problem of *population drift* or *concept drift* (see e.g. Kelly et al, 1999; Adams et al, 2010; Webb et. al. 2016). Typically, population/concept drift refers to gradual drift. However, we could also mean an abrupt change. Indeed, although this article considers predicting through a crisis, much of the discussion and methods proposed are relevant for any type of abrupt change, such as a sudden change in consumer behaviour due to the introduction of new technology or infrastructure (e.g. online grocery shopping).

The problem is also related to *domain adaptation* (see e.g. Redko et. al. 2020). In the field of machine learning for computer vision, it has been found that systems developed on

one set of image data do not always translate well to another (Farhadi and Tabrizi 2008). Hence, in general, domain adaptation is the problem of developing an AI system under one domain (we would say distribution), but applying to another.

Techniques from studies of population drift and domain adaptation may prove useful for tackling robustness of AI systems to a crisis.

3. Monitoring

3.1 Dynamic performance monitoring and Dashboards

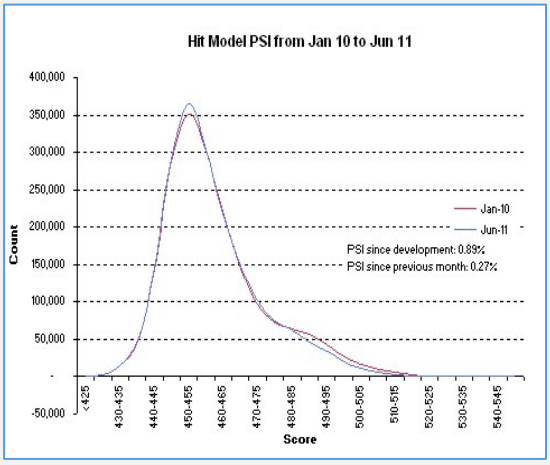
A fundamental starting point is to build and deploy AI systems enabled with an interface to allow monitoring of performance over time. This way it is possible to manually check the AI dynamically and pre-empt any degradation or shift in performance. The performance measures used should be considered carefully and should fit with the required decision-support task of the system. Some common measures for regression are mean square error (MSE), mean absolute error (MAE) and mean percentage error (PE). Some common measures for classification are accuracy, the Kolmogorov-Smirnov (KS) statistic and the area under the ROC curve (AUROC or AUC)[2]. As discussed earlier, two common measures of distributional change are divergence measures and population stability index (PSI). It is also important to consider measures specifically designed around the application; e.g. cost-based loss measures or using utility functions. A good discussion of business-specific measures can be found by Fawcett and Provost (2013). It is advisable to use an array of different measures, rather than relying on just one and using a dashboard helps visualize this information.

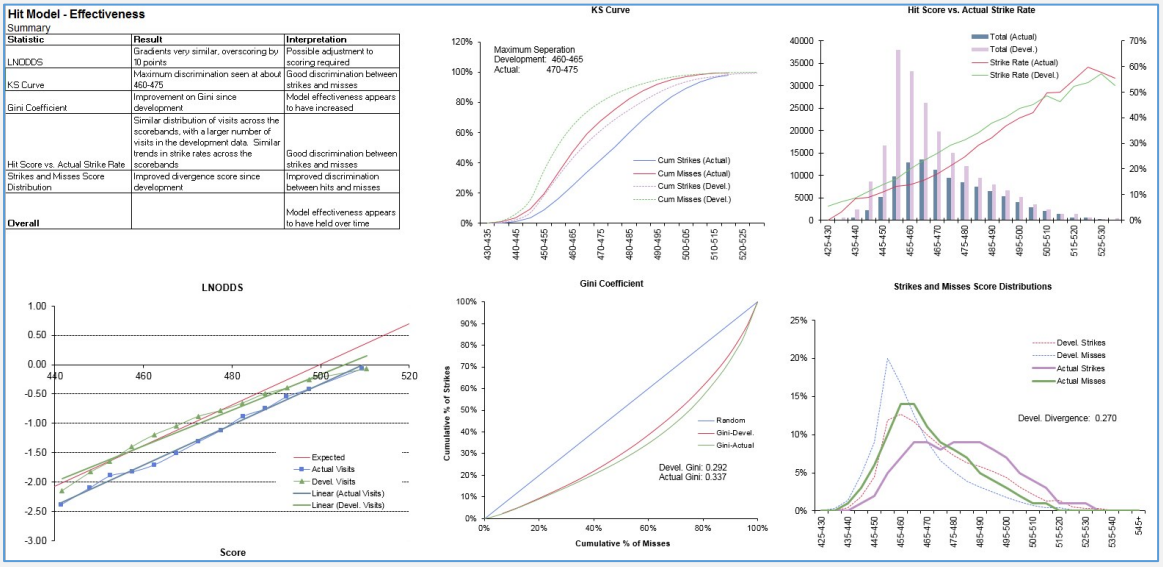
An example of a dashboard for monitoring is shown in Case Study 1.

[2] The AUC is a popular measure both in industry and academia, along with Gini which is a linear transformation of AUC. However, it is worth mentioning that AUC has a fundamental flaw in being an incoherent measure and the authors do not recommend its use. Alternatives such as H-measure or KS statistic are available. For further information see Hand (2009) and Hand (2010).

Case Study 1: Monitoring AI using a Performance Dashboard

A dashboard of diagnostics would be highly recommended in monitoring AI model performance. For example, a bank would typically monitor the performance of its classification models to predict personal loan default and develop such a system to help with this over time. This approach would be applicable in many other settings. Tax authorities is another example, used to maintain AI model performance to predict taxpayer non - compliance. Here are two example outputs from the dashboard user interface:





The first figure shows change in distribution of scores output by the model over a six-month period. This shows these are largely unchanged but with a small reduction in high scores over time. The second figure shows a range of measures of performance over time for a classifier including KS and Gini for class discrimination, and log-odds (LNODDS) for probability calibration. In particular, the log-odds graph suggests a shift in the bias of probability estimation over time which may require further investigation.

4. Approaches to evaluating AI Systems through a crisis

4.1 Augmented Testing

Re-evaluate predictive algorithms taking account of the expected new distributional structure. The predictive algorithm would have been developed under H, but data from this distribution can be reweighted to emulate the distribution F^* (our expectation of F). In particular, we consider reweighting of historical observations in test data to allow for a different marginal distribution over the predictor variables $\mathbf{x}$ so as to be closer to F^*.

+ Advantage: Allows the performance of the predictive algorithm to be assessed under novel conditions, so we see how good or bad it is under those conditions.

- Disadvantage: Augmented testing depends on how accurate our expectation F^* of F is.
- Disadvantage: The accuracy of the test depends on how much past data (from H) looks like F^*. This can however be measured using *effective sample size* (ESS) (Kish 1965) and this can be factored into the measure of our uncertainty in P.
- Disadvantage: Since the goal is to reweight so that the marginal distribution over predictor variables $\mathbf{x}$, this does mean that this method will not be affective for conditional changes in the relationship between $\mathbf{x}$ and the outcome variable y.
- Condition: Augmented testing will not work if the data change is due to a change in the way data is measured (e.g. measuring unemployment using a different formula). This type of change would need to be dealt with by a data transformation / conversion between the two measurement types, if that is possible.

The term *augmentation* has different senses in statistics. The term as used here refers to the augmentation method to re-weight observation from one distribution to match that of another, as used within credit scoring models to adjust for selection bias (Crook and Banasik, 2005).

4.2 Scenario testing

The augmentation approach can be refined to test a model based on past data for some specific scenario, rather than the whole F^* (for example, we might want to know how the model performed just on people who were unemployed if we are expecting unemployment rate to rise). A novel approach to scenario testing is given, eg, by Pesenti et al. (2019).

+ Advantage: Allows focus on particularly important or sensitive cases.
+ Advantage: Does not require a projection F^* of F.

- Disadvantage: Requires sufficient historical data that looks like the scenario, but again ESS can be used to measure this.

Even if insufficient data is available for a particular scenario, projections of predictions could be made from similar scenarios along with error on predictions. Consider the

simplified example in Figure 1. This shows estimates for some factor for different income bands. If we do not have sufficient data to infer an estimate for income band 96K+, projecting the estimates from the graph may suggest a point estimate around 10.2 with confidence intervals +/-3.0.

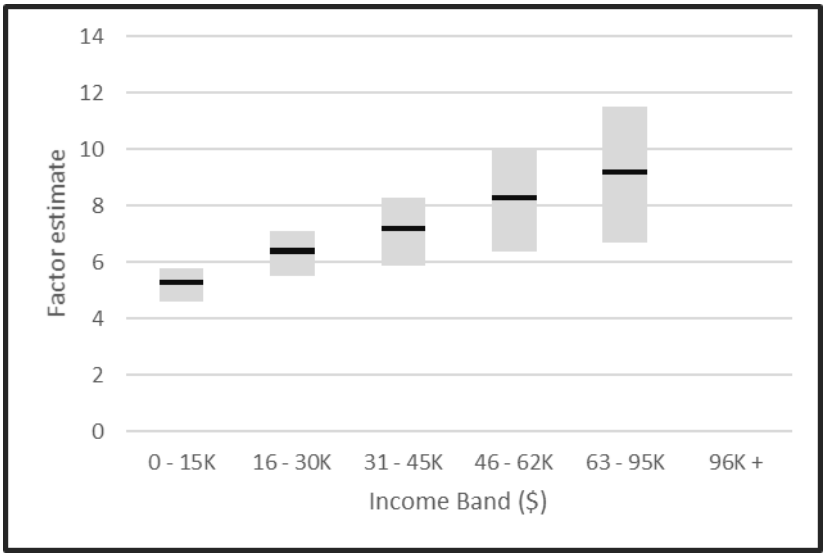


Figure 1. Factor estimate by income bands (simulated data). The black line is a point estimate and the grey bands are 99% confidence intervals.

Example: Augmented and Scenario Testing

Suppose our original distribution H includes only 4% unemployed people. After or during a crisis we expect (F^*) this will increase to 12%. Then apply **augmented testing** by reweighting all unemployed people in the Test Data by 3, relative to employed people, when computing the performance measure. The following table shows results. We can see that with the augmented test where we anticipate a new distribution, the mean absolute error rises to 0.148 from 0.116.

Example	**Employed**	**Unemployed**	**Test under H** (unaugmented)	**Augmented test, under F^***
N	960	40	960 : 40	880 : 120
Mean Absolute Error	0.1	0.5	0.116	0.148

The mean absolute error of 0.5 recorded in the 3rd column is a simple **scenario test** for data covering just unemployed people, demonstrating how poorly this model performs for this group (relative to the employed group).

4.3 Metamorphic Testing

Test the model with extreme cases, such as scenario testing, and determine if the results are sensible based on some given rules (e.g. if tax revenue is being predicted and the model predicts a negative value under some circumstance, this suggests some problem with the model). See Xie et al. (2009) and Segura et al. (2018).

+ Advantage: Does not require outcomes for test cases and if a predictive algorithm passes the metamorphic test it suggests it is behaving within realistic bounds.
+ Does not require projection F^* of F.

- Disadvantage: Requires a prior set of rules for the test. Expert judgement can be used for this purpose.

4.4 Stress Testing

Project how the model will behave and predict in future scenarios (e.g. a drop in GDP). This is slightly different from the scenario testing described above since scenario testing is based on historical data. In contrast, a stress test is a projection forward with new data (see eg Bellotti and Crook 2013 and Breeden 2016).

+ Advantage: Gives us a useful indication of how future outcomes will look under different scenarios, such as crisis or downturn period.

- Disadvantage: Assumes the underlying model is reliable and is suitable for projecting outcomes for new scenarios. However, augmented and scenario testing can be conducted to test the reliability of the model. Additionally, measuring uncertainty in the prediction can be used to adjust the outcome projected by the predictive algorithm to allow for *conservatism*. For example, a simplified example: if in some circumstances, a model predicts tax revenue from a particular source to be $10 million but our measure of uncertainty suggests an error of up to +/-20% then it may be preferable to provide a conservative prediction of $8 million.

- Disadvantage: The results of a stress test cannot be empirically tested (since the outcome, typically, is never measured). However, metamorphic testing and expert judgement can be used to ensure projected outcomes are sensible.

Case Study 2: Stress Test of Dairy Lenders in New Zealand

New Zealand has a large dairy farming sector and financial institutions lending to dairy farms are vulnerable to changes in the dairy market. Therefore in 2015, the Reserve Bank of New Zealand conducted a stress test to determine dairy lenders' vulnerability (RBNZ 2016).

Historic data could be used to associate default with break-even price for milk and loan-to-value ratio (LVR), which form the basis of a credit rating model. Two scenarios were considered to explore plausible downturn scenarios in the dairy market: scenario 1 involves a 20% drop in land prices and a market value for milk below $5.25/kgMS (milk solid) until 2018; scenario 2 was more severe with a 40% drop in land price and market value for milk remaining below $5.25 until 2019. These two factors affect the credit rating, based on the historic association with break-even price and LVR. Hence through the course of each scenario, the distribution of credit ratings changes with these changing economic values, leading to projected default rates reflecting these changes.

The figure below illustrates average credit rating projecting dairy loans through each scenario (grades CCC, B, BB and BBB represent high to lower relative risk).

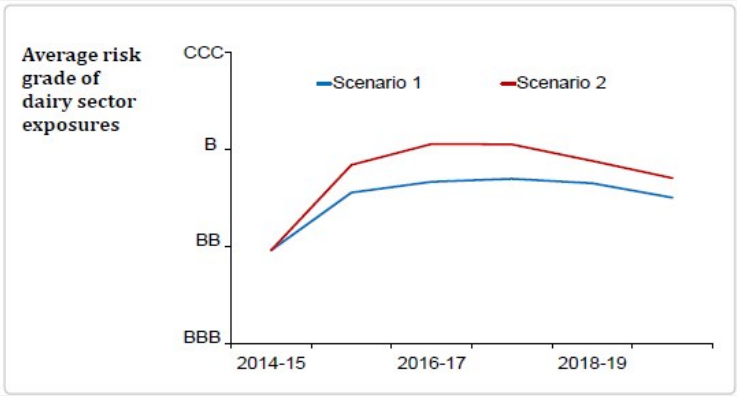


Figure 2. Stress test results taken from RBNZ (2016).

The outcome of the stress test was that scenario 1 projected to generate 3% bad debt expense (as percentage of exposure), whilst the more severe scenario 2 projected to generate 8% bad debt expense. Knowing this helped the Reserve Bank and the dairy lenders plan ahead, in particular in setting aside sufficient capital reserves.

For stress testing, it is important to consider the uncertainty inherent in the model. After all, the model is itself just an estimate of reality. During normal conditions, the effect of this uncertainty may be minimal. However, during a crisis, with a significant change in distributions, model uncertainty may exacerbate poor predictive performance. In a recent study, Wang et al. (2020) used a Bayesian approach, treating model parameters as random objects, showing that extreme risk measures (e.g. 99% VaR) is extremely under-estimated if model uncertainty is not taken into account.

5. Remedial solutions for AI Systems through a crisis

Once we have evaluated an AI system, we can consider alternative remedial solutions depending on how well or badly the system is expected to behave during the crisis and resources available. Several responses are described below.

We may be anticipating a crisis, in the middle of one or coming out of one. If we are entering a crisis period it certainly makes sense to take some remedial action. However, if we have already gone through and are coming out of a crisis, *the horse has already bolted* and perhaps it is too late to take action. However, it is important to recognise that coming out of the crisis may lead to another distributional change ($H \rightarrow F \rightarrow G$) and we should also anticipate and adapt to these new conditions.

5.1 Expert judgement

We can expect models to become less reliable over time, hence we can draw on the judgement of experts who work in the application field to help determine how we may expect future trends to play out.

+ Advantage: Experts will have broad experience across time and can apply their knowledge to determine future expectations.
+ Advantage: It is less important to have an accurate estimate F^* of F as some of the other analytic methods that follow. A human expert can make a decision with less precise expectations.

- Disadvantage: Experts can be subjective and do not always agree, or may miss the consequences of new developments.

This approach can be supplemented by economic models and psychological models to provide some objective content to expert decisions.

+ Advantage: Economic models are theory-driven models rather than empirical models, so may be less vulnerable to changes in data distribution.
+ Experts and/or economic models can be hybridized with existing empirical predictive models (e.g. Hand et al. 2008).

- Disadvantage: May miss changes in behaviour due to economic regime shift or major societal changes.

A popular approach is to apply **conservatism** to automated decision making. For example, when assessing individuals for credit, set a more risk-averse threshold than would normal be applied based on a credit risk model output. This approach is especially appropriate when there is a lack of good information (i.e. F^* cannot be estimated reliably) and, in particular, during the uncertain initial phases of a crisis. For example, considering the effects of COVID-19 on its credit card users and projecting a large drop in income, Capital One in the USA cut credit limits, making a conservative decision and essentially over-riding their automated credit limit algorithms (Picchi 2020).

5.2 Revert to simpler statistical models or use bagging

Simpler models with less model structure and fewer parameters will be more robust to changes; whereas complex models may be more vulnerable. Simpler models will express general features and be less misled by local patterns in H.

+ Advantage: Potentially more robust in the presence of change.

- Disadvantage: May be insufficiently rich to capture functionality of the predictive system. It will underperform a more complex algorithm, at least on the given data at the time of model development.

Some complex models may actually be more resilient under change. In particular, bagging and ensemble approaches may be robust since they average across several underlying predictors. Therefore, this leads to another strategy which is to use bagging or ensemble models to supplement existing predictive algorithms.

For further discussion and experiments in the context of Deep Learning and ensembles, see Qian et al. (2020).

5.3 Augmentation

If the existing predictive algorithm is quite poor, it may require a rebuild. Typically this will involve retraining a model. Typically only historical data, from H, will be available. However, the distribution can be reweighted to resemble the target expected distribution F^* of F (Crook and Banasik 2005), rather like with augmented testing described earlier.

+ Advantage: We may hope that the model will give better predictive performance under the new future distribution following this update.

- Disadvantage: Requires a rebuild of the model.
- Disadvantage: This depends on how accurate our expectation F^* of F is.
- Disadvantage: The training depends on how much past data (from H) looks like F^*. This can however be measured using *effective sample size* (ESS) which can be factored into our understanding of the uncertainty in the predictive algorithm.
- Caution: As with augmented testing, this will not work if the data change is due to a change in the way data is measured.

5.4 Use of proxy data

If insufficient data is available (or low ESS) for any of the methods outlined above, external data can be used as a proxy; e.g. in testing or model build. For example, if we want to test a novel GDP fall of 2% or more and this has already occurred in another country or region, then data can be *borrowed* from this region to supplement native data. This approach is also referred to as *knowledge transfer learning* within the domain adaptation community (Redko et. al. 2020).

+ Advantage: More data means more predictive power. Essentially, the predictive algorithm is learning from experiences in other locations.

- Disadvantage: This relies strongly on supposing that the proxy data is like the native population we are predicting on; i.e. that the distribution of the proxy data is sufficiently similar to H or F in some sense. For example, if I wish to predict credit risk amongst people in Massachusetts, it might make sense to use a proxy data set from New York, which is geographically close; but not from New Delhi.

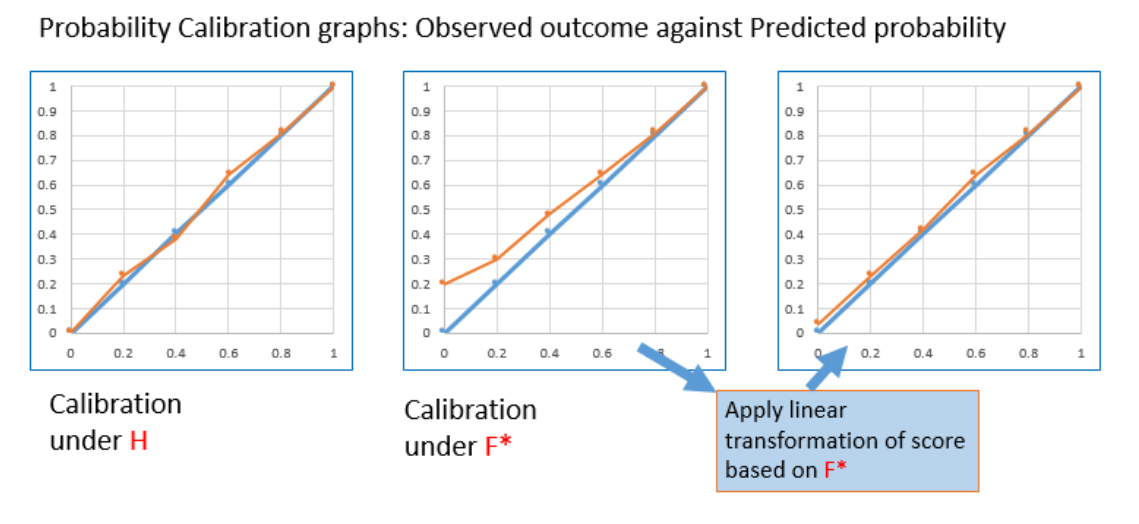
A good example of the use of alternative data was given by Church and Lucas (2020) who discussed economic forecasts through the COVID-19 crisis in UK. The official unemployment figures from the Office of National Statistics are lagged and hence unemployment affects due to the COVID-19 lockdown were not immediately apparent. Instead, HMRC payroll data was used which is up-to-date and hence does reveal unemployment effects due to COVID-19 lockdown.

5.5 Recalibration

Recalibration is already a popular technique to adjust models in the light of changing distributions. If a model rebuild using augmentation, or other method, is not plausible then the model can be recalibrated by applying a simple rule to adjust the output of P (typically a linear rule), so that the average outcome across different risk groups matches reality under the new distribution F. It is illustrated by the following example.

Recalibration of probability estimates

The following figure shows probability calibration. For different risk groups, the red graphs show average estimated probability of outcome (x-axes) against observed outcome (y-axes). The closer these are, the more reliable the predictor. Hence the diagonal blue line shows the target of a perfectly behaving predictor. We see how predictor P behaves under H (left) and how it behaves under F^* (centre) with a clear shift to under-estimating outcomes. We apply some transformation to the output of P and get a better calibrated outcome (right).



Recalibration is useful if we are interested in the probability estimates given by the predictor. If we are only interested in scoring or rank ordering of observations it will not have any impact. In particular, for applications where the predictor is used to make decisions for observations giving a score above a certain threshold, recalibration will not have an impact; e.g. in automated loan application approval. Also, probability calibration in itself does not guarantee a good model. For example, suppose I have a data set of observations where 80% of them are type A and 20% are type B. If I have a very simple predictor that outputs probability 0.8 for being type A for *all* observations, it is perfectly calibrated but it will be useless for classification. Therefore, probability calibration should not be a stand-alone measure: a suite of measures are needed (see Case Study 1).

5.6 Time-varying models

Many of these methods may be good for dealing with unconditional distributional change (i.e. on $\mathbf{x}$ or y), but may not be so successful on conditional distributional change ($y|\mathbf{x}$). Proxy data may be useful for this, since some external data may already reflect future changes in risk factors. A more ambitious method is to model the predictive algorithm's parameters over time and then project their development into the future. Such statistical methods exist, such as the use of time-varying coefficients in models (Djeundje and Crook 2019).

+ Advantage: Allows projection of changes of risk factors into the future scenarios.

- Disadvantage: Assumes that the transition from H to F is smooth and such projection can be made. If the change is expected to be abrupt, the change may not be possible to project.
- Disadvantage: Requires sophisticated modelling techniques and historical data over a long time period.

6. Summary

We have discussed the problem of deploying predictive algorithms through a crisis, evaluating performance and robustness and suggested several remedial methods in anticipation of performance degradation. Solutions need to consider changing data feeds impacting the relevance of existing model features and weights and to explore appropriate re-alignment, re-build or replacement options. We would advocate a three-pronged solution to a live AI system:

1) The monitoring of existing system performance
2) Remedial actions to address AI system instability
3) Checking for robustness to mitigate future risks

Careful consideration also needs to be taken in the commissioning of new AI systems as the condition that the '*future will resemble the past*' (the inductive principle or the "uniformity of nature") is an assumption and we may find that pre- and post- crises such as COVID-19, data may well be substantially different.

This paper has, of course, been motivated by the COVID-19 crisis. Note, however, that the observations made in this paper are not the full story and there other facets of the problem of predicting through a crisis that we have not been able to explore. We are learning more as the crisis develops and affects users of predictive algorithms. We hope

the paper can form a point for further discussion and we welcome comments and feedback. We also hope that this paper will lead to further engagement with interested practitioners and academics in the form of workshops and webinars.

Acknowledgements

We would like to thank the many people who contributed ideas and feedback for this paper. In particular, we thank Mathias Kern and Michael Mortenson who took part in a lively panel session on predicting through a crisis during the Operational Research Society online conference 62.

References

- Adams, N.M., Tasoulis, D.K., Anagnostopoulos, C. and Hand, D.J. (2010) Temporally-adaptive linear classification for handling population drift in credit scoring, Lechevallier, Y. And Saporta.(eds), *COMPSTAT2010, Proceedings of the 19th International Conference on Computational Statistics*, Springer, 167-176.

- Bellotti A. and Crook J. (2013), Forecasting and stress testing credit card default using dynamic models, *International Journal of Forecasting* Volume 29, Issue 4, October - December 2013, Pages 563 - 574

- Black R., Tsanakas A., Smith A.D., Beck M.B., Maclugash I.D., Grewal J., Witts L., Morjaria N., Green R.J. and Lim Z. (2018), Model risk: illuminating the black box, British Actuarial Journal, Vol. 23, e2, pp. 1–58.

- Box, G. E. P. and Hunter W. (1965), The experimental study of physical mechanisms, Technometrics, 7, 57-71

- Breeden, J.L. (2016), Incorporating lifecycle and environment in loan-level forecasts and stress tests, European Journal of Operational Research, Volume 255, Issue 2, 1 December 2016, Pages 649-658

- Church, K and Lucas, L. (2020) The Analytics behind Economic Forecast Scenarios for Credit Risk Modelling, OR62 Opetational Research conference, Online, 16th September 2020

- Cox, D. R. (1990), Role of models in statistical analysis, Statistical Science, 5, 169-174

- Crook, J. and Banasik, J. (2005), Does reject inference really improve the performance of application scoring models?, Journal of Banking & Finance, Volume 28, Issue 4, April 2004, Pages 857-874

- Dawid A.P. (1976), Properties of diagnostic data distributions. Biometrics, Vol. 32, No. 3 (Sep., 1976), pp. 647-658

- Djeundje V.B. and Crook J. (2019), Dynamic survival models with varying coefficients for credit risks, European Journal of Operational Research, Volume 275, Issue 1, 16 May 2019, Pages 319333

- Farhadi A., Tabrizi M.K. (2008) Learning to Recognize Activities from the Wrong View Point. In: Forsyth D., Torr P., Zisserman A. (eds) Computer Vision – ECCV 2008. ECCV 2008. Lecture Notes in Computer Science, vol 5302. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-88682-2_13

- Fawcett T. and Provost F. (2013), *Data Science for Business*, O'Reilly Media

- Hand D.J. (2009) Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, **77**, 103-123.

- Hand D.J. (2010) Evaluating diagnostic tests: the area under the ROC curve and the balance of errors. *Statistics in Medicine*, **29**, 1502-1510.

- Hand D.J. (2020) *Dark Data: Why What We Don't Know Matters*. Princeton University Press.

- Hand D.J., Brentnall A., and Crowder M.J. (2008) Credit scoring: a future beyond empirical models. *Journal of Financial Transformation*, **23**, 121-128.

  Hand D.J. and Yu K. (2001) Idiot's Bayes – not so stupid after all? *International Statistical Review*, **69**, 385-398.

- Kelly M.G., Hand D.J., and Adams N.M. (1999) The impact of changing populations on classifier performance. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ed. S.Chaudhuri and D.Madigan, Association for Computing Machinery, New York. 367-371.

- Kish, L. (1965), Survey Sampling, New York: Wiley

- Lehmann, E. L. (1990), Model specification: the views of Fisher and Neyman, and later developments, Statistical Science, 5, 160-168

- Overstreet, G.A., Bradley Jr, E.L. and Kemp Jr, R.S. (1992), The flat-maximum effect and generic linear scoring models: a test, IMA Journal of Management Mathematics, Volume 4, Issue 1, 1992, Pages 97–109

- Pesenti S.M., Millossovich P. , Tsanakas A. (2019), Reverse sensitivity testing: What does it take to break the model?, European Journal of Operational Research 274 (2019) 654–670.

- Picchi, Aimee (2020), Capital One cuts credit limits as millions struggle with income cliff, CBS News, August 28th 2020 (www.cbsnews.com/news/capital-one-lowering-credit-card-spending-limits).

- Qian S., Kalimeris D., Kaplun G. and Singer Y. (2020), Robustness from Simple Classifiers, https://arxiv.org/pdf/2002.09422.pdf, 21 Feb 2020

- Redko, I., Morvant, E., Habrard, A., Sebban, M., & Bennani, Y. (2020). A survey on domain adaptation theory. ArXiv, abs/2004.11829.

- RBNZ: Reserve Bank of New Zealand (2016) Bulletin, Vol. 79, No. 5 March 2016

- Segura S., Towey D., Zhou Z.Q. and Chen T.Y. (2018), Metamorphic Testing: Testing the Untestable, IEEE Software, December 2018, PP(99):1-1

- Xie X., Ho J., Murphy C., Kaiser G., Xu B., and Chen T.Y. (2009), Application of Metamorphic Testing to Supervised Classifiers, Proc. Int. Conf. Qual. Softw. 2010 January 15; 2009: 135–144.

- Wang, Z., Crook, J., and Andreeva, G. (2020), Reducing estimation risk using a Bayesian posterior distribution approach: Application to stress testing mortgage loan default, European Journal of Operational Research, Volume 287, Issue 2, Pages 725-738.

- Webb, Geoffrey I; Hyde, Roy; Cao, Hong; Nguyen, Hai Long; Petitjean, Francois (2016), Characterizing concept drift, Data Mining and Knowledge Discovery; New York Vol. 30, Iss. 4, (Jul 2016): 964-994.

- White, L. and Cruise, L. (2020), Model misbehaviour - coronavirus confounds bank risk systems, Reuters 22 June 2020.

- Wu, D. and D L Olson, D.L. (2010), Enterprise risk management: coping with model risk in a large bank, Journal of the Operational Research Society, Volume 61, 2010 - Issue 2

Contact us at: contact@validateai.org
Second white paper enquiries to: Anthony-Graham.Bellotti@nottingham.edu.cn
Follow us on twitter: @Validate_AI, #ValidateAI
Visit our website: https://validateai.org