How to ensure AI and Machine Learning scores are fair and ethical

28 November 2018

The author
Martin Benson
Head of Artificial Intelligence

The decision of whether to grant credit to an applicant is often life-changing to that individual. It determines whether or not they can buy their dream house, upgrade their unreliable car or even buy a washing machine. In making decisions about who to lend to, credit providers bear a huge responsibility to ensure that they're doing so in a fair and responsible manner, in addition to ensuring profitable outcomes.

What's more, with the rise of AI and machine learning in credit scoring, how can we ensure lending decisions follow common sense trends and are fair and ethical? In this blog, Martin Benson explores these questions to help you make the right lending decisions.

Back in 2008, an online credit card provider found themselves on the sharp end of this issue when they withdrew credit from 160k customers, many of whom had excellent credit profiles. The communication that went along with the cancellation stated that it was intended to limit the spending of people that they identified as being high credit risk, and directed them to a credit reference agency to check their status. But, many people did that and confirmed what they already knew – their credit profile was excellent, and yet they had still been targeted. Inevitably, they complained in strong terms, drawing the attention of MPs who called for them to be referred to the Office of Fair Trading and the Financial Services Authority.

Wherever a strange decision is made that inappropriately penalises a customer, they are highly likely to complain about it (justifiably so) and it needs a cluster of only a few such instances to draw the attention of the media and of regulators.

For decades, lenders have used regression models to automate and optimise their lending decisions via credit scores, in a way that is consistent with that responsibility. There are two main strands to how they have achieved that:

1. Ensure that the models used are robustly estimated and can be shown to be effective when applied to new data, and not merely on the dataset that was used to develop the model. This is key to ensuring that in aggregate the decisions that will be made using the model will be of the expected high quality when it is used in practice.

2. Ensure that the points that are assigned to each variable in the model follow an intuitive trend. For instance, people who have had prior CCJs should receive fewer points than people that have not. These rules ensure that – all other things being equal – the influence of each data item in generating the decision corresponds with common-sense expectations.

The second of these points is key to ensuring that the scores that are assigned are defensible and fair. Consider two applicants whose data is identical in all regards other than the fact that Applicant A has had prior CCJs, but Applicant B has not. Would it be fair to assign a lower score to Applicant B and decline them while accepting Applicant A? Clearly not, and ensuring that the response to each input variable follows an intuitive trend is the only way to avoid it.

There is frequent confusion about what is important in this regard - the **marginal relationship** between a variable in a model and the model output or the **average observed relationship** between the variable and the model output (or the outcome variable) measured on some dataset.

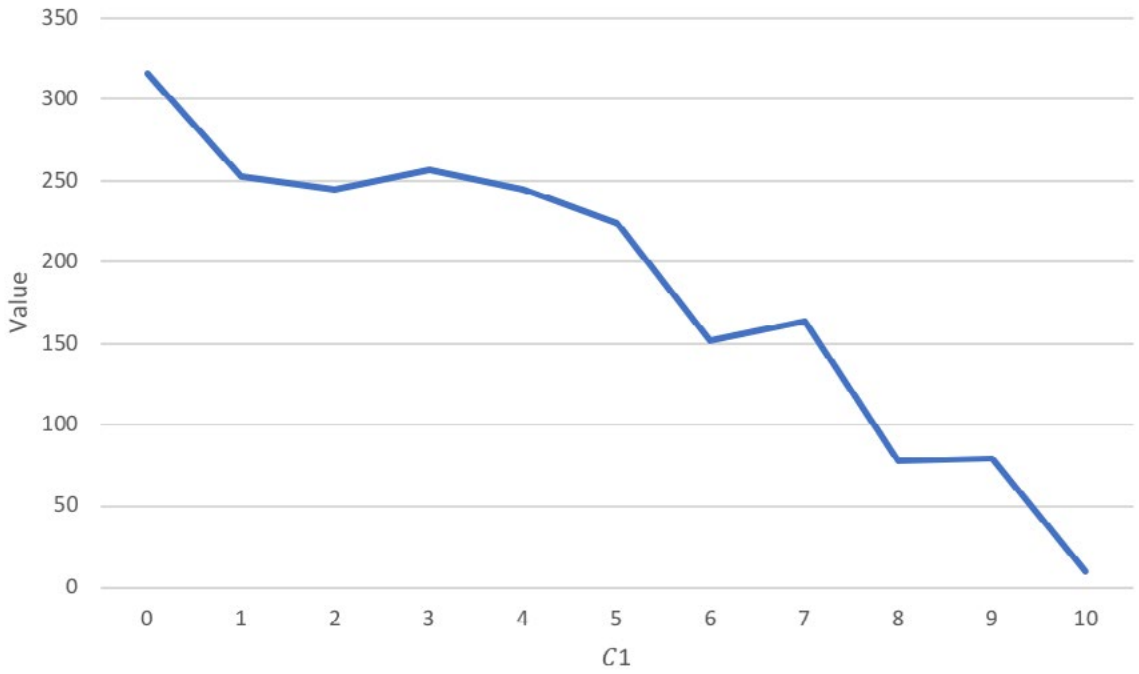
The **first** of these is the type of the relationship that we've already covered, and have noted to be critical – if the variable changes in isolation and all other variables are held constant, what happens to the model output.

The **second** type of relationship simply measures the average value of the model output (or the outcome variable) across the range of the variable. This second type of relationship is totally irrelevant to assessing the fairness of the model. A simple example will make this clear.

Imagine that I keep track of the coins in my pocket at the end of the day, every day for a year and let's suppose that there are only ever pound coins, 10p coins and 1p coins. I'm interested in measuring the total value (in pence) of the coins in terms of the number of each type of coin. This is clearly straightforward:

$$V = 100C_{100} + 10C_{10} + C_1$$

If you built a regression model for V in terms of C_{100}, C_{10} and C_1 it would have no trouble in producing the expected relationship (and in fact will be able to achieve 100% accuracy in this simple example). What happens, though, if we look at the average output split by C_1?



Uh? The trend doesn't make any sense - average value generally goes down as the number of coins goes up! What is going on? Well, the answer is straightforward – my pockets are only so big and only so many coins fit in there. This means that in the data there is a negative correlation between C_{100} and C_1 – the more pennies there are in my pocket, the fewer pounds there are, generally speaking.

The average observed relationship (shown on this chart) - with either the model output or the outcome variable (the example above covers cases both since this model is 100% accurate) - is not a useful guide as to how a variable should enter a model, because it is influenced by correlations that exist in the data.

In that first example it turned out that we got a perfect model, despite the strong correlations between input variables, but generally we aren't so lucky. In fact, if we attempt to model V in terms of only C_{100} and C_1 using the same data (let's say the C_{10} column got corrupted somehow and I can't use it any more) we see how things can go wrong. If you do that, the regression equation that you get is:

$$\hat{V} = 97.25C_{100} - 2.38C_1 + 36.89$$

Clearly that is not a good model of reality – it says that for every penny I put in my pocket, the value of the coins in there goes down by 2.38p. It doesn't. The negative sign in the parameter estimate is consistent with the negative trend in average actual outcomes (and model output), but that is not a signal that the model is reliable – because, well, it isn't reliable. If you had a lot of pennies in your pocket (or a lot of pounds for that matter) you wouldn't be happy to trade me your coins for a cheque for $\hat{V}$. And, while the model above is optimal in a statistical sense on the data it was trained on, I wouldn't expect it to work very well on new data gathered by someone who's pockets were a different size to mine – it's optimal only where correlations remain the same.

The example[1] is somewhat artificial, in the interest of simplicity, but the issues that it surfaces occur in every credit model development, and the only correct course of action is to ensure that the marginal relationships that the model expresses are sensible. To summarise, the consequences of not doing that are:

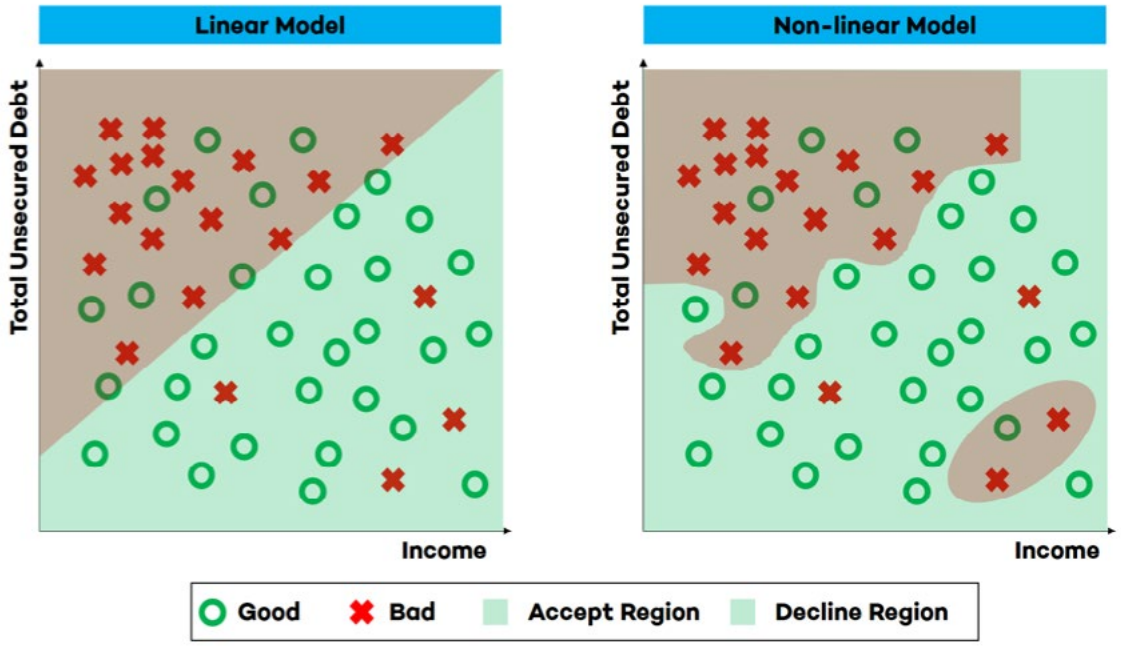
1. The outputs assigned will arguably be unfair in some cases – because an applicant may be declined despite their data being better in every dimension than another applicant who scored higher and was accepted.

2. The model is likely to perform worse when applied to new data that differs somewhat from the development data, because it relies on correlations between variables to correct for marginal relationships being wrong, and as those correlations shift over time, model performance will degrade.

That's why credit modellers have done it for decades.

While the discussion above was framed in terms of linear models it applies equally strongly to non-linear ones. But, things are more complicated there because marginal relationships cannot simply be read off the model specification and will typically vary in nature from applicant to applicant, making it harder to ensure that marginal relationships in the model are always sensible. It also significantly increases the risk that odd decisions might be made, because the decision boundary of a non-linear model can be much more complex than for a linear one. Let's again consider a simple example, for a hypothetical model that involves just two variables.

1. If you wish to explore this hypothetical data further, you can find it here: https://drive.google.com/file/d/1y1eW-vEQXrEivy8vORroezGG-S0ryye1/view?usp=sharing



In a linear model, the decision boundary is linear (by definition) and so any unusual incidences where applicants with an excellent credit profile go bad do not have a strong influence on who is accepted by the model and who is not – their impact is generally restricted to nudging the decision boundary by a small amount. A non-linear model, though, is free to draw a little circle around such cases, with the result being that on an ongoing-basis a group of applicants with an excellent credit profile will be declined by the model. For cases like those, it will be very difficult to justify why they were declined when they inevitably complain about the decision. Certainly, the generic wording that is generally employed will not cut the mustard: "Our decision making takes into account a variety of factors, including the information you supplied on your application and credit reference agency information".

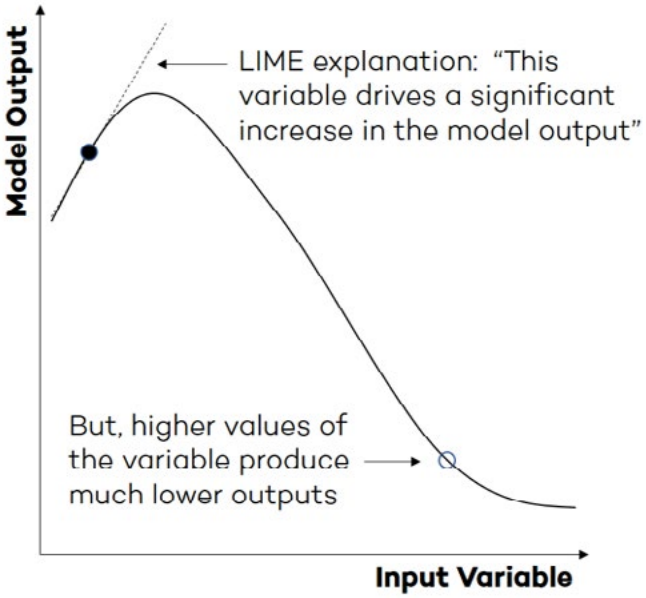
When you know that every decline decision you make must be associated with at least one piece of derogatory information it is an appropriate response. When you cannot be sure of that, it is not. In a model (even a non-linear one) where every variable is forced to adhere to sensible marginal relationships in every case you can be sure of it.

Whenever an applicant is declined it must be because they received a negative contribution from some variable in the model, and the reason that the contribution was negative will stand to reason. This means that even for non-linear models, if you are able to ensure that all marginal relationships are intuitive then you can be confident that the decisions that the model makes on your behalf are defensible.

A common approach to "explaining" the outputs of a non-linear model is to use the LIME algorithm[2]. While the LIME approach is undoubtedly useful in demystifying black-box models in general, it is not sufficient for ensuring that a non-linear credit score is fit-for-purpose. There are two main reasons for this:

1. It can only produce "explanations" after the fact – after the model has generated a prediction. Moreover, generating the explanation is computationally expensive and producing explanations in near real-time (in the hope that you could intervene and override the decision if you did not like the "explanation") at scale would be problematic at best.

2. The "explanations" that it produces merely quantify approximately how the model behaves in the local vicinity of the case in question[3]. Again, a simple example will help to explain why that is not a good basis for explaining a credit decision, (which is why I've been calling them "explanations").



The LIME explanations quantify the sensitivity of the model output to a variable in the immediate vicinity of a case (as indicated by the dotted line in the diagram). In this example the LIME explanation will say that the model output strongly increases as the variable increases, implying that a low value of that variable will weigh negatively on it. But that need not be the case at all - in general the global behaviour can be totally different to the local behaviour (indeed the inventors of the approach make this point explicitly in the paper where they introduced the technique[4]). In this example, higher values of the variable are associated with much lower model outputs. Again, this deliberately simple example may seem artificial, but some form of this effect is likely to exist in any real-world non-linear model that is produced.

2. https://homes.cs.washington.edu/~marcotcr/blog/lime
3. They describe an approximate tangent hyper-plane to the prediction surface of the model, if we're getting geeky about it.
4. http://arxiv.org/pdf/1602.04938v1.pdf

While non-linear modelling approaches can undoubtedly deliver more powerful credit models than more traditional approaches, they must be adopted with care, and ensuring that their outputs are defensible in all cases is a challenging problem. Solutions do exist though, such as our award-winning predictive modelling tool Archetype[5].

In summary, it's essential that existing high standards of care regarding the behaviour of credit scoring models are upheld - credit assessment cannot be allowed to descend into a 'computer says no' scenario where the decisions that are being made are not properly understood. The decisions that are being made are too important not to.

As other industries begin to adopt machine learning to streamline decision making they should also be mindful of this, where those decisions are similarly impactful – healthcare, crime prevention and human resources all spring to mind as examples. The key to ensuring fair treatment of customers and defensible decision making is to continue to ensure that for every model input, the model's marginal response to changes in it aligns with common-sense expectations.

For more information on using Artificial Intelligence in risk, read our latest news and views, or, visit our AI specialisms page.

5. https://risk.jaywing.com/specialisms/artificial-intelligence/archetype/