BY SVETLANA BOROVKOVA

AI and Machine Learning in Insurance: can we ensure Fairness and Explainability?

Artificial Intelligence (AI) and Machine Learning (ML) are rapidly marching into finance. Credit scoring, fraud detection and quant investing are just a few of the finance areas where ML-powered models are already used. Such models are also finding their way into the insurance sector.

Dr. S. Borovkova is a partner and Head of Quantitative Modelling at risk advisory firm Probability & Partners. She is also an Associate Professor of Quantitative Finance and Risk Management at Vrije Universiteit Amsterdam. Find her research on SSRN and her columns on various issues in finance in Financial Investigator.

THE WORRIES OF AFM AND OTHER REGULATORS

Recently, AFM (Autoriteit Financiële Markten) published a paper '*Technologie richting 2023: De toekomst van verzekeren en toezicht*', where it warns of risks associated with new technologies and digitalization, and outlines approaches to mitigating these risks. AFM argues that, by using vast amounts of (often personal) data and advanced models, it becomes possible for insurers to exclude certain customers or dramatically increase their insurance premia. Existing legislation such as acceptance obligation or privacy law GDPR do not offer a full solution to this: acceptance obligation can be avoided by charging an exorbitant premium for an insurance policy, and customers can be nudged to (unwittingly) give permission to the use of their personal data by a click of a mouse.

The two main objections of regulators such as ECB and DNB against use of ML models in material decisions are: their lack of explainability and potential unfairness of outcomes. These worries are also echoed by AFM for the insurance sector.

It is well-known that the outcomes of ML models are difficult to explain, since these models construct highly complex, non-linear relationships between the outcome (e.g., acceptance of a customer) and the inputs (customers' characteristics). This distinguishes them from traditional statistical models: where these relationships are typically linear and, hence, simpler and intuitive. Another well-known issue with machine learning models is that they are prone to unfair outcomes, which can be discriminatory against some groups, such as women or ethnic groups. This happens because machine learning models are very good at finding patterns in data (which carry historical biases, such as men earning on average more than women), carrying these patterns forward, and often amplifying them.

The AFM paper outlines some supervision-based solutions to these risks (such as analysis of individual outcomes, or testing organizations' processes and procedures in their decision making). However, there are also plenty of tools that modelers have at their disposal, to ensure both fairness and explainability. Often, the same tools that make ML models explainable, can be used to assess whether their outcomes are fair. In the remainder of this article, I will discuss some of these tools.

TOOLS FOR EXPLAINABLE ML

The most famous tool for explainable ML are the so-called *SHAP values* (SHapely Additive exPlanations). The SHAP values come from game theory and measure the importance of each input feature for the outcome of the model. SHAP values can do this for the whole dataset (showing what the effect of each feature is on the outcome *on average*), as well as for each individual case: this makes them particularly useful in finance applications. Take as an example a life insurance acceptance model. If the SHAP value for the death benefit amount is the highest among all features, it means that the benefit amount has the biggest effect (on average) on whether the policy is accepted or not. For an individual application, SHAP values allow us to see why that application was rejected: was it because its benefit was far above the average benefit, or because the individual's age was significantly higher than the average applicant's age? So SHAP values allow us to 'demystify' the outcomes of a ML model at the global as well as individual level.

Another powerful technique is called *counterfactual explanations* (*CE*). This technique explains the outcomes of a ML model on an individual (rather than global) level. It tells us, for each negative outcome (e.g., denied insurance policy or a loan), which input features must change in order for this individual to migrate from the negative to the positive class. For example, for a rejected life insurance policy, counterfactual explanations might tell us that, if the applicant reduced his death benefit size by 20%, the application will be approved. Often, there are several different counterfactual explanations possible, but not all of them are actionable (in the above example, instead of reducing the benefit size, the applicant might be advised to lower his age by 10 years, which is clearly impossible). Still, these counterfactual explanations – actionable or not – give us a lot of information about which features the ML model found important for generating a particular outcome.

These are just two of the best-known techniques from the explainable ML toolkit – there are several others, and new ones are being developed.

FAIRNESS: DEFINITIONS AND MEASUREMENT

At the heart of AI fairness is the principle of avoiding preferential treatment of certain groups of society – based on gender, race, age or other protected attributes (these can be also e.g., sexual orientation or religion). Protected attributes are determined by law, but financial institutions can set their own ethical standards (and hence, their own, larger set of protected attributes).

Is a particular ML model fair? The above-mentioned tools (SHAP and CE) can help us determine that. For example, if a model was trained on a full set of input features (so also including the protected attributes), and its SHAP values are high for those protected attributes, it might indicate that the model is unfair. Counterfactual explanations are even better at indicating unfairness: if in a particular case, CE tells an applicant to change her gender and then her loan application will be approved, this clearly indicates unfairness of the model.

An outcome can be either fair or unfair, but a model is not just fair or unfair: there are different degrees of unfairness, or bias. So it is important to measure this bias. There are two notions of fairness: group fairness, which means that the protected group is treated similarly to the advantaged group or the population as a whole, and individual fairness, which means that the negative outcome for a particular individual would not change if his or her protected attribute was different. Both notions are relevant in practice; however, we can only measure the group fairness.

There are three formal definitions of fairness (illustrated in Figure 1) and hence three ways of measuring it. These three bias measures are shown in Figure 2.

Group Fairness

Independence

> Requires that the acceptance rate is equal in all groups.

> The probability of being classified by the algorithm in each of the groups is equal for two individuals with different sensitive characteristics.

$$P(\hat{Y} = y \mid A = a) = P(\hat{Y} = y \mid A = b)$$
$$y \in \{0,1\};\ a,\ b \in A$$

Example: strive for an equal outcome of men and women.

Separation

> Requires that all groups experience equal true pos. rates and false pos. rates.

> The probability of being classified in each of the groups is equal for two individuals with different sensitive attributes given that they belong in the same group

$$P(\hat{Y} = 1 \mid Y = y,\ A = a)$$
$$= P(\hat{Y} = 1 \mid Y = y,\ A = b)$$
$$y \in \{0,1\};\ a,\ b \in A$$

Example: give men and women equal opportunity, regardless the outcome.

Sufficiency

> Requires consistency of pos./neg. predictive values across all groups.

> The probability of being in each of the groups is equal for two individuals with different sensitive characteristics given that they were predicted to belong to the same group.

$$P(Y = y \mid \hat{Y} = 1,\ A = a)$$
$$= P(Y = y \mid \hat{Y} = 1,\ A = b)$$
$$y \in \{0,1\};\ a,\ b \in A$$

Example: both men and women in a range of outcomes predicted should find the same average realised value.

Figure 1: Three notions of group fairness

Group Fairness: Measures

Independence

> Statistical (Demographic) Parity

$$P(\hat{Y} = y \mid A = a) - P(\hat{Y} = y \mid A = b) < 0.2$$

> Disparate Impact

$$\frac{P(\hat{Y} = y \mid A = a)}{P(\hat{Y} = y \mid A = b)} > 0.8$$

Separation

> Equal Opportunity

$$P(\hat{Y} = 0 \mid Y = 1,\ A = a)$$
$$-\ P(\hat{Y} = 0 \mid Y = 1,\ A = b) < 0.2$$

> Equalized Odds

$$P(\hat{Y} = 1 \mid Y = 1,\ A = a)$$
$$-\ P(\hat{Y} = 1 \mid Y = 1,\ A = b) < 0.2$$

Sufficiency

> Predictive Parity

$$P(Y = 1 \mid \hat{Y} = 1,\ A = a)$$
$$-\ P(Y = 1 \mid \hat{Y} = 1,\ A = b) < 0.2$$

> Calibration

$$P(Y = 1 \mid \hat{S} = \hat{s},\ A = a)$$
$$-\ P(Y = 1 \mid \hat{S} = \hat{s},\ A = b) < 0.2$$

where $\hat{s} \in \hat{S}$ is the predicted probability score

Four-fifth rule: prescribes that a selection rate for any disadvantaged group that is less than four-fifths of that for the group with the highest rate.

Figure 2: Measures of bias

MITIGATION OF MODEL UNFAIRNESS

If a ML model is deemed unfair, this does not mean you have to discard it. There are plenty of modern bias mitigation toolkits such as AI Fairness 360 by IBM and other open source tools, which can help model builders reduce or even completely eliminate bias from their models.

There are three points in a model where bias can be reduced. First is the model's input: one can modify the data used to train the ML model, by the so-called 'massaging' (swapping some of the outcomes between the advantaged and disadvantaged groups), re-weighting or changing features to increase fairness. Second is the ML algorithm itself. Changing the algorithm to be more fair leads to the best outcomes, but is difficult and costly, since most models are built using ready-made algorithms and packages, which are not easy to change. The third option is to change the model outcomes to increase fairness – this is bias mitigation in the post-processing stage.

Any bias mitigation results in some loss of the model performance. So this is a balancing act between improving fairness while still having an adequate model. The good news is that such balance is easily achieved: the modern bias mitigation techniques do not require much of the performance loss, while significantly improving fairness of a ML model.

TO CONCLUDE

The AFM and other regulators express fair concerns about the use of AI and ML, powered by large quantities of data, in finance and also in the insurance sector, citing lack of explainability and potential unfairness of outcomes. However, there are plenty of modern tools and techniques for ensuring explainable ML, bias measurement and mitigation – the only issue is awareness of them and how to apply them appropriately. ■