Bias in Insurance Fraud Risk Prediction

Chiara Manganini and Tommaso Terragni
Kube Partners Italy Srl, Monza (MB), Italy
{chiaramanganini, tommasoterragni}@kubepartners.com

What is Detector and how it works

Detector is a system that predicts the fraud risk of insurance claims. Insurance companies use Detector as a decision-support **tool to optimise** the resources (time, money, and personnel) they allocate for **uncovering and contrasting claim fraud**. By giving each claim a score, Detector provides the human expert with an **examination order** which helps prioritise manual inspection of suspect refunding requests. For each of the claims Detector selected as suspect, the human expert carries out an evaluation which either confirms or disproves Detector's prediction and is **taken to be the ground truth** for the fraud label of that claim.

KEY CONCEPTS

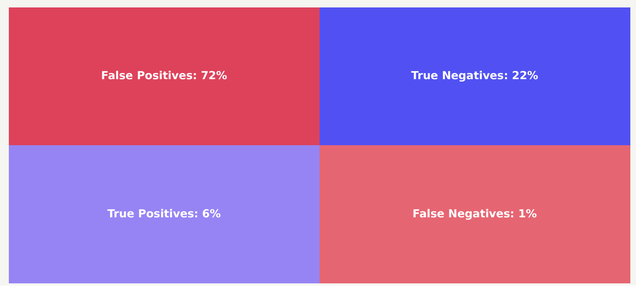
- **RARITY** Insurance fraud is a **rare and nuanced phenomenon**.
- **PRIORITY** Not every claim can be looked into by anti-fraud experts. It is necessary to **sort** between suspect and not suspect claims and **prioritise** the examination of the suspect ones based on a fraud risk order.
- **SCORE** Detector associates a **score** with every claim: the higher the score, the higher the priority for it to be examined by a human anti-fraud expert, who ultimately decides whether it is fraudulent or not.
- **PLURALITY** No single indicator alone has the required precision. Detector's total score combines 4 different indexes, respectively based on **heuristic, geographical, socio-relational, and ML analysis** of the claim.

Goal

In this work, we address three crucial questions about the specific case of bias occurrence in Detector:

1. What constitutes an **instance of bias** in Detector?
2. What constitutes an adequate **measure** for it?
3. What **characterises** this particular instance?

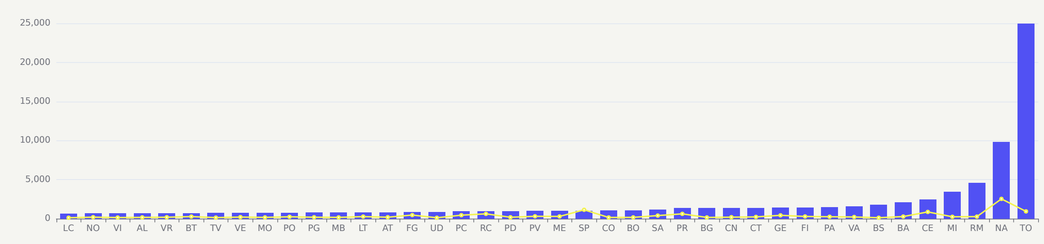
Confusion matrix



Sample and Method

STEP 1 We prepared a synthetic sample of **100,000 closed car insurance claims**. The sample was generated to reflect realistic estimates on the Italian incidence of fraud in car claims (about 6%). For each data point, the following attributes were considered: the **policyholder's province of residence, the claim fraud risk score given by Detector, the more specific geographical risk score, and the outcome of the human examiner's investigation on it.**

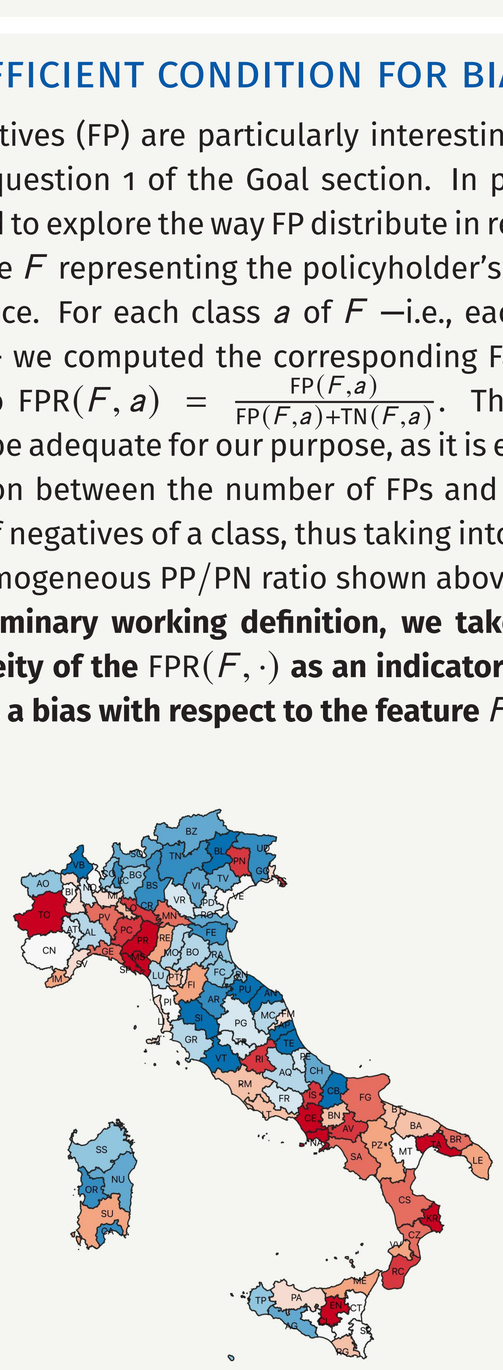
STEP 2 We then computed the confusion matrix of Detector's predictions for this sample. The label given by the human examiner is the **ground truth** (fraud/not fraud), while the total score computed by Detector constitutes the **predicted value**. Although continuous, the score can be considered as a binary prediction by fixing a threshold T: if a claim receives a score greater or equal to T, then it is considered fraudulent, genuine otherwise. The figure summarises some key features of the sample in relation to the policyholder's province of residence: the frequency of classes and their Predicted Positives/Predicted Negatives (PP/PN) ratio.



1. A sufficient condition for bias

False positives (FP) are particularly interesting for addressing question 1 of the Goal section. In particular, we wanted to explore the way FP distribute in relation to the feature F representing the policyholder's province of residence. For each class a of F —i.e., each Italian province— we computed the corresponding False Positive Ratio $\mathrm{FPR}(F, a) = \frac{\mathrm{FP}(F,a)}{\mathrm{FP}(F,a)+\mathrm{TN}(F,a)}$. This metric seems to be adequate for our purpose, as it is expresses the relation between the number of FPs and the total number of negatives of a class, thus taking into account the in-homogeneous PP/PN ratio shown above.

As a preliminary working definition, we take the in-homogeneity of the $\mathrm{FPR}(F, \cdot)$ **as an indicator that Detector has a bias with respect to the feature** F.



The map of the FPR.

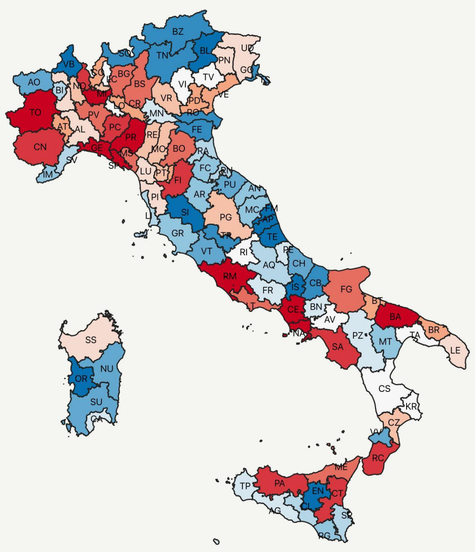
2. A measure for bias

A second step consists in devising a metric to appropriately quantify Detector's bias w.r.t. to F in terms of ineffective resource allocation in the anti-fraud process. Starting from the FPR, this metric must factor in a weight that expresses the relative frequency with which class a is actually encountered by the human inspector. We therefore define this weight as the percentage of predicted positives (PP) belonging to that class:

$$\mathrm{wFPR}(F, a) = \mathrm{FPR}(F, a) \cdot \frac{\mathrm{PP}(F, a)}{\mathrm{PP}}.$$

The figure shows the distribution of the $\mathrm{wFPR}(F, \cdot)$ across the Italian provinces: wFPR associates higher values to the most frequent provinces of the sample. This emphasises the fact that the presence of clusters of FP in those provinces have an amplified negative impact on performance. By aggregating over the classes of F, we define the measure of Detector's bias w.r.t. the feature F by

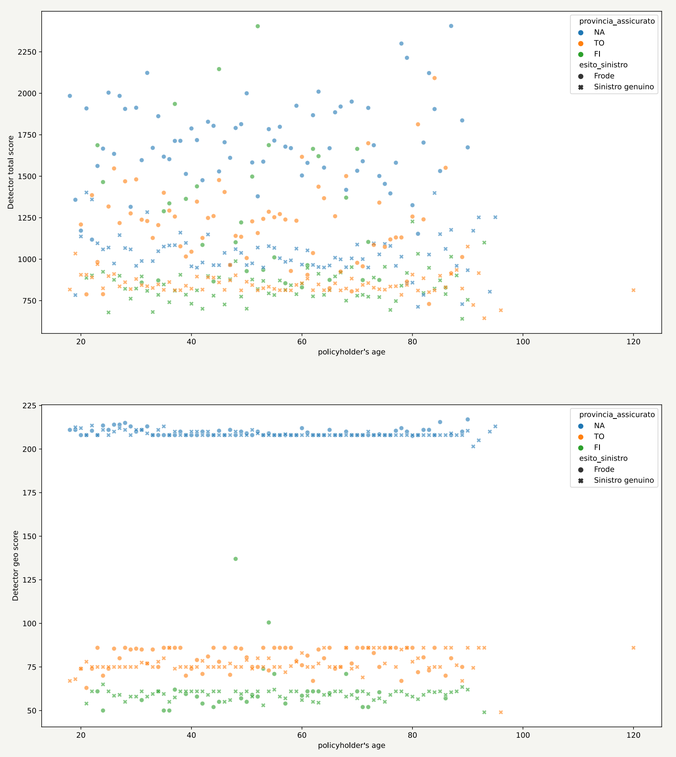
$$\mathrm{wFPR}(F) = \sum_{a \in F} \mathrm{wFPR}(F, a).$$



The map of the $\mathrm{wFPR}(F, \cdot)$.

3. Qualifying geographical bias

Finally, we argue that the bias analysed in the present work w.r.t. to the feature F, the policyholder's province of residence, can be qualified as a geographical bias of the system. An analysis of the distribution of data points along Detector's total (upper image) and geographical (lower image) median scores reveals a stratification based on the geographical attribute of interest.



Takeaways

1. Given a feature F, bias in relation to F occurs in Detector whenever the distribution of false positive rates (FPR) is uneven among F-classes.
2. An adequate metric to measure bias in relation to a specific class a of F is the weighted version wFPR of the false positive rate.
3. A bias with respect to a feature F is called geographical if F is correlated to a geographical variable.