|  |  |
| --- | --- |
| Desiderata | How to test |
| Model-agnostic | - (by design) |
| Valid | Check that new prediction is of the target class |
| Robust | - (by design: deterministic) |
| Sparse | Count how many features are shown |
| Similar | Calculate the distance to the data point |
| Realistic | Calculate the distance to training data point |
| Informative | Show that the proposed feature changes are in line with the SHAP values |

Smart way to generate sparser explanations;

* Only show feature changes that are in line with the SHAP values.

Qualitative evaluation

## Distance kernel idea

How? Add the feature values of each training point together, weighted by the training point’s distance to the profile we are trying to explain. This creates a new data point which is not guaranteed to be realistic. If we then search for the most similar training data point to this average, we require a second run through all the training data points which is slow.

This was brought up because the initial implementation often resulted in counter-intuitive explanation; e.g. adjusting a feature value in the opposite direction than expected. For instance, decreasing the TAC value when increasing the NOC. We tried to come up with ideas to counteract this problem. E.g. by creating clusters and presenting the median as a datapoint. However, clustering comes with extra parameters that are difficult to accurately optimize. A simpler approach would be to not change features that are against the median of the closest *n* features, which would again require the optimization of *n*, and no guarantees that the prediction is still valid.

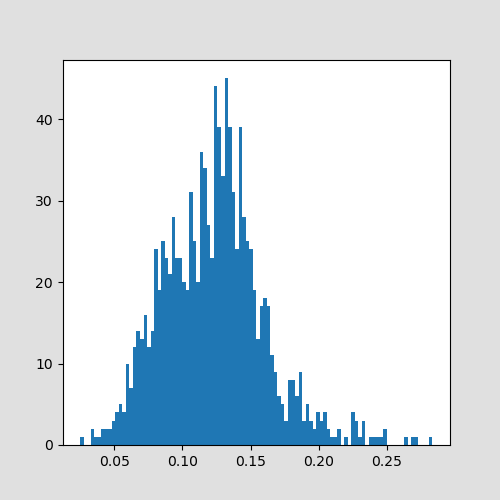
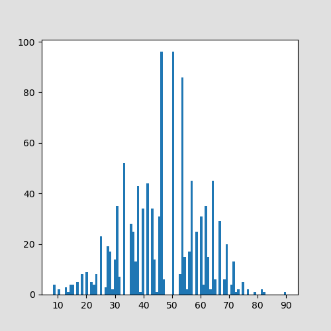
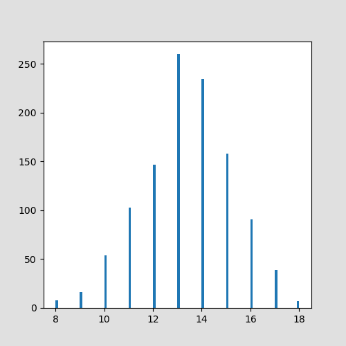
We noticed this issue mainly before we sampled the extra 5000 data points, showing how it could have mainly been caused by the sparsity of the data. It has improved since sampling.

These also stray from the concept of a counterfactual explanation i.e. finding *the most similar world* to the original profile, yet with the target prediction. By simply presenting an average data point, this loses all the quirks of what makes a profile unique. Presenting a profile that has similar quirks as the profile to be explained, even though those do not fit the average, is preferred.

For the following two approaches, we use a heuristically chosen value for the weight of the two scores of distance and number of features changed. (10 \* distance + 1 \* num\_features\_changed).

Using the weighted kernel approach, (1 – distance scores) we obtain the following measures on 1118 test samples.

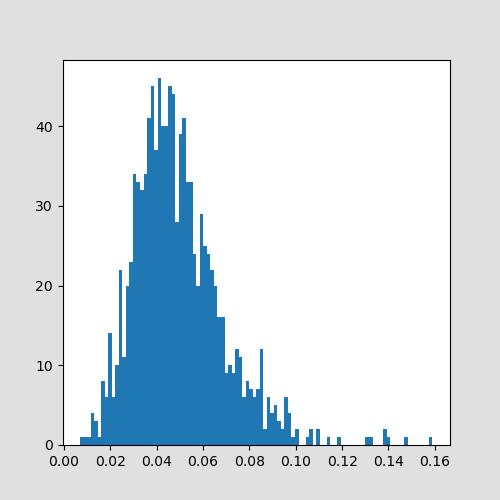
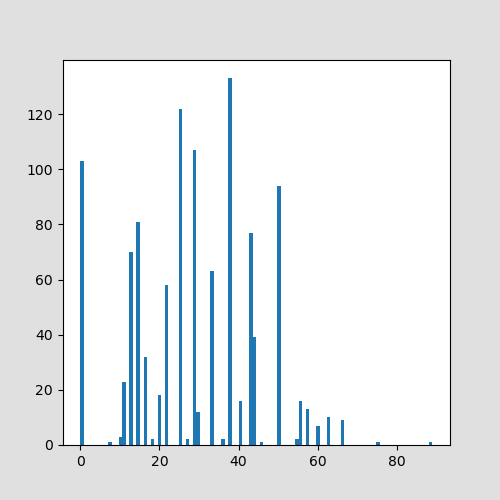
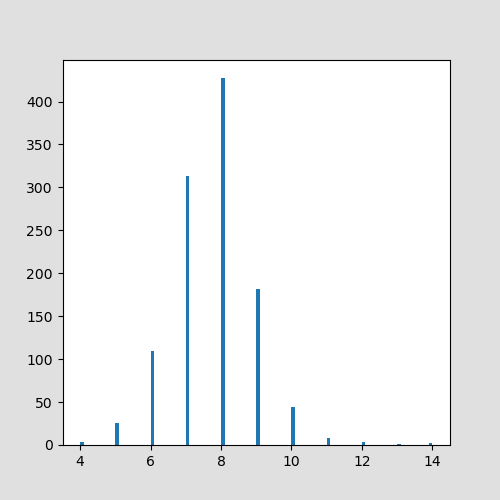
|  |  |
| --- | --- |
| Metric | Score |
| Average number of features changed | 13 (/19) |
| Average number of features changed with counterintuitive SHAP changes | 9 (/19) |
| Average distance to the profile | 0.123 (/1) |
| Average distance to the closest training point | 0 (boolean) |



All in all, the CF profiles picked here have a high number of changed features, a high distance to the profile we want to explain, and a large number of feature changes that do not move the prediction in the right direction.

In contrast, when we simply pick the closest profile according to this weighted sum, we obtain the following scores:

|  |  |
| --- | --- |
| Metric | Score |
| Average number of features changed | 8 (/19) |
| Average number of features changed with counterintuitive SHAP changes | 5 (/19) |
| Average distance to the profile | 0.049 (/1) |
| Average distance to the closest training point | 0 (boolean) |



This weight was chosen according to the observed values of the training dataset to approximately center the means.

Perhaps there could be a filter that shows the full example profile versus the feature changes that have helped towards the target prediction. In this way, we 1. Show a 100% real training data point, but also 2. Present the most influential changes.