Making an IMPACT

Date: 18 November 2022
Authors: Anna Theorin Johansson, Jacob Rothschild, Andrew Zehr
In collaboration with: *IMPACT Initiatives*
Developed in the context of Hack4Good, 4th Edition

Introduction

The mission of IMPACT Initiatives is simple: survey households in crisis regions, collecting data to inform and improve the local distribution of humanitarian aid. Survey collectors, called enumerators, are sent out in the field to conduct household surveys, storing the answers in an app, and the responses are used, in the respective countries, as the foundation for humanitarian planning and decision-making. In inaccessible regions, the data collection is often done by third parties – for example, local NGOs with better access to the area – which ultimately requires plenty of human intervention to guarantee the accuracy of the data.

The organization's current approach to data cleaning is very conservative: should you find a suspicious looking survey response, you have no choice but to delete all entries by that enumerator. The old approach is a waste of both time and important information, and an improvement is therefore long overdue. To rectify the situation, we have developed an end-to-end interpretable anomaly detection system that can be used across different projects.

Should our proposed tool be implemented by IMPACT, it would substantially improve both the quality and the efficiency of their data cleaning, diminishing the factor of human error, and making the data more accurate, more reliable and more affordable. In the big picture, the solution improves the very core function of IMPACT Initiatives, thereby improving the distribution of humanitarian aid in areas of crisis.

Approach

What IMPACT ideally wants is an interpretable fraud detection system that they can reuse across all of their projects. However, their approach so far has been to base their anomaly detection on the actual responses to the survey. A central problem with this approach is that the questions asked and answers expected differ between questionnaires and questionnaires differ between projects, so the approach doesn't generalize. This means that IMPACT has until now been forced to tailor-make different sets of analyses for their different projects.

What has now changed is that IMPACT for the first time has access to data on how the enumerators were interacting with the app when filling out their survey responses. More specifically, for each survey response, we have access to what time-intervals each page of the questionnaire was visited. While questions and answers very easily become too specific to a certain questionnaire, the time information can reveal patterns that are suspicious no matter what the content of the questionnaire is. For example, if a questionnaire has a question on which all survey responses spent around 2 minutes except one that spent 1 second, that last

response is suspicious no matter what the actual question was, thus no matter which questionnaire the question belonged to. Thus, a feature capturing such a pattern can be used for any questionnaire to detect survey responses that are suspicious upon comparison with the other responses for that questionnaire.

The approach thus boiled down to engineering good features to assign the survey responses and fitting an interpretable anomaly detector on these. The good features would then be characterized by having an unusual value if the corresponding survey response is showing some kind of suspicious pattern upon comparison with the other responses of its questionnaire, no matter the specific questionnaire. For finding such features, we performed a three-step brainstorming process, answering the following questions:

1. What are the ways in which an enumerator can try to falsify a survey response?
2. What patterns will that generate in the time information?
3. What features can capture these patterns?

The corresponding features and brainstorming table can be found in the GitLab.

Once the features were calculated, an anomaly detection algorithm was used to identify the survey responses that were unusual when compared to other responses. In our implementation, we used the Isolation Forest algorithm for anomaly detection, which operates similar to a Random Forest of decision trees with a few differences. At each step, it selects a feature and a random value at which to "cut" or separate that data. It continues making random cuts until the data points are isolated from one another. The average number of cuts it takes to isolate a given data point is then a sign of how anomalous or typical that data point is. Anomalous points require, on average, fewer cuts to be isolated than normal data points. An illustration of the algorithm can be seen in Figure 1. One benefit of the Isolation Forest is that it works well on high-dimensional data, which was important as many features were calculated.

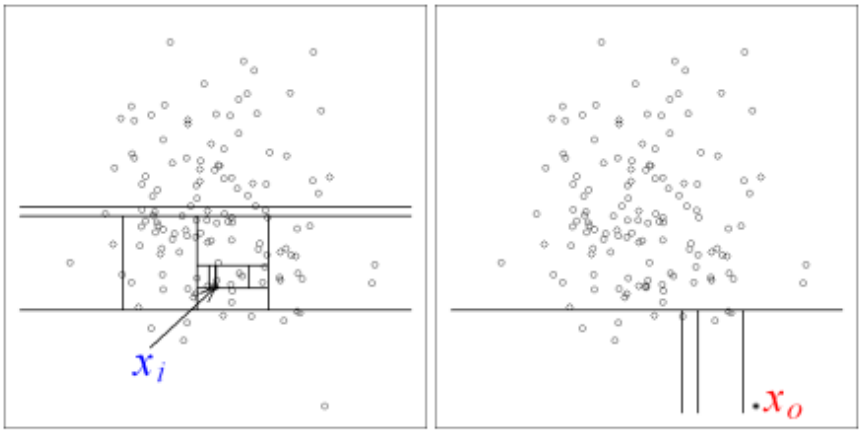

Figure 1: Illustration of isolation forest algorithm. It takes more cuts, on average, to isolate point x_i from the rest of the data points than it does to isolate point x_0 [1].

Isolation Forests are black-box algorithms, meaning that although the input and output to the model are clear, it is difficult or even impossible to understand the inner-workings of the

algorithm and the reasons for its decisions. This poses a practical problem to implementation. In order for IMPACT Initiatives to evaluate the reliability of the algorithm or take action on potentially suspicious survey responses, it must be clear why the response is being classified as suspicious. To solve this problem, we employed Shapley values, a game-theoretic method of assigning importance to all of the features in a machine learning model. With Shapley values we were able to report which features contributed most heavily to a given survey response being classified as an anomaly. A visualization of the average importance by feature (across all responses) can be seen in Figure 2.

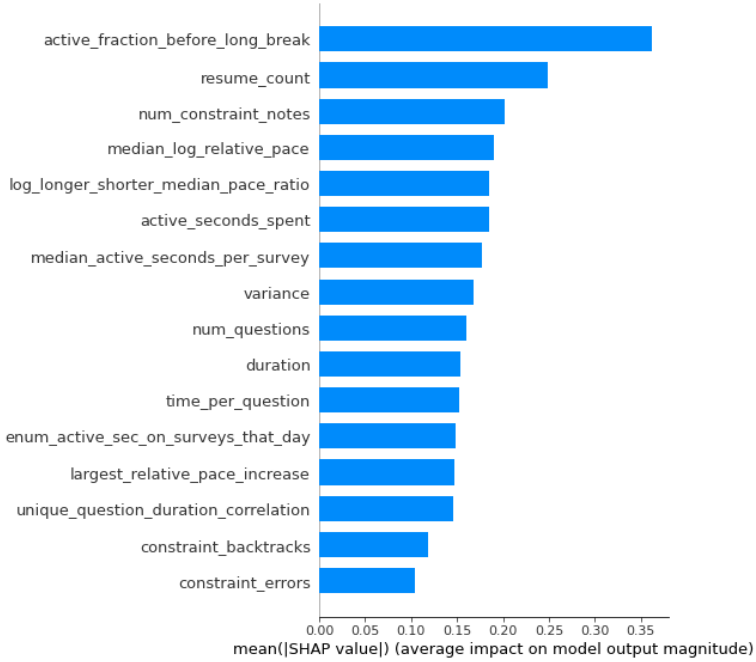

Figure 2: Average Shapley values across all survey responses

Difficulties, Limitations & Risks

One of the chief difficulties of this project arose from the fact that it is impossible to be sure if a given survey response is fraudulent or not. Thus, it was impossible to use supervised learning techniques to classify fraudulent survey responses. Likewise, it was difficult to objectively assess the performance of the methodology.

A further difficulty came from the fact that the questionnaire used differs from region to region and from year to year. On top of this, substantial domain knowledge would be needed to fully understand the questions and expected answers. Any effective solution would need to be generalizable between different questionnaire formats, regardless of the questionnaire content. This problem was addressed by focusing all features on the audit files which are consistent across all data collection efforts.

Results & Deliverables

The project resulted in a well-documented, modular code repository containing the entire data science pipeline from start to finish. From a raw data file, the pipeline automatically calculates all features and outputs the anomalous survey responses to a file on the user's

computer. Along with the anomalous survey responses, an interpretation for why they are anomalous is also outputted. The type and number of features to be calculated as well as the operation of the anomaly detection algorithm can be easily edited in the future as needs arise or goals change. The code plots comparisons of the feature distributions for typical survey responses versus those of suspicious ones. Particularly suspicious enumerators are also highlighted. Examples of these plots can be seen in Figure 3.

Roughly 4% of the total survey responses we investigated were classified as anomalies, which aligns with the prior expectation that IMPACT Initiatives had on the number of suspicious survey responses.

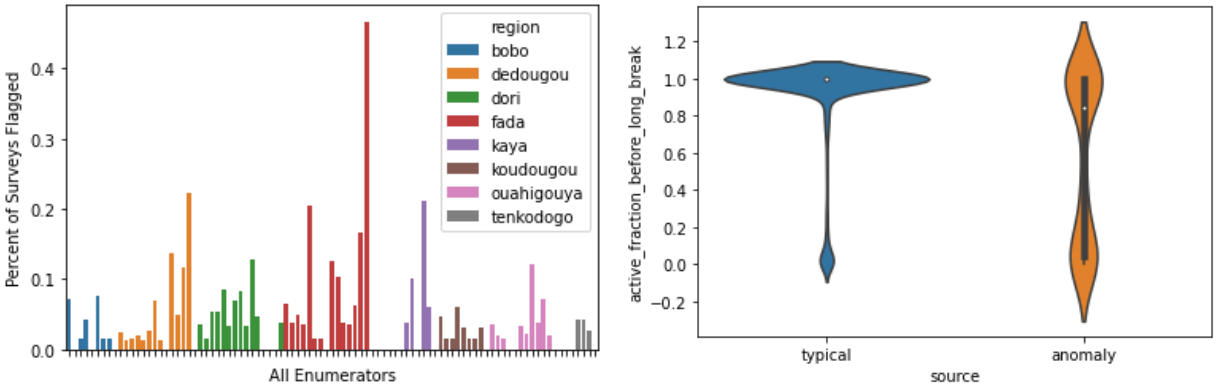

Figure 3: (Left) Percentage of suspicious survey responses by enumerator grouped by region. (Right) Sample comparison of feature distributions from typical versus anomalous responses.

Recommendations & Conclusion

After a short tutorial, the code itself should be relatively simple to use, and will not require any substantial amount of extra work to implement. We would urge IMPACT to make the solution part of their official data cleaning guidelines, and use the extensive feature documentation provided to regularly review the process and update the features used.

One idea for improvement is to extend the current algorithm to put more weight on deviations from normal in certain directions – as it is right now, the algorithm looks for outliers on both extremes for each feature, but naturally some directions are more suspicious than others. Our next suggestion is to use the speed and ease of use of this approach to look into online real-time anomaly detection, to have suspicious responses flagged immediately upon registration in the app, so that they could be dealt with quickly. For example, the supervisors in the field could confront the enumerator at the end of the day, thereby getting a direct explanation to the reasoning behind the suspicious answer, reducing the need to later have to go through the entire survey.

References:

[1] Liu, Fei Tony & Ting, Kai & Zhou, Zhi-Hua. (2009). Isolation Forest. 413 - 422. 10.1109/ICDM.2008.17.