Introduction

C:*The manuscript presents an assessment of two machine learning methods for weather index insurance. My overall impression is that this is a methodologically sound albeit not overly innovative study. It avoids many pitfalls that are sometimes overlooked even in peer-reviewed publications. The train - validation - test split is performed accurately, the problem of class imbalance is tackled adequately (using suitable performance metrics), and presentation quality in terms of figures is good. My recommendation is that this paper can be accepted subject to minor revisions.*

R: Dear Reviewer,

Thank you very much for your time and effort reviewing our manuscript. This response (R) carefully addresses all the comments (C). Where deemed appropriate, modifications to the manuscript are proposed (red underlined text indicates additions to the manuscript, blue strikethrough text indicates removed text).

General Remarks

C: *That the authors have put some effort into making the study is easily understandable for people not intimately familiar with machine learning methods. While this is commendable in a journal with a core focus on natural hazards, I feel that the manuscript is a bit lengthy at times. Some parts of sections 1 (Introduction) and particularly 2 (Methodology) could be shortened in order to make them more concise. Some parts read like an introductory book on machine learning. I suggest to go over section 2 again and streamline some of the rather basic parts.*

R: We agree that parts of Section 2 may be streamlined. We propose the following changes, which we believe will improve the readability of the manuscript while still providing some of the key concepts of the methods that are used.

At line 215:

"~~Neural networks drew inspiration from the behaviour of biological neurons in the human brain, where neurons interconnected by synapses are able to perform a function when activated (Vaiserman and Lushchak, 2018).~~ Neural networks are a machine learning algorithm composed by nodes (or neurons) that are typically organized into three types of layers: input, hidden and output. Once built, a neural network is used to understand and translate the underlying relationship between a set of input data (represented by the input layer) and the corresponding target (represented by the output layer). In recent years and with the advent of big data, neural networks have been increasingly used to efficiently solve many real-world problems, related for example with pattern recognition and classification of satellite images (Dreyfus, 2005), where the capacity of this algorithm to handle nonlinearity can be put to fruition (Stevens and Antiga, 2019). ~~The number of neurons in the input layer is uniquely determined by the input data (e.g. number of environmental variables source), while the output layer, for binary~~

~~classification problem, contains one hidden node returning a prediction about either of the two classes.~~ A key problem when applying neural networks is defining the number of hidden layers and hidden nodes. This must usually be done specifically for each application case, as there is no globally agreed-on procedure to derive the ideal configuration of the network architecture (Mas and Flores, 2008). ~~Depending on the number of layers, the neural network takes different names: artificial neural networks (ANN) are usually defined as networks with only one hidden layer; deep neural networks (DNN) are composed of two or more layers. The difference between ANN and DNN is not perfectly defined in literature, and for sake of simplicity.~~ Although different terminology may be used to refer to neural networks depending on their architectures (e.g., Artificial Neural Network, Deep Neural Network), in this paper they are addressed simply as neural networks, specifying where needed...”

At line 297:

“Moreover, data preprocessing ~~preparation~~ usually generates leaner and more reliable datasets, boosting the efficiency of the ML algorithm (Zhang et al.,2003). The literature presents several operations that can be adopted to transform the data depending on the type of task the model is required to carry out (Huang et al., 2015; Felix and Lee, 2019). ~~For instance, images and video analysis might require previous cropping or blurring through Gaussian convolution to better identify the edges of an image (Getreuer, 2013), while machine learning model used for time-series forecast benefit more from the detection of outliers and duplicate instances (Kotsiantis and Kanellopoulos, 2006)~~.”

At line 307:

“~~Also, this process was used to identify any incoherence amid the dataset for example by checking the spatial patterns of precipitation in the days leading to flood events.~~”

At line 336:

“~~Both scikit-learn and TensorFlow allow for the implementation of class weight into the model construction through an explicit parameter. The weighting values can easily be tweaked to find the optimal settings for a given problem~~”

At line 390:

“~~Also, while here we focus on performance-based evaluation measures, an alternative approach may be to quantify the utility of the predictive systems. By taking into account actual user expenses and thus specific weights for different model outcomes, a utility-based approach may potentially lead to different decisions regarding model selection and definition of the trigger threshold (Murphy and Ehrendorfer, 1987; Figueiredo et al., 2018). This aspect is outside the scope of the present article and warrants further research.~~

Table 5 summarizes the metrics described above used in this paper to evaluate model performances.

In the context of performance evaluation, it is also relevant to discuss ~~the issue of class imbalance~~. ~~Class imbalance refers to the difference between positive and negative instances~~

~~with the latter usually outnumbering the former. Thus, it is important to keep in mind~~ how class imbalance might affect measures that use true negative in their computation."

C:*Please also double check the language throughout the text, especially syntax (e.g. line 289: Hereinafter is proposed a procedure (...)).*

R: In the revised manuscript we will carefully check and improve the writing and syntax, as suggested.

C: *I would refrain from using the term 'big data' in this context and adjust the title accordingly. Simply because the authors use larger data sets, this is not a novel big data problem per se.*

R: We agree and propose to change the title to:

"The potential of machine learning for weather index insurance"

Specific Remarks

C: *Line 15: I recommend to avoid the term 'significant' in a methodology- oriented paper. This might lead to confusion with respect to statistical significance.*

R: We agree and will replace the word "significant" with "substantial" here.

C: *Line 65 ff: This section states the core aim of the paper. Please add information on the input data source that is used. Currently, this essential statement is missing. In addition, I suggest to more precisely refer to flood and drought in this statement: '(...) is capable of objectively identifying and classifying extreme flood and drought events from satellite and gauge data products in near-real time (...)'.*

R: We suggest the following change:

"we propose and apply a machine learning methodology that is capable of objectively identifying and classifying extreme weather events, namely flood and drought, in near-real time, using quasi-global gridded climate datasets derived from satellite imagery or a combination of observations and satellite imagery. This methodology is then used to address the following research questions"

C: *Line 129ff: These 5 criteria are important. I would welcome a reference of the actual values for these five criteria in the text. Maybe add a table featuring spatial and temporal metadata of the datasets used?*

R: We agree. We would like to point out that Table 1 and Table 2 already report the information regarding the 5 criteria listed. An effort to highlight the tables will be made in the article adding a reference. We suggest the following change to the manuscript to make the connection more direct.

"... Based on a comprehensive review of available datasets, we found six rainfall datasets and one soil moisture dataset, comprising 4 layers, matching the above criteria. The main features of the selected datasets are reported in Table 1 and Table 2..."

C: *Line 325: Missing year in Mueller and Massaron*

R: The reference will be corrected to: "(Mueller and Massaron, 2016)"

C:*It is not ultimately clear which method for tackling class imbalance was used. I realized this when reading the results section, but the authors might want to add a sentence that this was also tuned as a model parameter in the methods section.*

R: We agree with the reviewer that this should be made more clear. In the manuscript, we tried to specify this aspect at line 342 using the term "data augmentation technique", which we reckon might create some confusion. We propose the following change at line 342:

"... all the ~~data augmentation techniques~~ resampling techniques previously introduced were tested…"

C:*Since different methods for approaching class imbalance were used: Is there a reasons why procedures for undersampling were not considered? Or combinations such as SMOTE + undersampling?*

R: When evaluating which techniques were more suitable to tackle class imbalance, we concluded that using undersampling would have reduced our datasets to such a dimension that was not deemed appropriate for the training of the ML algorithm. Accordingly, for the same reason, a combination of SMOTE and undersampling was not used. Since SMOTE generates synthetic samples "close" to the sample it is trying to replicate, undersampling from a group built as such could lead to the loss of some real events.

We propose to add the following sentence at line 335 to clarify:

"(…) Oversampling, SMOTE and class weight were the resampling techniques deemed more appropriate to the scope of this work, namely, identifying events in the minority class. (…)"

C:*Line 366f: Please check reference (M. and M.N., 2015)*

R: Will be corrected in "(Hossin and Sulainman, 2015)"

C:*Similar to the class imbalance method, it is unclear in the methods section which performance metrics have been used to compare the performance of the model. Was one specific metric used, or was the decision reached using all metrics presented in Tab. 5 by comparing all of them somehow? This is mentioned in the results section, but it is not clear when reading the methods section. I would argue that this is a methodological decision, not a result of the analysis.*

R: Thank you for pointing out this missing information. We agree that this is a methodological decision and we'd like to propose the following changes at line 402:

"Lastly, once the domain of all configurations is well established and the best settings of the ML algorithms were selected based on the highest values of F1 score and area under the PS curve ~~are extracted from it through the aforementioned metrics~~, the predictive performances of the models are compared to those of logistic regression (LR) models. (…)"

C:*The reference model on logistic regression is not ultimately clear. Did the authors use simple logisitic regression? Which link function was used? Did the authors include interaction effects? Did the authors use nonlinear effects? Simple logistic regression is fine as a reference model, but I think this could be stated more clearly.*

R: We suggest the following addition to the manuscript to clarify the doubts raised regarding the properties of the logistic regression

" The logistic regression adopted as a baseline takes as input multiple environmental variables, in line with the procedure followed for the ML methods and used a logit function (eq.6) as link function, neglecting interaction and nonlinear effects amid predictors. The logistic regression is a more traditional statistical model whose application to index insurance has recently been proposed, and can be said to already represent in itself an improvement over common practice in the field (Calvet et al., 2017; Figueiredo et al., 2018)."

C:*Is there any particular justification why these two methods were selected specifically? My guess would be that a simple random forest with default parameters would probably perform equally well.*

R: We limited our analyses to these two types of ML methods for the sake of brevity. We reckon that other ML methods might enable comparable results. We propose adding a comment about this in the Conclusion:

“It is also worth noting that although this work focuses on the application of neural network and support vector machine models, we expect that comparable results could be obtained using other machine learning algorithms, which calls for further research.”

C:*I think more focus on the discussion would be beneficial. Results are described in this section, and findings are briefly commented. However, I am under the impression that there is some imbalance between the first half of the manuscript, which is quite extensive, and the discussion of the results, which is quite sparse. What have we learned from this study? Which novel aspects does this analysis show? What do the results mean for the Dominican Republic? Which impacts do the findings have on the study area?*

R: In agreement with what was discussed in the general remarks section about the length of the methodology section, an effort was made to slim down the introductory part and to improve the results and discussion section, to strike a better balance between the two parts. In an effort to provide also an answer to the reviewer’s questions, we propose the following additions in the conclusion section along with the cuts already mentioned for Section 2:

At line 599:

“It is also worth noting that although this work focuses on the application of neural network and support vector machine models, we expect that comparable results could be obtained using other machine learning algorithms, which calls for further research.”

We found appropriate moving the following paragraph from the methodology section to conclusion:

“Also, while here we focus on performance-based evaluation measures, an alternative approach may be to quantify the utility of the predictive systems. By taking into account actual user expenses and thus specific weights for different model outcomes, a utility-based approach may potentially lead to different decisions regarding model selection and definition of the trigger threshold (Murphy and Ehrendorfer, 1987; Figueiredo et al., 2018). This aspect is outside the scope of the present article and warrants further research.”

At line 601:

“Although several issues raised in this article warrant further research, there is clear potential in the application of machine algorithms to take advantage of increasing amounts of available environmental data within the context of weather index insurance. The capability of these algorithms to reduce basis risk with respect to traditional methods could play a key role in the adoption of parametric insurance in the Dominican context and more generally for those countries that detain a low level of information about risk. Indeed, being able to rely on global data that are disentangled from the resources of a given territory, both from the point of view of climate data (e.g., lack of rain-gauge network) and from the point of view of information about past natural disasters, is an appealing feature of the work presented that would make the approach proposed feasible for other countries. The framework presented and topics discussed in this study provide a scientific basis for the development of robust and operationalizable parametric risk transfer products.”