Natural Hazards and Earth System Sciences Discussions

Nat. Hazards Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/nhess-2022-227-RC2, 2022
© Author(s) 2022. This work is distributed under
the Creative Commons Attribution 4.0 License.

Comment on nhess-2022-227

Anonymous Referee #2

Referee comment on "Development of a seismic loss prediction model for residential buildings using machine learning – Ōtautahi□/□Christchurch, New Zealand" by Samuel Roeslin et al., Nat. Hazards Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/nhess-2022-227-RC2, 2022

The authors have presented a novel ML based approach to estimating the loss to buildings after an earthquake is 3 categories - low, medium, and high. As part of model training, the authors have performed spatial data merging between 3 datasets, and only selected the subset of buildings with the least chance of erroneous data attributes. The authors have also focused on the 4 earthquakes in NZ around 2011 with the maximum number of data points. The paper is well structured, and is fairly easy to follow.

However, I found some of the key information about the ML approach missing or confusing in the paper, and have highlighted it below. I believe that the paper would be further improved substantially by adding or clarifying the ML approach. I have listed both my major and minor concerns below.

- The selection of the test set is unclear in the paper. It also appears that the test set has been erroneously used as a validation set. If that is the case, then it is difficult to assess the generalizability of the authors' conclusions. It would be helpful to clarify how the test set was selected and used in this study. Additional comments regarding test set are also included below with specific line references.
- While it is a suitable approach to only select the 4 events with the highest number of claims during model training, the other events with fewer claims could be used for testing purposes. This would not only ensure that no data leakage occurred between the training and test sets, but also enable the authors to validate the generalizability of their models more effectively.
- It would improve the paper if the authors added their thoughts on some of the potential use cases of this research. While the authors' conclusions indicate a promise for using ML within this domain, it was unclear how this model and approach could be used in the future. For example, if training data is needed each time an earthquake occurs, then is one of the use cases to manually collect a subset of ground truth data for building losses, train a model, and then apply it widely to the rest of the buildings?
- Further discussion of the model metrics such as recall and precision would be helpful.

For example, a recall of only 20% for overcap, and 49% for low loss category indicates that 80% and 51% of these losses, respectively would be missed when implementing this model. Depending on the model's use cases, this could have a significant impact on the model's utility. Further discussion of the most appropriate metric (or their combinations), given the model's use cases would also improve the paper. For example, why was accuracy selected as the primary evaluation metric for choosing the best performing model?

- I appreciated that the authors listed the distribution imbalance of different features, such as construction type. However, the paper could be further improved by adding the model performance in those different feature categories. This would enable the reader to understand in which categories the model performs better than others.
- Given the relatively low performance of the ML model (as highlighted above for recall), adding a section on error analysis would substantially improve the paper. In error analysis for ML, the objective is to identify the cases in which the model does not perform well. This error analysis is often used in ML modeling to improve model performance and generalizability.
- Figure 13 is missing, and appears to be a repeat of Figure 12. Hence, Section 9 - Insights - could not be reviewed.
- It would further improve the paper if the authors added some information about their hyperparameter tuning methodology, and which search strategy they used.
- Line 50 - While the authors are completely correct in the paragraph at line 50, this paper deals with ML for structured data, for which the goal is often to surpass human performance since humans are generally unable to identify all patterns in millions of data points with hundreds of features, often found in these problems. Hence the paragraph does not apply to the ML scope of this paper. It may be suitable to remove the paragraph within the scope of this paper, or change "human-level performance" to "baseline model", which would be a more suitable term in this case.
- Line 65 - latter -> later
- Line 73 - Suggest adding reference/url for the source of the data.
- Line 83 - It would be helpful to further describe Figure 2. Why is there a difference in the number of claims and buildings?
- Line 95 - I was curious about the accuracy of the Riskscape dataset. For example, are the building characteristics determined statistically from Census data similar to HAZUS in the US, or was it based on collecting data from building records so that it is expected to be fairly accurate? If possible, it would be helpful in the paper to include some information describing Riskscape's data collection methodology and comment on its expected accuracy.
- Line 115 - Although a reference is provided to the authors' previous work, it would be helpful to summarize the major reasons for incorrect merging using direct spatial joins within this paper to help understand the issue without having to read the previous work.
- Line 122 - It would be helpful if the authors added the percentage of addresses in each of the 3 categories - 1-1 match with titles, 0-1 match, and many-1 match.
- Line 131 - It would be helpful if the authors added the percentage of RiskScape data that was discarded.
- Line 132 - I was unable to understand the intent described in this paragraph, especially the first and the last sentences.
- Table 1 - The table is very helpful. However, the action taken for 2 points LINZ and 1 point Riskscape was unclear. The above mentioned percentages of data could also be added to Table 1 instead.
- Line 150 - It would be helpful if the authors added the methodology used to merge soil conditions, and liquefaction occurrence with street address. Did they use the same inverse distance weighted interpolation as seismic demand?
- Line 172 - The reason for discarding claims with maximum value lower than or higher than $115,000 is unclear. Is it because this wasn't possible and hence the data is erroneous?

- Line 240 - It is unclear which event was selected as the test set. From my understanding of line 243, one of the 4 events was selected as test set, and the other 3 events as training+validation sets. However, it also appears from the sentence that in different instances of the model, a different event was selected as a test set so as to determine the most generalizable model. If that is the case, the test set was erroneously used as a validation set, since the model cannot be changed at any point after evaluating its performance on the test set. It would be helpful to clarify the selection of the test set, and ensure that it was only used once at the end to evaluate the performance of the final developed model.
- Line 254 - It would be helpful if the authors added how the min-max scaling was implemented with respect to training, validation, and test sets.
- Line 286 - It is unclear which limitations related to random forest model the authors are referring to.
- Figure 11 - The SVM model does not appear to have been modeled correctly as its output prediction is always the medium category, hence it has been reduced to a trivial model.
- Line 295 - It appears that the model was selected based on the best performing model on the test set. This indicates that the test set was not used correctly, as the model selection can only be done using validation sets. The test set must only be used to show the performance of an already selected model on it.
- Line 326 - While the authors raise an accurate point about the lack of claims information exceeding $115,000, it is not clear how that data could have benefitted this study since the claims have been bucketed and all those claims greater than $115,000 are already expected to be included in the over-cap category.