https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022

Natural Hazards and Earth System Sciences
Discussions

Development of a Seismic Loss Prediction Model for Residential Buildings using Machine Learning - Christchurch, New Zealand

Samuel Roeslin[1], Quincy Ma[1], Pavan Chigullapally[1], Joerg Wicker[2], and Liam Wotherspoon[1]

[1]Department of Civil and Environmental Engineering, The University of Auckland, Auckland, New Zealand
[2]School of Computer Science, The University of Auckland, Auckland, New Zealand

Correspondence: Samuel Roeslin (s.roeslin@auckland.ac.nz)

Abstract. This paper presents a new framework for the seismic loss prediction of residential buildings in Christchurch, New Zealand. It employs data science techniques, geospatial tools, and machine learning (ML) trained on insurance claims data from the Earthquake Commission (EQC) collected following the 2010-2011 Canterbury Earthquake Sequence (CES). The seismic loss prediction obtained from the ML model is shown to outperform the output from existing risk analysis tools for New Zealand for each of the main earthquakes of the CES. In addition to the prediction capabilities, the ML model delivered useful insights into the most important features contributing to losses during the CES. ML correctly highlighted that liquefaction significantly influenced buildings losses for the 22 February 2011 earthquake. The results are consistent with observations, engineering knowledge, and previous studies, confirming the potential of data science and ML in the analysis of insurance claims data and the development of seismic loss prediction models using empirical loss data.

1 Introduction

In 2010-2011, New Zealand experienced the most damaging earthquakes in its history, known as the Canterbury Earthquake Sequence (CES). It led to extensive damage to Christchurch buildings, infrastructure and its surroundings, affecting both commercial and residential buildings. The entire CES led to over NZ$40 billion in total economic losses. Owing to New Zealand's particular insurance structure, the insurance sector contributed to approximately 80% of the losses for a total of more than NZ$31 billion. NZ$21 billion and NZ$10 billion of the losses as a result of the CES were supported by the private insurers and the Earthquake Commission (EQC) respectively (King et al., 2014; Insurance Council of New Zealand (ICNZ), 2021). Over NZ$11 billion of the losses arose from residential buildings. Approximately 434,000 residential building claims were lodged following the CES and were covered either partially or entirely by the NZ government backed EQCover insurance scheme (Feltham, 2011; Insurance Council of New Zealand (ICNZ), 2021).

In 2010-2011, EQC provided a maximum cover of NZ$100,000 (+GST) per residential building for any homeowner who previously subscribed to a private home fire insurance (New Zealand Government, 2008). In the process of resolving these claims, EQC collected detailed financial loss data, post-event observations, and building characteristics. The CES was also an opportunity for the NZ earthquake engineering community to collect extensive data on the ground shaking levels, soil condi-

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022


Natural Hazards and Earth System Sciences Discussions

tions, and liquefaction occurrence throughout wider Christchurch (Cousins and McVerry, 2010; Cubrinovski et al., 2010, 2011; Wood et al., 2011).

This article presents the development of a seismic loss prediction model for residential building in Christchurch using data science and machine learning (ML). Firstly, a background on ML and some of its applications in earthquake engineering is provided. Key information regarding ML performance and interpretability are also introduced. Then, details regarding the data that was collected following the CES are given. The challenges posed by the raw data are also highlighted. The following section details the merging process required to enrich the data collected. The data preprocessing steps necessary before the application of ML are then described and the paper expands on the actual ML model development. It subsequently describes the algorithm selection, model evaluation, and presents the insights derived from the previously trained ML model. The next section discusses the current limitations and challenges in the application of ML to real-world loss damage data. Finally, the ML performance is compared to outputs from risk analysis tools available for New Zealand.

2 Machine Learning

2.1 Machine learning applied in earthquake engineering

In recent years, the application of ML to real-world problems increased significantly (Sarker, 2021). Similarly, the use of ML in structural and earthquake engineering gained in popularity. Sun et al. (2020) gave a review of ML applications for building structural design and performance assessment and Xie et al. (2020) presented an extensive review of the application of ML in earthquake engineering. A few notable relevant ML studies include the evaluation of post-earthquake structural safety (Zhang et al., 2018), rapid loss assessment (Stojadinović et al., 2021), the derivation of fragility curves (Kiani et al., 2019), quality classification of ground motion records (Bellagamba et al., 2019), classification of earthquake damage to buildings (Mangalathu et al., 2020; Mangalathu and Burton, 2019) and bridges (Mangalathu and Jeon, 2019; Mangalathu et al., 2019).

2.2 Machine learning performance

The performance of a ML model relates to its ability to learn, from a training set, and generalize predictions on unseen data (test data) (Hastie et al., 2009). To achieve this objective, it is important to find a balance between the training error and the prediction error (generalization error). This is known as the bias-variance trade-off (Burkov, 2020; Ng, 2021).

The performance of a ML model might, among other parameters, be improved by using more complex algorithms and feeding more training data to the model. However, despite more training, the accuracy of a ML model will plateau and never surpasses some theoretical limit which is called the Bayes optimal error. It is complex to exactly define where the Bayes optimal error lies for a specific problem. In some cases, the perfect accuracy may not be 100%. Therefore, it is often simpler to compare the accuracy of a ML model to human-level performance on a particular task. In some cases, ML is even capable of surpassing the human-level performance. Before evaluating the actual performance of a ML model on a specific task, it is thus important to clarify the context, background, and current human-level performance.

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022


Natural Hazards and Earth System Sciences Discussions

2.3 Interpretable machine learning

Depending on the aim and purpose of a ML model, obtaining correct predictions only may be satisfactory. However, recent applications of ML showed that interpretability of the model could help the end-user (Honegger, 2018). A ML model can be inspected to identify relationships between input variables, derive insights, and/or find patterns in the data that may be hidden from conventional analysis (Géron, 2019).

Model interpretability is achievable in two main ways. It could come from the possibility for humans to understand the parameters of the algorithm (intrinsic interpretability). This is for example the case for linear regression which remains interpretable due to its simple structure. For complex models, interpretability could come from methods that analyze the ML model after it has been trained (post hoc methods). One method used to explain predictions from ML models is the SHapley Additive exPlanations (SHAP) tool. SHAP is a methodology originally conceived in game theory for computing the contribution of model features to explain the prediction of a specific instance (Lundberg and Lee, 2017). The SHAP methodology has latter been extended to the interpretation of tree-based ML algorithms (Lundberg et al., 2018). It can be used to rank the importance of the model features. SHAP relies on the weight of feature attribution rather than on the study of the decrease in model performance. It is thus more robust compared to the permutation of features in tree-based models (Lundberg et al., 2018; Molnar, 2022). Developing post hoc solutions to make complex model decisions understandable to humans remains a topical research endeavor (Du et al., 2020; Molnar, 2022; Ribeiro et al., 2016a, b, 2018).

3 Data acquisition

3.1 Residential building loss data: EQC claims data set

This study uses the March 2019 version of the EQC claims database. Over 95% of the insurance claims for the CES had been settled by that time. The raw version of the EQC data set contains over 433,500 claims lodged for the CES. Prior to any further data manipulation step, any instance missing information about the building coordinates and unique property identifier were filtered out as these attributes are essential for mapping and merging. This led to a 5% loss in the number of claims related to the CES leaving 412,400 instances in the filtered dataset. However, this includes all the claim statuses, among others, claims that were declined and instances settled on associated claims. To maximize the accuracy of the developed loss prediction model, only claims for which the payment was complete were selected. This ensures that the ML model learns from instances for which the claim amount is final. Figure 1 shows the number of instances for different claim statuses. The selection of the complete claims induced a loss of approximately 50% in the number of instances for the 4 September 2010 and 22 February 2011 events. Figure 2 presents the number of instances for earthquakes in the CES following the selection of settled claims. Prior to merging and data processing, only four events have more than 10,000 instances (i.e., 4 September 2010, 22 February 2011, 13 June 2011, and 23 December 2011). As supervised ML requires a significant amount of data to be able to learn, only those events are selected for the development of the loss prediction model.

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022


Natural Hazards and Earth System Sciences Discussions

The EQC claims data set provided included 62 attributes. The data set contained information such as the date of the event, the opening and closing date of a claim, a unique property number, and the claim amount for the building, content and land. Among the 62 variables, the data set also included information about the building (e.g., construction year, primary construction material, number of stories). However, for those critical features that identify the building characteristics, more than 80% of the instances were not collected as it was not necessary for settlement purposes. The scarce information for building characteristics combined with the necessity to have full data for key variables led to the need to add information from other sources.

3.2 Building characteristics

The RiskScape 'New Zealand Building' inventory data set (RiskScape, 2015) had been adopted by this project to deliver critical information on buildings characteristics. The 'New Zealand Building' inventory collected building asset information for use within the RiskScape software (NIWA and GNS Science, 2017). This data set contained detailed engineering characteristics and other information for every building in New Zealand.

3.3 Seismic demand

A key input for the damage prediction model is the seismic demand for each individual building. This project utilized recordings from the GeoNet strong motion database for the CES earthquakes at 14 strong motion stations located throughout Christchurch (GeoNet, 2012; Kaiser et al., 2017; Van Houtte et al., 2017). Whilst there are many possible metrics to describe the seismic demand, this study focused on using summary data such as the peak ground acceleration (PGA). For this study, the GeoNet data was interpolated across Christchurch for the four main events using the inverse distance weighted (IDW) interpolation implemented in ArcMap (Esri, 2019).

3.4 Liquefaction occurrence

During the CES, extensive liquefaction occurred during four events: 4 September 2010, 22 February 2011, 13 June 2011, and 23 December 2011. The liquefaction and related land damage were the most significant during the 22 February 2011 event. The location and severity of the liquefaction occurrence was based on interpretation from on-site observations and LIDAR surveys. Geospatial data summarizing the severity of the observed liquefaction were sourced from the New Zealand Geotechnical Database (NZGD) (Earthquake Commission (EQC) et al., 2012). The land damage and liquefaction vulnerability due to the CES has been extensively studied. The interested reader is directed to the report from Russell & van Ballegooy (2015).

4 Data merging

The final merging approach made use of the Land Information New Zealand (LINZ) NZ Property Titles data set (Land Information New Zealand (LINZ), 2020a) as an intermediary to constrain the merging process between the EQC and RiskScape data within property boundaries. This was necessary as initial merging attempts using built-in spatial join functions and spatial nearest neighbor joins led to incorrect merging (Roeslin et al., 2020).

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022


Natural Hazards and Earth System Sciences Discussions

As the LINZ NZ Property Titles did not directly include information about the street address, it was first necessary to merge the LINZ NZ Street Address data (Land Information New Zealand (LINZ), 2020b) with the LINZ NZ Property Titles before being able to use the street address information related to a property. Once the LINZ NZ Street Address data (points) and the LINZ NZ Property Titles (polygons) merged, it was found that some properties did not have a matching address point and some properties have multiple address points within one polygon (see Figure 3). Polygons with no address point were filtered out. Properties with multiple address induced challenges regarding the merging of the EQC claims and RiskScape information. The merging process was thus started with instances having a unique street address per property.

The RiskScape database contains information for residential buildings as well as secondary buildings (e.g., external garages, garden shed). Therefore, some properties contain multiple RiskScape points within a LINZ property title (Figure 4). All RiskScape points present in a property were merged to LINZ street address. The data was then filtered to remove points associated with secondary buildings. The RiskScape database includes two variables related to the building size (i.e., building floor area and building footprint). For properties having only two RiskScape points and under the assumption that the principal dwelling is the building with the largest floor area and footprint on a property, it was possible to filter the data to retain RiskScape information related to main dwelling only. Some of the properties have three or more RiskScape points. Automatic filtering of the data using the largest building floor area is unreliable for those instances. In the aim of retaining only trusted data, where one street address had more than three RiskScape instances in a property the data was discarded.

7% of the LINZ property titles have two street address points. As the number of instances used to train a supervised ML model often affects the model accuracy, an attempt was made to retrieve instances that were not collected via the previously mentioned approach. Nevertheless, the philosophy followed here was to put emphasis on the quality of the data rather than the number of points. The effort is focused on retaining the cases when there are two LINZ street addresses and two Riskscape points in the same properties. Following the selection of RiskScape points merged to their unique single LINZ points, the data was appended to the previous RiskScape data set.

Table 1 summarizes the merging steps depending on the number of LINZ street address points and RiskScape points per LINZ property title. While the current selection approach is conservative, it ensured each EQC claim can automatically be assigned to the corresponding residential building using the street address. For cases with multiple street addresses or residential buildings within the same property, a manual assignment of RiskScape points to LINZ street address points would enable the inclusion of more instances. However, this was impractical and was applicable to only 4% of the overall LINZ property titles.

The overall merging process of EQC claims points to LINZ street address points is similar to the process merging RiskScape to LINZ. The limitations related to the combination of the LINZ NZ street address data with the LINZ NZ property titles apply here as well. Hence, it was only possible to merge EQC claims to street address for points contained within LINZ property titles with one street address and to some extent retain claims for properties with two street addresses per title. Once the LINZ NZ street address information added to RiskScape and EQC, these data sets were merged in Python using the street address as a common field.

The final step of preparing the EQC claims data was to add information related to the seismic demand, the liquefaction occurrence, and the soil conditions. This was achieved within ArcMap (Esri, 2019) by importing each of the data sets as a

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022


Natural Hazards and Earth System Sciences Discussions

separate GIS layer. The information contained within each GIS layer was merged with the EQC claims previously combined with RiskScape. Finally, using the street address as a common attribute, the information was combined in one merged data set.

Figure 5 shows the evolution of the number of instances for the 4 September 2010 and 22 February 2011 after each step in the merging process. In its original form, the EQC raw data set entails almost 145,000 claims for the 4 September 2010 and 144,300 claims for 22 February 2011. Following all the aforementioned merging steps, 38,607 usable instances remain for 4 September 2010 and 42,486 instances for 22 February 2011.

5 Data preprocessing

5.1 Feature filtering

Before fitting a ML model to a data set, it is necessary to remove any instance with missing values as many of the ML algorithms are unable to make predictions with missing features. Underrepresented categories within attributes are also carefully examined. Categories with few instances introduce challenges for the ML algorithms as the model will have difficulties "learning" and generalizing for a particular category. In some cases where the meaning is not changed, it is possible to combine instances from different categories. However, whenever a combination of multiple classes is not possible, categories entailing a few instances are removed. This section explains the filtering steps performed on the EQC, RiskScape and additional attributes.

The EQC claims data set contains an attribute specifying the number of dwellings insured on a claim. To avoid any possible issue with the division of the claim value between the multiple buildings, only claims related to one dwelling were retained. Despite the previous selection of claims with the status "Claims Payments Complete" (see section 3.1), another attribute capturing the status of the claims indicated that some selected claims were not closed. To avoid any issues that could be caused by non-closed claims, such instances were discarded. Another important attribute from the EQC data set is the building sum insured. At the date of the CES, EQC provided a maximum cover of NZ$100,000 (+ GST) or NZ$115,000 (including GST) for a residential dwelling for each natural event (Earthquake Commission (EQC), 2019b). To ensure data integrity for the ML model, only the instances with a maximum cover of exactly NZ$115,000 were selected. Finally, two similar attributes related to the claim amount paid were not exactly matching for some claims. To train the ML model on reliable data, instances where the amount indicated by the building paid attribute did not exactly match with the value of building net incurred were excluded from the data set.

Section 4 presented the merging of the EQC data with additional information related to the building characteristics from RiskScape. Building characteristics encompassed the use category, floor area, construction and floor type, wall and roof cladding, and deprivation index. An exploratory analysis of these attributes revealed that initial filtering was required before further use. The use category was the first RiskScape attribute explored. All instances not having the use category defined as residential dwellings were discarded. Once residential dwellings were selected, the size of the building was examined. The analysis of the floor area revealed the presence of outliers, with values reaching up to 3,809 sqm for a single house. To avoid the induction of edge cases in the training set of the ML model, a filtering threshold was set at 1,000 sqm. This led to a minimal loss of instances (0.1%) but eliminated outliers. The following attribute inspected was related to the material of construction.

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022


Natural Hazards and Earth System Sciences Discussions

Figure 6 shows the number of instances for each construction type in the merged data set. Light timber was the most prevalent construction type. Conversely, steel braced frame, light industrial, reinforced concrete (RC) moment-resisting frame, and tilt-up panel only appeared in very few instances. Given that these categories have less than 100 instances, it is unlikely ML models can make correct predictions for those construction types. As a result, these underrepresented categories were filtered out of the data set. Selected, along with light timber dwellings were buildings where the main construction type classified as RC shear wall, concrete masonry, and brick masonry. While the latter category only entails 347 and 371 instances for 4 September 2010 and 22 February 2011 respectively, it was deemed necessary from an engineering point of view to retain brick masonry as possible construction type in the model. Along with the building material, RiskScape also entailed information for the floor type. This attribute had two categories: concrete slab and timber floor. Sufficient instances were present in both categories such that no filtering was required. The wall and roof cladding attributes however, had several underrepresented categories. When possible similar categories were combined together (e.g., fibre cement plank and fibre cement sheet combined together in a category fibre cement) and categories with insufficient entries were discarded (e.g, corrugated iron, plastic, glass). The last attribute sourced from the RiskScape data was the deprivation index. The deprivation index set describes the socioeconomic deprivation of the neighborhood where the building is located. The deprivation index is defined according to ten categories ranging from 1 (least deprived) to 10 (most deprived) (Atkinson et al., 2020). Nine of the ten categories were well represented. Only the category for the deprivation index 10 (most deprived) had a lower 279 instances for 4 September 2010 and 316 for 22 February 2011. Nevertheless, all data was kept in order to capture the full possible range of values related to the deprivation index attribute.

The final merging included information about the seismic demand, liquefaction occurrence, and soil. To ensure that the ML can generalize, the soil types having less than a hundred instances for 4 September 2010 or 22 February 2011 were removed. Each filtering operation induced a loss in the number of instances. Figure 7 documents the evolution of the number of points through the data preprocessing steps.

5.2 Processing of the target attribute

At the time of the CES in 2010-2011, EQC's liability was capped to the first NZ$100,000 (+GST) (NZ$115,000 including GST) of building damage. Costs above this cap were borne by private insurers if building owners previously subscribed to adequate insurance coverage. Private insurers could not disclose information on private claims settlement, leaving the claims database for this study soft-capped at NZ$115,000 for properties with over NZ$100,000 (+GST) damage. Despite the data set having been previously filtered, an exploratory analysis of the attribute 'BuildingPaid' showed that some instances were above NZ$115,000 and other even negative. To be consistent with the coverage of the EQCover insurance, only instances with BuildingPaid between NZ$0 and NZ$115,000 were selected. Figure 8 shows the distribution of 'BuildingPaid' within the selected range for 4 September 2010 and 22 February 2011. Following the filtering of the BuildingPaid attribute, 27,932 instances remained for 4 September 2010 and 27,479 instances for 22 February 2011.

In the original EQC claims data set, 'BuildingPaid' is a numerical attribute. Initial modelling attempts using 'BuildingPaid' as a numerical target variable produced poor model predictions in terms of both accuracy and ability for generalization.

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022


Natural Hazards and Earth System Sciences Discussions

'BuildingPaid' was thus transformed into a categorical attribute. The thresholds for the cut-offs were chosen according to the EQC definitions related to limits for cash settlement, the Canterbury Home Repair Programme, and the maximum coverage provided (Earthquake Commission (EQC), 2019). Any instances with less than and equal to NZ$11,500 was classified as the category 'low', reflecting the limit of initial cash settlement consideration. Next, while the maximum EQC building sum insured was at NZ$115,000, it was found that many instances that were over-cap showed a 'BuildingPaid' value close to but not exactly at NZ$115,000. In consultation with the risk modelling team at EQC, the threshold for the category 'over-cap' was set at NZ$113,850 as this represents the actual cap value (nominal cap value minus 1% excess). Instances with 'BuildingPaid' values between NZ$11,500 and NZ$113,850 were subsequently assigned the category 'medium'. Figure 9 shows the number of instances in each category for 4 September 2010 and 22 February 2011.

6 Model Development

The selected data for the model development included nine attributes (construction type, construction year, floor area, floor type, wall cladding material, deprivation index, PGA, a flag for liquefaction occurrence, soil type) plus the target attribute 'BuildingPaid'. The preprocessed data is complete with no missing value for all the instances.

6.1 Training, validation, and test set

For ML, the data was split into three distinct sets, the training, validation (or development), and test set. Figure 10 shows a schematic overview of the splitting and their use in the development of the ML model. The training and validation sets were coming from the same data set using 80% of the data for training and 20% for validation. The 4 September 2010 preprocessed data had 27,932 instances. Thus, there were 22,345 instances in the training set and 5,587 instances in the validation set. The 22 February 2011 entailed 27,479 instances in total, thus leading to 21,983 examples in the training set and 5,496 in the validation set. The next most represented events were 13 June 2011 and 23 December 2011 (see Figure 2). Instances from those two events were also merged and preprocess to enable their use in the training, validation, and testing process.

Unlike the 'traditional approach' where the test set is held out from the same data as the training and validation set, the test set here employed came from another event in the CES (limited to the four main events). Testing the model using data from another earthquake (preprocessed in the same way as the training and validation set) enabled to evaluate the model capacity to generalize to other events. Thus, changing the earthquake from which the input and test data set comes from, it was possible to study multiple combinations and find the model which generalized best for the entire CES.

6.2 Handling categorical features

Categorical attributes were transformed into binary arrays for adoption by ML algorithms. For the model in this study, strings in categorical features were first transformed into an ordinal integer using the scikit-learn Ordinal encoder (Pedregosa et al., 2011). Once converted to integers, the scikit-learn One Hot Encoder (Pedregosa et al., 2011) was used to encode the categorical features as one-hot numeric array.

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022


Natural Hazards and Earth System Sciences Discussions

6.3 Handling numerical features

Numerical features were checked against each other for correlation prior to the ML training. If two features are correlated, best practice is to remove one of them. The numerical data was also normalized prior to the training process according to best practice. This step is called feature scaling. The most common feature scaling techniques are min-max scaling (also called normalization) and standardization. Both these techniques are implemented in scikit-learn (Pedregosa et al., 2011). In this study, a min-max scaling (normalization) approach was used to scale the numerical features.

6.4 Addressing class imbalance

Figure 9a shows the number of instances for each category in the target variable 'BuildingPaid' for the 4 September 2010 data. While the categories 'low' and 'medium' had respectively 16,558 and 9,970 instances, the category 'over-cap' had only 1,404 instances. The 'over-cap' category was thus the minority class with a significant difference in the number of instances compared to the two other categories. Training a ML algorithm using the data in this form would lead to poor modelling performance for the over-cap category. Thus, before training the model, the imbalanced-learn Python toolbox (Lemaitre et al., 2017) was applied to address the class imbalance. The toolbox encompasses several under-sampling and oversampling techniques, however not all of them apply to multiclass problem. The following over-sampling and under-sampling techniques suitable for multi-class problem were trialed: random oversampling (ROS), cluster centroids (CC) and random undersampling (RUS). For the 4 September 2010 data, ROS delivered the best results regarding the overall model predictions as well as the prediction for the minority class over-cap.

7 Algorithm selection and training

The model was trained using the merged data set which included information on the model attributes as well as the target attribute 'BuildingPaidCat', thus making the training a supervised learning task. Given the nine attributes selected for the model development, the objective of the model was to predict if a building will fall within the category 'low', 'medium', or 'overcap' (expressed via the target variable 'BuildingPaidCat') thus leading to a categorical model for three classes. Several ML algorithms can perform supervised learning task for categories (e.g., logistic regression, support vector machine (SVM), random forest (RF) artificial neural networks (ANN)). Those algorithms differentiate themselves by their complexity. More complex algorithms can develop more detailed models with a potential improved prediction performance, but complex algorithms are also more prone to overfitting. For this study, the prediction performance was an essential metric. Nevertheless, the human interpretability of the model was also of significant interest. The goal was to produce a 'greybox' model enabling for the derivation of insights. In this project, the logistic regression, decision trees, SVM, random forest were trialed.

As mentioned in section 6.1, training data was obtained from the four main events in the CES (4 September 2010, 22 February 2011, 13 June 2011, 23 December 2011). Once the model trained, the validation set was used for the tuning of the

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022


Natural Hazards and Earth System Sciences Discussions

hyperparameters. This paper only presents outputs and findings for the 4 September 2010 and 22 February 2011. For findings related to the 13 June 2011 and 23 December 2011, the reader is directed to Roeslin (2021).

8 Model evaluation

Figure 11 shows confusion matrices for the logistic regression, decision trees, SVM, and random forest trained and validated on data from 22 February 2011. For each confusion matrix, the diagonal in the green area represents the correct predictions. The top integer numbers in each of the upper left boxes display the number of instances predicted, and the percentage in the bottom rows represent that instance as a percentage of the population. The closest the value on the diagonal sum to 100%, the better the prediction. Mistakenly predicted instances are shown off the diagonal. Despite limitations, random forest showed the best overall prediction performance and was deemed as the best performing algorithm in this study.

Figure 10 showed the process for the model development. Each model was tested on instances from the other main events of the CES. Figure 12a and Figure 12d show the confusion matrix for the random forest model for the 4 September 2010 and 22 February 2011 validated on the same event respectively. Figure 12b shows the confusion matrix for the random forest model developed with the 4 September 2010 data and tested on the 22 February 2011 instances. Figure 12c presents the confusion matrix for the random forest model developed with the 22 February 2011 data and tested on the 4 September 2010 instances. Similar confusion matrices were generated for the model trained 13 June 2011 and 23 December 2011. All combinations and permutations between the models trained on the four mains events in the CES were tested. It was found that the model trained on data from the 22 February 2011 performs the best on testing data from other main events in the CES and was thus deemed as the model which generalized best. Nevertheless, despite the thorough attribute filtering, attribute selection, attribute preparation, and model development addressing class imbalance and carefully checking for under- and over-fitting, the prediction accuracy of the random forest algorithm on any test set did not exceed 0.62.

9 Insights

The SHapley Additive exPlanations (SHAP) post-hoc method was applied on the random forest models for analyzing the relative influence of the different input features. Figure 13 shows the SHAP feature importance for the random forest models for 4 September 2010 and 22 February 2011. The influence of PGA on the residential building losses was highlighted for all the key events of the CES. This validated the probabilistic seismic loss estimation methodology which relies on PGA and the spectral acceleration at selected periods as intensity measures (IM) as the key input. It was satisfying to observe that ML, which has no physical understanding or prior knowledge related to building damage and loss, was capable of capturing the importance of PGA from empirical data alone.

For the 22 February 2011, PGA significantly stood out and was followed by the liquefaction occurrence and soil type. It thus seemed that the building damage and losses due to the 2011 Christchurch earthquake were driven by liquefaction. This result corroborated the findings from previous studies, which highlighted the influence of liquefaction on building damage

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022


Natural Hazards and Earth System Sciences Discussions

for the 22 February 2011 event (Rogers et al., 2015; Russell and van Ballegooy, 2015). The year of construction appeared second for the 4 September 2010 event, however, it was only fifth for the 22 February 2011 event. It is possible that the feature 'ConstructionYear' captured information related to the evolution of the seismic codes which appears more significant for the events less affected by liquefaction.

The study of the feature importance of the ML models seemed to distinguish two types of events: shaking dominated events such as the 4 September 2010 event and liquefaction dominated like the 22 February 2011 earthquake.

10 Current challenges

There are numerous possible reasons for the limited ML model accuracy. Some are listed here.

Having more direct information collected on-site about the building characteristics would improve the completeness of the EQC data set, which could benefit the model performance.

The issues faced during the merging of the EQC dataset with RiskScape building characteristics and LINZ information highlighted the need for an improved solution to identify each building in New Zealand. It is believed that the establishment of a unique building identifier common to several databases will introduce consistency, thus opening new opportunities for the application of data science techniques and the derivation of insights.

At the time of the CES, EQC only provided building coverage up to NZ$100,000 (+GST) which led to the EQC dataset being capped at NZ$115,000. Losses above the NZ$115,000 threshold were covered by private insurers, given that the building owner subscribed to appropriate private insurance. Any detail for building loss above NZ$115,000 was not available for this study. For 4 September 2010, there was a significant class imbalance between the classes of the target variable with over-cap instances being mostly underrepresented. Despite the use of the Python imbalance toolbox to address the imbalance, having more instances in the over-cap category would be beneficial. The access to data from private insurances would enlarge the range of the target attribute 'BuildingPaid' giving more information on the buildings which suffered significant losses.

A more in depth analysis of the actual value of 'BuildingPaid' might also bring an improved model performance. Taking into account apportionment between the events in the CES would provide a more accurate allocation of loss to each event and enable to capture more details about over-cap instances. To mitigate issues related to sequential damage throughout the CES, the data could be segregated by geographical area where the majority of damage occurred for each event. This might lead to a "cleaner" training set and thus might deliver more accurate predictions.

The prediction accuracy also depends on the attributes present in the model. Section 6 presented the target variable and nine selected model attributes. These attributes were selected based on domain knowledge as possible features that could affect the building losses. There may be other attributes that were not considered in this study that have direct and indirect impacts on the value of a claim. It is thus possible that the inclusion of additional attributes might be beneficial to the overall model accuracy. The introduction of additional parameters related to properties and social factors for example might deliver an improved model accuracy as well as new insights.

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022


Natural Hazards and Earth System Sciences Discussions

11 ML loss model performance vs current tools

Section 8 showed that for all the models trained on data from the main events of the CES, random forest the best performing algorithm reached a maximum accuracy of 0.62. Section 2.2 highlighted the importance of providing context and information related to the maximum achievable performance of ML for a specific task. While it was difficult to give an exact value of the Bayes error for this task due to the inherent complexity of loss prediction, it was possible to compare the accuracy of the developed ML model to the performance of current tools employed for the damage and loss prediction.

The outputs of the ML model were compared to predictions obtained from the RiskScape v1.0.3 software (NIWA and GNS Science, 2017). Loss prediction scenarios for the 4 September 2010 and 22 February 2011 were performed in RiskScape using the hazard information and building data available within the software. RiskScape output loss predictions for all the buildings in the Canterbury region. To enable a comparison, samples of 26,500 residential buildings located in Christchurch were selected for both the 4 September 2010 and 22 February 2011 events. The buildings in the sample were carefully selected to only encompass buildings for which at least one claim was lodged to EQC during the CES. This later enabled the comparison of the RiskScape software predictions to the actual level of building loss captured by EQC. To obtain prediction from ML, the samples were then passed through the ML model previously trained on 22 February 2011 as it was found that this model generalizes better for the CES.

Table 2 shows an overview of the accuracy of the ML loss model and RiskScape for the selected building sample. Despite limitations in the ML model, it significantly outperformed the accuracy from the RiskScape predictions.

12 Conclusions

This paper introduced a new framework for the seismic loss prediction of residential buildings. It used residential building insurance claims data collected by the Earthquake Commission following the 2010-2011 Canterbury earthquake sequence to train a machine learning model for the loss prediction in residential buildings in Christchurch, New Zealand. The random forest algorithm trained on claims data from 22 February 2011 delivered the most promising outputs. Results from the machine learning model were compared to the performance of current tools for loss modeling. Despite limitations, it was found that the machine learning model outperformed loss predictions obtained using the RiskScape software. It was also shown that machine learning was capable of extracting the most important features that contributed to building loss.

Overall, this research project demonstrated the capabilities and benefits of applying machine learning to empirical data collected following earthquake events. It showed that machine learning was able to extract useful insights from real-world data and outperformed current tools employed for the damage and loss prediction of buildings. It confirmed that data science techniques and machine learning are appropriate tools for the development of seismic loss prediction models.

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022

Natural Hazards and Earth System Sciences
Discussions
Open Access
EGU

Acknowledgements. We acknowledge the Earthquake Commission, especially the Risk Modelling team for the help with data interpretation.

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022


Natural Hazards and Earth System Sciences Discussions

References

Atkinson, J., Salmond, C., and Crampton, P.: NZDep2018 Index of Deprivation, Final Research Report, Tech. Rep. December, University of Otago, Wellington, New Zealand, https://www.otago.ac.nz/wellington/departments/publichealth/research/hirp/otago020194.html, 2020.

Bellagamba, X., Lee, R., and Bradley, B. A.: A neural network for automated quality screening of ground motion records from small magnitude earthquakes, Earthquake Spectra, 35, 1637–1661, https://doi.org/10.1193/122118eqs292m, 2019.

Burkov, A.: Machine Learning Engineering, True Positive Inc., 2020.

Cousins, J. and McVerry, G. H.: Overview of strong-motion data from the Darfield earthquake, Bulletin of the New Zealand Society for Earthquake Engineering, 43, 222–227, https://doi.org/10.5459/bnzsee.43.4.222-227, 2010.

Cubrinovski, M., Green, R. A., Allen, J., Ashford, S., Bowman, E., Bradley, B., Cox, B., Hutchinson, T., Kavazanjian, E., Orense, R., Pender, M., Quigley, M., and Wotherspoon, L.: Geotechnical reconnaissance of the 2010 Darfield (Canterbury) earthquake, Bulletin of the New Zealand Society for Earthquake Engineering, 43, 243–320, https://doi.org/10.5459/bnzsee.43.4.243-320, 2010.

Cubrinovski, M., Bradley, B., Wotherspoon, L., Green, R., Bray, J., Wood, C., Pender, M., Allen, J., Bradshaw, A., Rix, G., Taylor, M., Robinson, K., Henderson, D., Giorgini, S., Ma, K., Winkley, A., Zupan, J., O'Rourke, T., DePascale, G., and Wells, D.: Geotechnical aspects of the 22 February 2011 Christchurch earthquake, Bulletin of the New Zealand Society for Earthquake Engineering, 44, 205–226, https://doi.org/10.5459/bnzsee.44.4.205-226, 2011.

Du, M., Liu, N., and Hu, X.: Techniques for interpretable machine learning, Communications of the ACM, 63, 68–77, https://doi.org/10.1145/3359786, 2020.

Earthquake Commission (EQC): EQC Insurance, https://www.eqc.govt.nz/what-we-do/eqc-insurance, 2019.

Earthquake Commission (EQC), Ministry of Business Innovation and Employment (MBIE), and New Zealand Government: New Zealand Geotechnical Database (NZGD), https://www.nzgd.org.nz/Default.aspx, 2012.

Esri: ArcGIS Desktop 10.7.1, 2019.

Feltham, C.: Insurance and reinsurance issues after the Canterbury earthquakes, Parliamentary Library Research Paper, pp. 1–2, 2011.

GeoNet: GeoNet strong-motion FTP site, ftp://ftp.geonet.org.nz/strong/processed/, 2012.

Géron, A.: Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, O'Reilly Media, 2 edn., 2019.

Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning, Springer Series in Statistics, Springer, New York, 2009.

Honegger, M.: Shedding Light on Black Box Machine Learning Algorithms, Ph.D. thesis, Karlsruhe Institue of Technology, Germany, 2018.

Insurance Council of New Zealand (ICNZ): Canterbury Earthquakes, https://www.icnz.org.nz/natural-disasters/canterbury-earthquakes/, 2021.

Kaiser, A., Van Houtte, C., Perrin, N., Wotherspoon, L., and Mcverry, G.: Site Characterisation of GeoNet Stations for the New Zealand Strong Motion Database, Bulletin of the New Zealand Society for Earthquake Engineering, 50, 39–49, https://doi.org/10.5459/bnzsee.50.1.39-49, 2017.

Kiani, J., Camp, C., and Pezeshk, S.: On the application of machine learning techniques to derive seismic fragility curves, Computers & Structures, https://doi.org/10.1016/j.compstruc.2019.03.004, 2019.

King, A., Middleton, D., Brown, C., Johnston, D., and Johal, S.: Insurance: Its Role in Recovery from the 2010–2011 Canterbury Earthquake Sequence, Earthquake Spectra, 30, 475–491, https://doi.org/10.1193/022813EQS058M, 2014.

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022

Land Information New Zealand (LINZ): LINZ Data Service – NZ Property Titles, https://data.linz.govt.nz/layer/50804-nz-property-titles/, 2020a.

Land Information New Zealand (LINZ): LINZ Data Service – NZ Street Address, https://data.linz.govt.nz/layer/53353-nz-street-address/, 2020b.

Lemaitre, G., Nogueira, F., and Aridas, C. K.: Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning, Journal of Machine Learning Research, 18, 559–563, 2017.

Lundberg, S. M. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, in: Advances in Neural Information Processing Systems 30, edited by Garnett, I. G., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., pp. 4765–4774, Curran Associates, Inc., 2017.

Lundberg, S. M., Erion, G. G., and Lee, S.-I.: Consistent Individualized Feature Attribution for Tree Ensembles, 2017 ICML Workshop, http://arxiv.org/abs/1802.03888, 2018.

Mangalathu, S. and Burton, H. V.: Deep learning-based classification of earthquake-impacted buildings using textual damage descriptions, International Journal of Disaster Risk Reduction, 36, 101 111, https://doi.org/10.1016/j.ijdrr.2019.101111, 2019.

Mangalathu, S. and Jeon, J.-S.: Machine Learning–Based Failure Mode Recognition of Circular Reinforced Concrete Bridge Columns: Comparative Study, Journal of Structural Engineering, 145, 04019 104, https://doi.org/10.1061/(ASCE)ST.1943-541X.0002402, 2019.

Mangalathu, S., Hwang, S. H., Choi, E., and Jeon, J. S.: Rapid seismic damage evaluation of bridge portfolios using machine learning techniques, Engineering Structures, 201, 109 785, https://doi.org/10.1016/j.engstruct.2019.109785, 2019.

Mangalathu, S., Sun, H., Nweke, C. C., Yi, Z., and Burton, H. V.: Classifying earthquake damage to buildings using machine learning, Earthquake Spectra, 36, 183–208, https://doi.org/10.1177/8755293019878137, 2020.

Molnar, C.: Interpretable Machine Learning. A Guide for Making Black Box Models Explainable, Independently published, https://christophm.github.io/interpretable-ml-book/, 2022.

New Zealand Government: Earthquake Commission Act 1993 - 1 April 2008, 2008.

Ng, A.: CS230 Deep Learning - C3M1: ML Strategy (1), https://cs230.stanford.edu/files/C3M1.pdf, 2021.

NIWA and GNS Science: RiskScape, https://www.riskscape.org.nz/, 2017.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12, 12, 2825–2830, https://doi.org/10.1007/s13398-014-0173-7.2, 2011.

Ribeiro, M. T., Singh, S., and Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144, ACM, https://doi.org/10.18653/v1/N16-3020, 2016a.

Ribeiro, M. T., Singh, S., and Guestrin, C.: Model-Agnostic Interpretability of Machine Learning, in: 2016 ICML Workshop on Human Interpretability in Machine, 2016b.

Ribeiro, M. T., Singh, S., and Guestrin, C.: Anchors: High-Precision Model-Agnostic Explanations, in: Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18), pp. 1527–1535, 2018.

Roeslin, S.: Predicting Seismic Damage and Loss for Residential Buildings using Data Science, Ph.D. thesis, University of Auckland, Auckland, New Zealand, https://hdl.handle.net/2292/57074, 2021.

Roeslin, S., Ma, Q., Wicker, J., and Wotherspoon, L.: Data Integration for the Development of a Seismic Loss Prediction Model for Residential Buildings in New Zealand, in: Machine Learning and Knowledge Discovery in Databases, edited by Cellier, P. and Driessens, K., vol.

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022

1168 of *Communications in Computer and Information Science*, pp. 88–100, Springer, Cham, Switzerland, https://doi.org/10.1007/978-3-030-43887-6_8, 2020.

Rogers, N., van Ballegooy, S., Williams, K., and Johnson, L.: Considering Post-Disaster Damage to Residential Building Construction - Is Our Modern Building Construction Resilient?, in: Proceedings of 6th International Conference on Earthquake Geotechnical Engineering, Christchurch, New Zealand, 2015.

Russell, J. and van Ballegooy, S.: Canterbury Earthquake Sequence: Increased Liquefaction Vulnerability assessment methodology, Tech. rep., Tonkin & Taylor Ltd, Auckland, New Zealand, https://www.eqc.govt.nz/ILV-engineering-assessment-methodology, 2015.

Sarker, I. H.: Machine Learning: Algorithms, Real-World Applications and Research Directions, SN Computer Science, 2, https://doi.org/10.1007/s42979-021-00592-x, 2021.

Stojadinović, Z., Kovačević, M., Marinković, D., and Stojadinović, B.: Rapid earthquake loss assessment based on machine learning and representative sampling, Earthquake Spectra, p. 875529302110423, https://doi.org/10.1177/87552930211042393, 2021.

Sun, H., Burton, H. V., and Huang, H.: Machine Learning Applications for Building Structural Design and Performance Assessment: State-of-the-Art Review, Journal of Building Engineering, https://doi.org/10.1016/j.jobe.2020.101816, 2020.

Van Houtte, C., Bannister, S., Holden, C., Bourguignon, S., and Mcverry, G.: The New Zealand strong motion database, Bulletin of the New Zealand Society for Earthquake Engineering, 50, 1–20, https://doi.org/10.5459/bnzsee.50.1.1-20, 2017.

Wood, C. M., Cox, B. R., Wotherspoon, L. M., and Green, R. A.: Dynamic site characterization of Christchurch strong motion stations, Bulletin of the New Zealand Society for Earthquake Engineering, 44, 195–204, https://doi.org/10.5459/bnzsee.44.4.195-204, 2011.

Xie, Y., Ebad Sichani, M., Padgett, J. E., and DesRoches, R.: The promise of implementing machine learning in earthquake engineering: A state-of-the-art review, Earthquake Spectra, p. 33, https://doi.org/10.1177/8755293020919419, 2020.

Zhang, Y., Burton, H. V., Sun, H., and Shokrabadi, M.: A machine learning framework for assessing post-earthquake structural safety, Structural Safety, 72, 1–16, https://doi.org/10.1016/j.strusafe.2017.12.001, 2018.

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022

Natural Hazards and Earth System Sciences
Discussions
Open Access
EGU

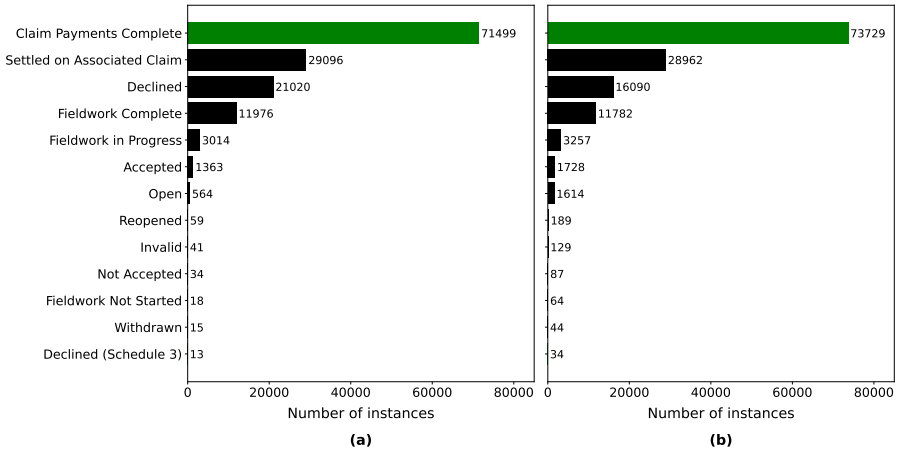


Figure 1. Number of instances grouped by the status of the claim: (a) 4 September 2010, (b) 22 February 2011

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022

Natural Hazards and Earth System Sciences
Discussions
Open Access
EGU

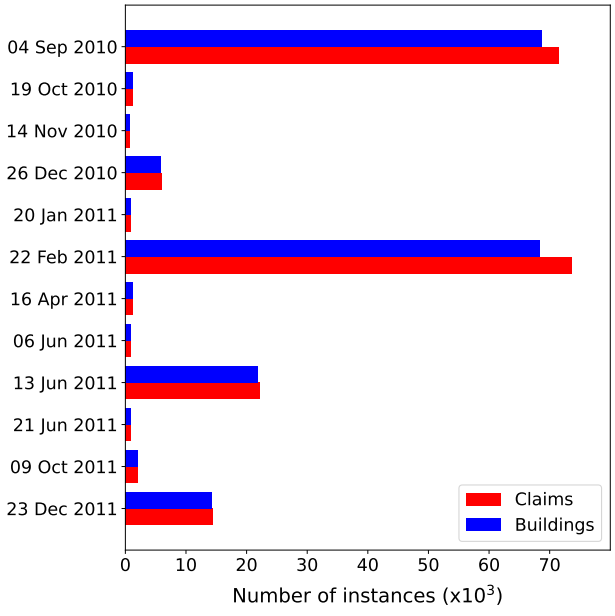


Figure 2. Number of claims and property for events in the CES after filtering for ClaimStatus. Only events with more than 1,000 instances prior to cleansing are shown.

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022

Natural Hazards and Earth System Sciences
Discussions

Figure 3. Satellite image of urban blocks in Christchurch overlaid with the LINZ NZ Street Address and LINZ NZ Property Titles layers. The polygons with a bold red border represent LINZ NZ property titles having only one street address (from Eagle Technology Group Ltd (NZ Esri Distributor)).

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022

Natural Hazards and Earth System Sciences
Discussions
Open Access
EGU

Figure 4. Satellite view of an urban block in Christchurch with RiskScape points and selected LINZ NZ Property Titles (from Eagle Technology Group Ltd (NZ Esri Distributor)).

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022

Natural Hazards and Earth System Sciences
Discussions
Open Access
EGU

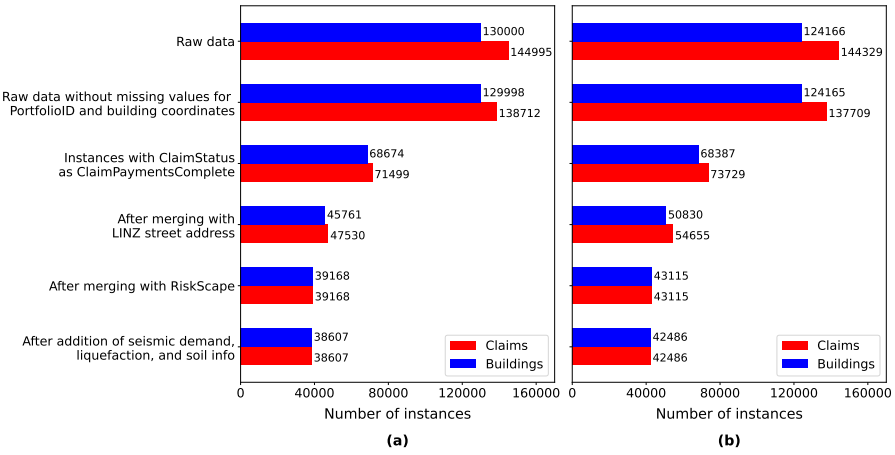


Figure 5. The number of data points after each processing step for event on 4 September 2010 and 22 February 2011: (a) 4 September 2010, (b) 22 February 2011

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022

Natural Hazards and Earth System Sciences
Discussions
Open Access
EGU

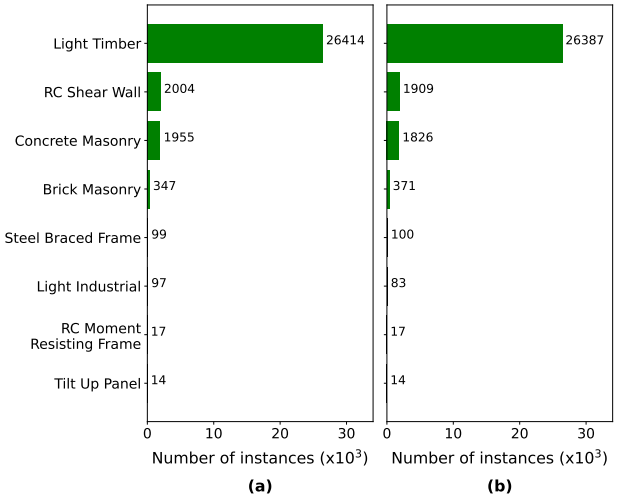


Figure 6. Number of instances for each Construction Type category: (a) 4 September 2010, (b) 22 February 2011

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022

Natural Hazards and Earth System Sciences
Discussions
Open Access
EGU

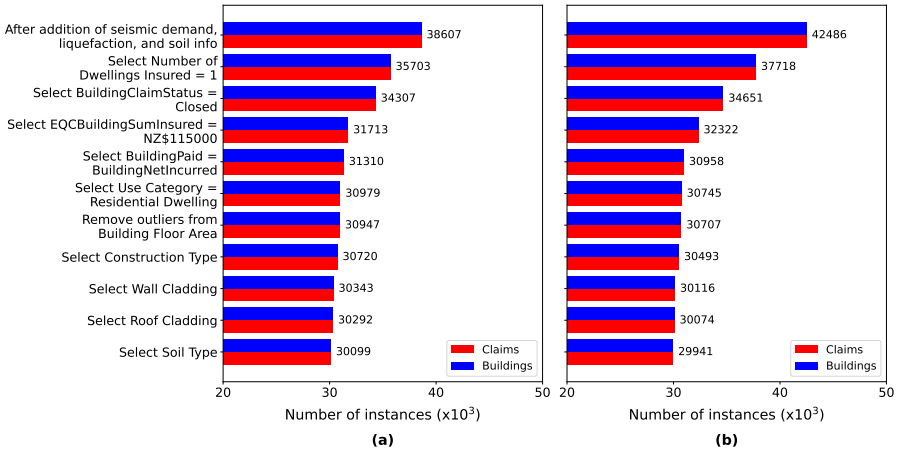


Figure 7. Evolution of the number of instances after each feature filtering step: (a) 4 September 2010, (b) 22 February 2011

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022

Natural Hazards and Earth System Sciences
Discussions
Open Access
EGU

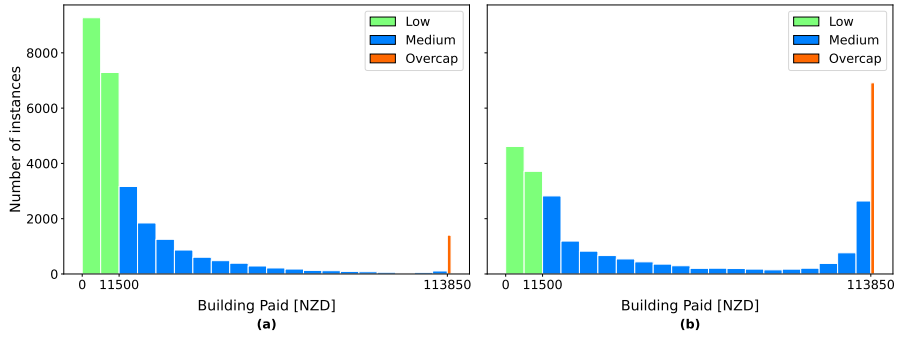
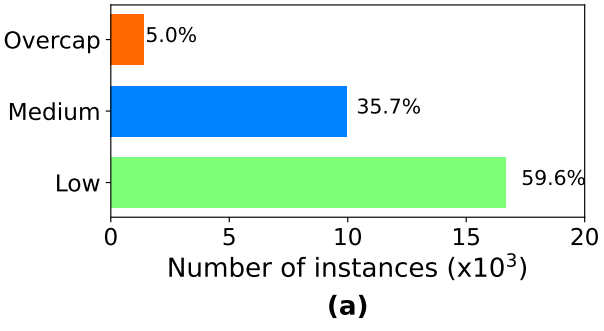


Figure 8. Distribution of BuildingPaid after selection of the instances between NZ$0 and NZ$115,000: (a) 4 September 2010, (b) 22 February 2011

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022



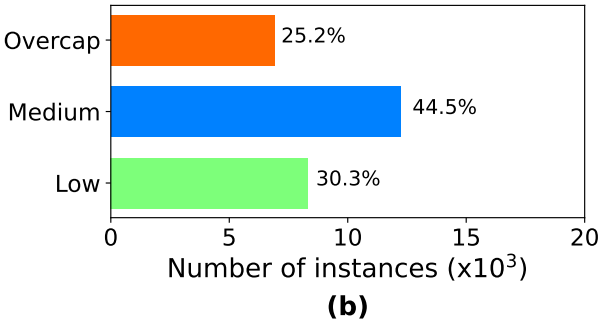


Figure 9. Number of instances in BuildingPaid categorical in the filtered data set: (a) 4 September 2010, (b) 22 February 2011

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022

Natural Hazards and Earth System Sciences
Discussions
Open Access
EGU

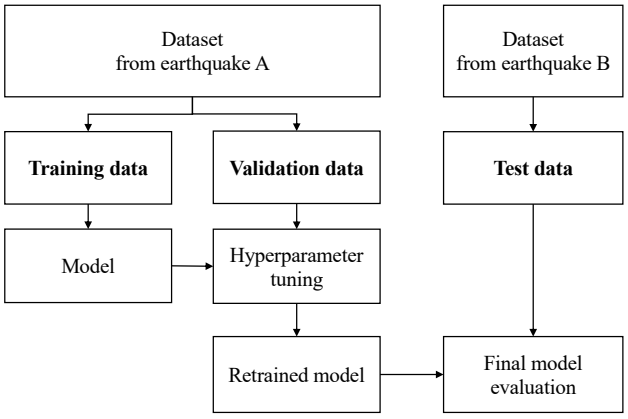


Figure 10. Overview of the training, validation, and test data sets and their usage in the development of a ML seismic loss model for Christchurch

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022

Natural Hazards and Earth System Sciences
Discussions

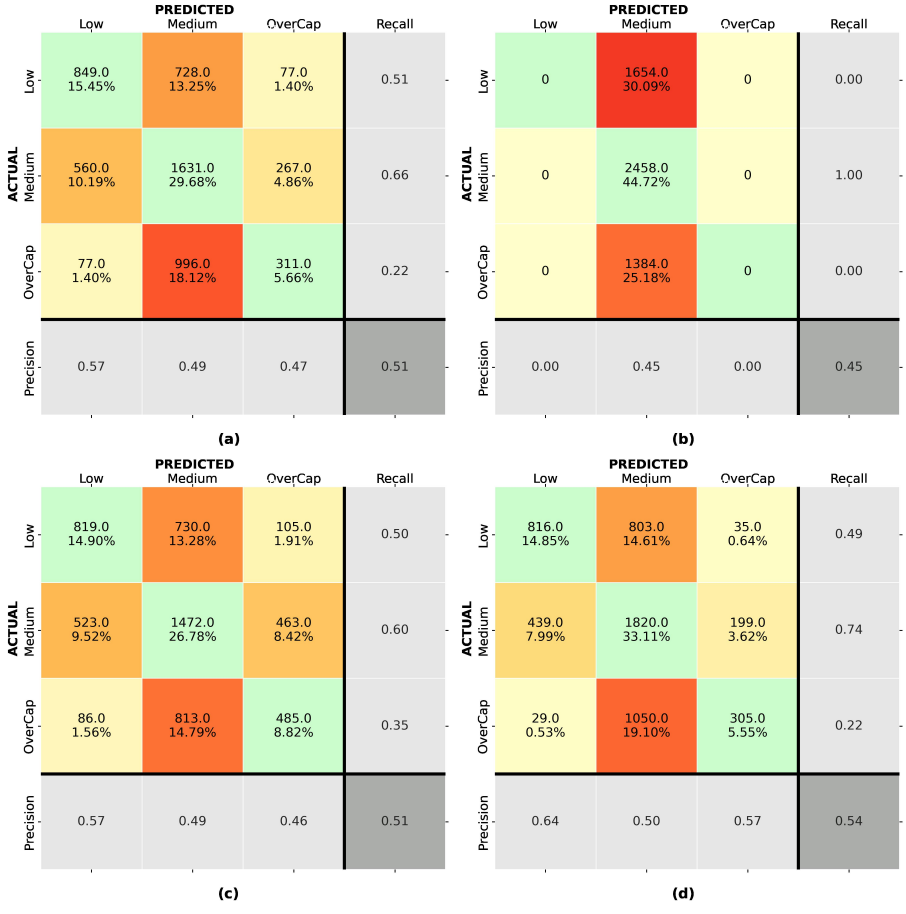


Figure 11. Confusion matrices for models trained and validated on data from 22Feb2011: (a) Logistic Regression, (b) Support Vector Machine (SVM), (c) Decision Tree, (d) Random Forest

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022

Natural Hazards and Earth System Sciences
Discussions

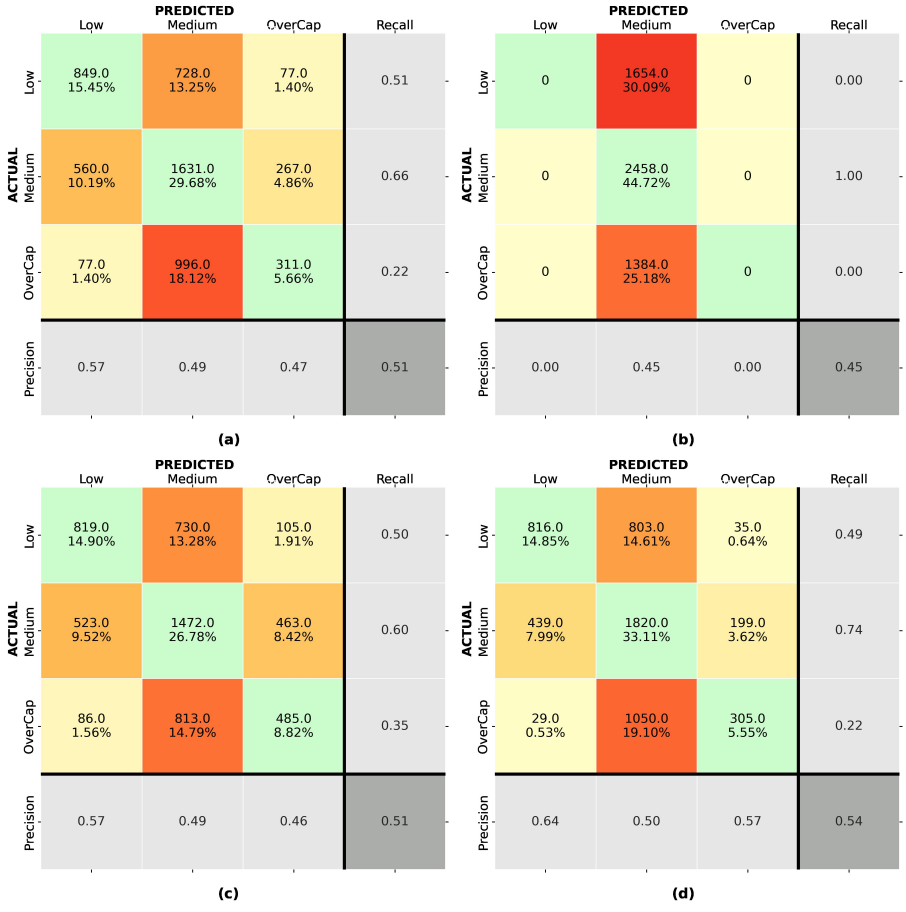


Figure 12. Confusion matrices for the random forest algorithm: (a) 4 September 2010 model tested on 4 September 2010, (b) 4 September 2010 model tested on 22 February 2011, (c) 22 February 2011 model tested on 4 September 2010, (d) 22 February 2011 model tested on 22 February 2011

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022

Natural Hazards and Earth System Sciences
Discussions

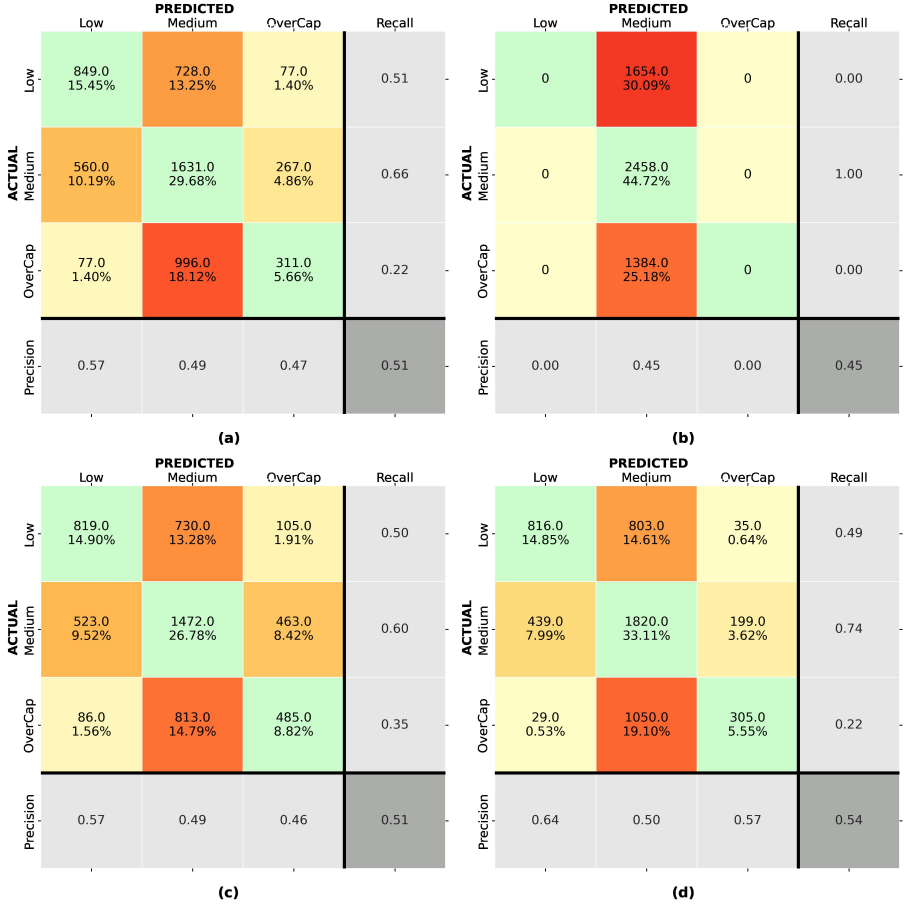


Figure 13. SHAP feature importance for the random forest model: (a) 4 September 2010, (b) 22 February 2011

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022

Natural Hazards and Earth System Sciences
Discussions

Table 1. Overview of the action taken depending on the number of LINZ NZ street address and RiskScape point present per LINZ NZ property title

LINZ NZ street address	RiskScape	Action
1 point per LINZ property title	1 point per property title	Direct selection
1 point per LINZ property title	2 points per property title	Select the RiskScape point with the largest building floor area
1 point per LINZ property title	3 or more points per property title	Discarded
2 points per LINZ property title	1 point per property title	The automatic selection and filtering did not retain those instances as it could not differentiate this specific case
2 points per property LINZ title	2 points per property title	Retain these instances based on "spatial join" (closest) combined with filtering.
2 points per property LINZ title	3 or more points per property title	Discarded
3 or more points per LINZ property title	Any configuration	Discarded

https://doi.org/10.5194/nhess-2022-227
Preprint. Discussion started: 1 September 2022

Natural Hazards and Earth System Sciences
Discussions

Table 2. Comparison of the accuracy of the ML model vs RiskScape v1.0.3 for the 4 September 2010 and 22 February 2011 events for a sample of 26,500 buildings located in Christchurch, New Zealand

Event	ML model	RiskScape
4 Sep 2010	61.7%	48.3%
22 Feb 2021	57.5%	35.0%