INTRODUCTION
In 2010-2011, New Zealand experienced the most damaging earthquakes in its history, known as the Canterbury Earthquake sequence (CES). The CES led to extensive damage to Christchurch buildings, infrastructure, and its surroundings, affecting commercial and residential buildings. The total economic losses of more than NZ$40 billion (1) accounted for 20% of New Zealand’s GDP (2). Owing to New Zealand’s particular insurance structure, the insurance sector contributed to over 80% of losses for a total of more than NZ$31 billion (3). Losses from residential building accounted for 50% of the total economic losses (4). The residential building losses were covered either partially or entirely from the NZ government-backed Earthquake Commission (EQC) cover insurance scheme. Following the CES, EQC collected detailed financial loss data and building characteristics for more than 500,000 claims (5).

EQC DATA
- EQC claim data set is a wide dataset with 62 features.
- For some of the attributes (e.g. construction year, primary construction material, number of stories), more than 80% of the data points was not collected as it was not necessary for settlement purposes.
- The first step of data cleansing consisted of selecting claims where the payment is complete.
- After the claim status selection, four earthquake events remain with enough instances to apply machine learning.
  - 4 September 2010 event
  - 22 February 2011 event
  - 13 June 2011 event
  - 23 December 2011 event
- Even if the claims are related to one event, the amounts paid or repaired may represent damage from multiple events (due to the short time between events resulting in ambiguity about which event caused the damage).

DATA Merging
- The primary database from the Earthquake Commission (EQC) contains property information and insurance claim data for residential buildings.
- This project merged additional information from private and open-source databases on top of EQC’s claim database.
  - Building characteristics from RiskScape
  - Liquefaction occurrence from the New Zealand Geotechnical database
  - Peak Ground Acceleration (PGA) from GeoNet
  - Soil conditions Land Resource Information Systems (LRIS)
- The data integration was challenging due to the non-presence of a common feature between EQC and RiskScape. The merging was performed using the building location. Nevertheless, the merging process entailed limitations which led to the loss of instances (6).

DATA PRE-PROCESSING
- In the original EQC claim data set ‘Building Paid’ is a numerical attribute.
- Prediction for a regression machine learning model using ‘BuildingPaid’ as a numerical target variable did not deliver satisfactory outputs.
- Data pre-processing included the transformation of ‘BuildingPaid’ from a numerical attribute to categories.
- Thresholds for the cut-off were chosen according to the EQC definitions related to the cash settlement of the claim, the Canterbury Home Repair Programme, and the maximum coverage provided (7).

TARGET ATTRIBUTE
- 4 September 2010:
  - 59.3% of the claims were low
  - 35.7% of the claims were medium
  - 5.0% of the claims reached the maximum cap
- 22 February 2011:
  - 44.5% of the claims were medium
  - 30.3% of the claims were low
  - 25.2% of the claims reached the maximum cap.

DATA PREPARATION
- Before starting training a machine learning model, it is necessary to split the data into distinct sets known as the training, validation (or development), and test set.
- Unlike a ‘traditional’ approach where the test set is held out from the same data as the training and validation set, the test set here employed comes from another earthquake.
- Testing the model using data from another earthquake in the CES (pre-processed in the same way as the training and validation set) enables to evaluate the model capacity to generalise to other events and find the model which works the best for the entire CES.

MACHINE LEARNING
- The availability of the target and observation makes this project a supervised learning problem for classification.
- Four algorithms were trialled: logistic regression, decision trees, SVM and random forest.
- Random forest is the best performing algorithm.

INSIGHTS
- The Shapley additive exPlanations (SHAP) post-hoc method was applied on the Random Forest models.
- PGA stands out as being the most important feature for all events.
- The liquefaction occurrence is second for 22 February 2011 model and fourth for 4 September 2010 model confirming the influence of liquefaction on the building damage/loss.
- The construction year and the floor area of the building appear in the top five most important features, however at a different position depending on the event.

CURRENT CHALLENGES/FUTURE WORK
- Machine learning requires complete and clean data.
- Key building characteristics are missing in the initial EQC claim database.
- Need for a unique building identifier to facilitate the merging of information.
- Once developed, a machine learning pipeline can easily be retrained. This facilitates future studies employing different combinations of building parameters.
- Taking into account appraisal(6) would provide a more accurate allocation of loss to each event and enable to capture more details about over cap instances.
- For each event, segregating the data by geographical area where the majority of damage occurred might lead to a “cleaner” train set and thus might lead to more accurate predictions.
- The introduction of additional parameters related to properties and social factors might deliver an improved model accuracy as well as new insights.

CONCLUSION
- We acknowledge the Earthquake Commission (EQC), especially the Risk Modeling Team for the help with data interpretation.
- This research was financed by New Zealand’s Marsden Fund (grant number 18-UOA-171).