

# Machine learning application to disaster damage repair cost modelling of residential buildings

Nadeeshani Wanigarathna<sup>a</sup>, Ying Xie<sup>b</sup>, Christian Henjewe<sup>a</sup>, Mariantonietta Morga<sup>a</sup> and Keith Jones<sup>a</sup>

<sup>a</sup>School of Engineering and the Built Environment, Anglia Ruskin University, Chelmsford, UK; <sup>b</sup>Cranfield School of Management, Cranfield University, Bedfordshire, UK

## ABSTRACT

Restoring residential buildings following earthquake damage requires a significant level of resources. Being able to predict these resource requirements in advance and accurately improves the effectiveness of disaster preparedness and subsequent recovery activities. This research explored how the latest ML algorithms could be used for antecedent earthquake loss modelling. A cost database for repairing residential buildings damaged by the Emilia Romagna earthquake in Italy was analysed using six state-of-the-art ML models to explore their ability to predict repair cost rates (cost per floor area) for a domestic building damaged by earthquakes. A Gradient Boost Regression model outperformed five other models in predicting earthquake damage repair cost rate. The performance of this model was significantly accurate and covers about 76% of the cases. A further SHAP analysis revealed that operational level, damage level and non-housing area of the buildings as top 3 important features when predicting the resultant damage repair cost rate. Overall this research advanced antecedent earthquake loss modelling approaches to increase the accuracy of estimates by incorporating more variables than the widely used damage level based simple methodology.

## ARTICLE HISTORY

Received 21 December 2023  
Accepted 16 October 2024

## KEYWORDS

Earthquake; cost modelling; machine learning; damage repair costs; disaster preparedness

## Introduction

Earthquakes cause devastating impacts on individuals, communities and nations at large. Preparing for a potential future earthquake could help people to effectively anticipate, respond to and recover from the impacts quickly when disasters strike (Sutton and Tierney 2006, UNISDR 2017). Preparation aims to build the capacities needed to manage disasters efficiently (UNDRR, nd). Financial capability is vital in effectively responding to a disaster (Wanigarathna *et al.* 2022). Antecedent loss assessments during the preparedness phase help individuals, organisations and governments to understand the potential financial capacity required following a disaster (Yu *et al.* 2015). Antecedent disaster loss assessment refers to assessing the economic value of damage resulting from disasters before an earthquake occurs as a part of disaster preparedness activities. Disaster effects and damage could be in the form of economic losses resulted by direct damage to physical assets such as building, infrastructure or indirectly through business interruption; environmental losses resulted by damage to watercourses, eco

systems and habitats), social losses (deaths and injuries, increase in crime, family violence etc), and heritage losses (such as damage to cultural, historic and world heritage assets) (De Groeve *et al.* 2015). The focus of this research is limited to direct losses associated with restoration costs. Reconstruction in this paper refers to all forms of restoration to damage buildings. Antecedent loss assessments are helpful for various stakeholders such as local authorities, private organisations and insurance companies, and individual owners to estimate the vulnerabilities of their buildings (Hill and Rossetto 2008, Vona *et al.* 2018) and make decisions on resilience intervention strategies or model reliable loss scenarios (Del Vecchio *et al.* 2020), before proceeding into expensive detail assessment which may require commissioning specialists (Wanigarathna *et al.* 2022). Simple antecedent loss assessment methodologies are used in rapid loss assessment immediately following earthquakes to assess the loss for resource allocation purposes (Stojadinovic *et al.* 2017) and avoid delays that impede repair works (Almufti and Willford 2013).

**CONTACT** Nadeeshani Wanigarathna  [Nadeeshani.wanigarathna@aru.ac.uk](mailto:Nadeeshani.wanigarathna@aru.ac.uk)  School of Engineering and the Built Environment, Anglia Ruskin University, Chelmsford, UK

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

Currently, there is a wide range of earthquake loss assessment methodologies (such as FEMA P-58 methodology, Hazus MR4 loss assessment methodology) and accompanying software (such as HAZUS-MH Earthquake Model, SELENA Seismic Loss Estimation software, PACT software for P-58; PAGER). (Ansal 2015, Wanigarathna *et al.* 2018, Kahandawa *et al.* 2021). Whilst these methodologies and tools are extremely supportive in understanding the financial requirements following a disaster, a number of previous research have highlighted weaknesses of some widely used tools by computing the differences between predicted losses based on these methodologies against observed losses for past earthquake events (see Spence *et al.* 2003 – For HAZUS-MH; Del Vecchio *et al.* 2017 – for FIMA P –58). Most of these differences are related to existing methodologies' inability to consider a large range of building and related variables within their estimates. For instance, Del Vecchio *et al.* (2017) reported that the underestimation of cost is related to the way damages are calculated for non-structural systems within buildings.

Advanced modelling techniques including Machine Learning (ML) are becoming widely used techniques for decision making in many areas of disaster management to eliminate weaknesses associated with existing methodologies. In a recent systematic review, Sun *et al.* (2020), found a total of 26 ML based applications and 17 application areas related to disaster mitigation, preparedness, response and recovery. ML applications and related data analytics offer an unprecedented opportunity to analyse existing real data sets such as insurance claims, post-disaster assessments, and similar (Kim *et al.* 2016, Sun *et al.* 2020) to produce data driven and more accurate decision support frameworks by extracting relations between available data presented (Hashemi *et al.* 2020). In this context, Machine Learning (ML) based cost models can eliminate some of the weaknesses of the existing methodologies by unpicking insights from existing data sets and incorporating a large range of variables within prediction models. The research presented here analysed an existing data set using 6 ML techniques to identify determinants of repair costs based on the repair works and building characteristics. In particular, a macro data set of reconstruction costs details for 6537 residential properties recorded from 60 towns in the Italian region of Emilia Romagna, damaged by the 2012 Earthquake were analysed to predict the damage repair cost rate, i.e. CostPerSQM, and analyse effects of individual factors on the predicted outcome. SHAP analysis was used to enhance the interpretability of the black box machine learning models, offering global and

local interpretability in visualised manner. Global interpretability offered an overview of feature (factor) importance, while local interpretability shows the effect of a single factor, or interaction effects between two factors.

The results would improve the antecedent disaster repair cost modelling process by identifying factors to consider when converting physical damage into monetary losses.

## Literature review

### *Antecedent earthquake loss assessment and weaknesses of current approaches*

Antecedent earthquake loss assessment involves computing physical damage levels by considering hazard characteristics, exposure levels and building characteristics and then converting the damage into monetary values termed as losses. Existing efforts to damage assessment for buildings can either focus on computing damage and losses at the building level or computing damage and losses at the individual component level (e.g. FEMA P-58 methodology) which could then sum up to calculate the losses at the building level. This research focuses on the building-level loss assessment, which is considered as the traditional and most widely used approach (Vona *et al.* 2018), as opposed to component-level damage and loss assessments, which involve complex procedures (Del Vecchio *et al.* 2020) and require substantial expertise based on a multidisciplinary view of structural assessment (Vona *et al.* 2018). Ansal (2015) provided a comprehensive review of existing earthquake damage and loss assessment methodologies and software solutions. Whilst, these methodologies can generate valuable information for decision makers, number of researchers have highlighted inaccuracies of the cost estimates developed based on existing methodologies. For example, Del Vecchio *et al.* (2017, 2018) evaluated FEMA P-58 methodology for L'Aquila Earthquake and found that the total predicted repair costs are significantly lower than actual observed repair costs. Bird and Bommer (2004) and Spence *et al.* (2003) presented the results of a study comparing predicted loss estimation based on the HAZUS methodology against observed damage of the 1999 Kocaeli Earthquake in Turkey. Spence *et al.* (2003) reported a significant over prediction of damage levels by 60-100% for two observed zones in Kocaeli.

The inaccuracies of loss assessments generated based on existing methodologies are related to several reasons. Firstly, generic assumptions and calculations behind these methodologies lack consideration of regional differences in building morphology and the

ability to account for the nature of individual building elements and their materials, hence resulting in inaccurate damage estimates (Baker *et al.* 2016, Cremen and Baker 2019, Del Vecchio *et al.* 2020). Secondly, there are weaknesses associated with the economic models (termed as damage cost functions or damage cost ratios) used to convert pre computed damage levels into monetary losses. Damage cost functions represent the relationship between damage levels (such as slight damage, complete damage etc.) and a property's (building or component) replacement cost. For example, the Hazus MR4 loss assessment methodology assumes 2%, 10%, 44.7%, 100% of building replacement costs as losses related to 'Slight', 'Moderate', 'Extensive', and 'Complete' damage states (respectively). Whilst these estimates may provide reasonably accurate estimates for large building stocks at a regional level, estimates for individual buildings or for smaller estates could be highly inaccurate as these functions do not consider individual building characteristics (see Wanigarathna *et al.* 2022 for more details). Finally, a number of external factors following disasters could also contribute to inaccurate estimates. For instance, past examples showed evidence of significant increases in construction cost levels in the aftermath of disaster events due to various reasons associated with surge in demand causing shortage in labour, professional and material resources (Ruddock *et al.* 2010, Wedawatta *et al.* 2018, Kahandawa *et al.* 2021, Ortiz *et al.* 2021). Other factors such as additional costs incurred due to demolition and debris disposal charges of heavily damaged buildings (De Martino *et al.* 2017); increases in repair requirements due to damages caused by earthquake-induced secondary hazards such as aftershocks, liquefaction (Kahandawa *et al.* 2021); delays in starting repair works for disaster related factors (Almufti and Willford 2013, Kahandawa *et al.* 2021); changes to rebuilding regulations following disasters (Daniell *et al.* 2017); and other factors associated with positioning, work accessibility, interactions with other buildings, and maintenance patterns (Vona *et al.* 2018). Further research is therefore required to advance the practice of antecedent earthquake loss assessments to avoid inappropriate assumptions being made by decision makers based on inaccurate estimates (Hill and Rossetto 2008, De Martino *et al.* 2017).

### **Cost modelling in construction and modelling techniques**

A model is a simplified abstract of a real system. Cost modelling in the construction sector intends to predict

acceptable output costs or prices for construction works based on an established series of input data (Seeley 1997). Construction professionals and other stakeholders utilise a range of simple to advanced cost modelling techniques to estimate costs or prices for construction works. For example, simple approaches such as the superficial area method, could be utilised to estimate the cost of a simple standard building (output) by multiplying the gross internal floor area of the building by a carefully established cost per area rate for the selected type of building. However, cost modelling becomes more complicated when the construction work is diversified, and the cost is affected by many factors (Ji *et al.* 2019) such as location, building morphology, building and individual storey heights, economic conditions, procurement arrangements, regulations, environmental factors (Ashworth & Perera 2015, Cunningham 2017, Parameswaran *et al.* 2019).

Advanced modelling techniques are extensively used in cost modelling to help to improve the accuracy of cost estimates when the cost is affected by a large number of variables (Kim *et al.* 2018, Ji *et al.* 2019, Geekiyanage and Ramachandra 2020). These include statistical methods (such as Regression Analysis and Monte Carlo simulations), advanced ML applications (such as Artificial Neural Networks – ANN, Fuzzy Neural Networks, Case-Based Reasoning (CBR), Expert systems, Decision Tree models), and Intuitive models. An example of using statistical methods could be found in the work of Parameswaran *et al.* (2019). They used AHP to determine the impact of location factors on construction costs. Geekiyanage and Ramachandra (2020) have used hedonic regression modelling to investigate the nexus between running costs and building characteristics of commercial buildings. Hashemi *et al.* (2020), in their recent systematic literature review, identified the extent to ML techniques are used in construction cost modelling. The authors concluded that the ANN and RA, respectively, are the most widely used cost modelling techniques, and the hybrid models of ANN with fuzzy logic, CBR, and Genetic Algorithm (GA), are the next most used methods. Authors reported a trend in increased use in ML applications for cost modelling over the years. They also reported that many applications are related to predicting building costs compared to other infrastructure types. Applications of ANN have been considerably popular around the beginning of the 21<sup>st</sup> century. As early as 1999, Li and Love (1999), in their highly cited paper, demonstrated the benefits of ANN and rule-based expert systems that could be used for estimating markup for construction projects. Emsley

*et al.* (2002) used ANN to analyse data from 300 building projects to develop a cost model to predict total construction costs. Emsley *et al.* (2002) also found that ANN was beneficial and reduced the prediction error significantly due to their ability to cope with nonlinearity of data compared to traditional regression analysis. Using data from 177 projects, Wanous *et al.* (2003) developed an ANN model to support bid or not to bid decision for construction contractors based on 18 selected variables. In a relatively recent example, Ahiaga-Dagbui and Smith (2014) analysed a comparatively large data set of data from 1600 completed projects. Using non-parametric bootstrapping and ensemble modelling in ANN authors developed a significantly accurate final cost forecasting model for construction projects. Applications of other advanced ML methods were limited. For example, Fuzzy logic applications have been used in cost modelling in conditions of uncertainty, and when the information and data available are only partial or incomplete, whilst neural networks have been used for learning and predicting the relationship between a number of inputs (factors affecting cost) and cost outputs when the functions are not explicit (Canesi and Marella 2017). Rafiei and Adeli (2018) used an unsupervised deep Boltzmann machine (DBM) learning approach to estimate building construction costs using a wide range of variables related to 372 multi story building projects.

Whilst advanced modelling methods are widely used for cost estimation for new build construction project decision making, their application in disaster damage and repair cost modelling is limited. Sun *et al.* (2020) reviewed applications of machine-learning applications used in disaster damage modelling. The authors identified the instances of using Linear regression (4 examples), nonlinear regression (3 examples), decision tree (1 example), and Neural networks (2 examples) in evaluating loss and repair cost. However, these early applications have limitations in their applications. Firstly, some applications have not used state-of-the-art ML modelling techniques. For example, Kim *et al.* (2016) have developed a multiple regression model to predict the relationship between insurance claim payout and 7 building and hazard related variables by analysing 4150 insurance claims related to a past Hurricane event. This analysis has only used traditional statistical analysis techniques such as Multi-collinearity analysis, ANOVA tests for the regression modelling. Secondly, some applications (Ferguson and Martin 2015, Stojadinovic *et al.* 2017) have used ML applications to predict damage levels and converting them into losses were computed using other methods.

For example, Stojadinovic *et al.* (2017) used random forest machine-learning algorithm to analyse 2010 Kraljevo earthquake data to develop a repair cost model. In this work, ML modelling was used to predict the damage level and repair cost was subsequently added based on expert opinion. In other words their focus has not necessarily been repair cost modelling. Previously, Kousky and Michel-Kerjan (2017) used regression analysis to analyse 1 million flood claims to unpick the nature and drivers of flood claims. Polese *et al.* (2015) used regression analysis and Montecarlo simulation to analyse L'Aquila earthquake data to explore the relationship between performance loss and repair costs, not necessarily to model the repair cost. Therefore, to the best of authors knowledge, so far, state-of-the-art ML techniques have not been utilised for earthquake damage repair cost prediction.

### **Comparison of traditional prediction models and machine learning models**

The difference between machine learning models and traditional prediction models lies primarily in their approach to learning and making predictions. Firstly, traditional prediction models are based on predefined rules or algorithms to make predictions and do not adapt or learn from data. Machine learning models iteratively learn patterns and relationships directly from data and improve performance over time. Machine learning models can detect complex patterns and make predictions in situations where traditional models may struggle (Xie *et al.* 2020). Secondly, traditional models often require human experts to manually identify and define features or variables that are relevant to the prediction task. Machine learning models can automatically discover and learn relevant features from raw data. They can handle high-dimensional and complex data and can often identify subtle patterns that may not be apparent to human experts (Kong *et al.* 2019). Finally, traditional models struggle in complex, high-dimensional data environments or in situations where relationships are non-linear or difficult to model (Lü *et al.* 2015). Machine learning models are capable of capturing complex, non-linear relationships in data and can adapt to changing patterns over time (Liu *et al.* 2019). Compared to traditional models, they often achieve better performance and generalisation on unseen data, especially in large-scale or high-dimensional datasets (Shirzadi *et al.* 2017, Wang *et al.* 2023).

The innovation of using machine learning models in earthquake engineering lies in its ability to provide

accurate and data-driven predictions quickly and efficiently. Machine learning applications in earthquake engineering have been categorised into four topic areas (Xie *et al.* 2020), including: (1) seismic hazard analysis, (2) system identification and damage detection, (3) seismic fragility assessment (Tocchi *et al.* 2023), and (4) structural control for earthquake mitigation (Kameshwar *et al.* 2020). The intersection of machine learning and earthquake engineering is an emerging field with growing dynamism, offering opportunities for impactful research. This study demonstrates the application of machine learning algorithms in estimating post-earthquake repair costs, based on a wide range of relevant factors, such as the extent of damage, building characteristics, geographical location, and local construction costs. As the next generation of spatiotemporal data emerges, characterised by large scalability, high dimensionality, nonlinearity and heterogeneousness, advanced machine learning models such as reinforcement learning and deep learning (Kong *et al.* 2020), are being developed. These models enable scalability and redeployment of machine learning models to handle large-scale disaster scenarios, providing timely and reliable estimates of repair costs for a wide range of buildings (Xie *et al.* 2020).

### ***Suitability of advanced ML approaches to predict disaster impacts using secondary data***

Advanced ML applications are increasingly being used to predict earthquake impacts based on secondary data sources. Whilst traditional ML methods have several limitations related to learning from the raw data, Linardos *et al.* (2022) provide a review of advanced ML and Deep Learning techniques used very recently (since 2017) in all phases of disaster management across a range of disasters. Authors claimed analysing big datasets and crowdsourcing could improve the efficiency and accuracy of laborious modelling by human. Lin *et al.* (2020) have demonstrated how ML techniques and big data sources could be used in estimating relief supplies demand for urban flood disasters. By utilising high-level data about 211 destructive earthquakes in mainland China from 2001 to 2019, Chen and Zhang (2022) demonstrated how automated machine learning can identify best-performing modelling techniques and generate a model to perform casualty rate and economic loss from earthquakes. Zhao *et al.* (2024) demonstrated how macro economic losses could be modelled using advanced machine

learning techniques of Support Vector Machines (SVM) and Extreme Gradient Boosting (XGBoost) with the intention of providing stakeholders with timely and accurate insights for risk mitigation. Iglesias *et al.* (2024) showed how ML could be used as a low-cost computing method to accurately predict disaster incidents at the municipality level using existing secondary data. Utilising secondary data, Tocci *et al.* (2023) used decision tree based ML approach to predict the usability of buildings following an earthquake event. The authors claimed their approach was able to incorporate traditionally neglected uncertainty of subjective decisions. ML based usability data were subsequently converted to usability related economic losses using non ML approaches. Overall, these latest research shows the ability of advanced ML approaches to utilise secondary data and improve disaster preparedness activities by predicting different types of impacts. The repair cost to buildings following earthquakes remains an unexplored research area and a similar approach could be taken to provide accurate antecedent loss assessment models using secondary data.

### **Materials and methods**

This research used a cost database for repairing residential buildings damaged by the Emilia Romagna earthquake in Italy. The Italian region of Emilia Romagna was hit with a 5.9Mw earthquake in May 2012. After the earthquake and damage to the buildings, separate restoration programmes were set up for different types of buildings, such as residential buildings, production buildings and infrastructures. This research analysed reconstruction costs to residential buildings obtained from the MUDE. The MUDE (Unique Digital Building Model) was a specifically deployed IT solution to manage applications for funding for the reconstruction of residential buildings damaged by the earthquake. Government funds and charitable donations were used to fund repairs to the damaged buildings. Owners of the damaged buildings made applications to request funding for their damaged buildings via MUDE. The deadline for making applications was December 2017. The research presented in this paper used an exported copy of the MUDE data on request in 2016 while the applications were still in progress. Additional data was accessed later when it was made publicly available via the Open portal (<https://openricostruzione.regione.emilia-romagna.it/ricostruzione-privata>).

**Table 1.** Example of the original dataset.

| ID    | AMOUNT-ASSIGNED | PRACTICAL-STATUS    | HOUSE NUMBER | AREA              | INSTRUCTION-DAYS | CostPerSQM |
|-------|-----------------|---------------------|--------------|-------------------|------------------|------------|
| M5784 | 75103.00        | DECRETI CONCESSIONE |              | BAGNOLO IN PIANO  | 441              | 360.12     |
| M3009 | 12332.00        | DECRETI CONCESSIONE | 25           | CAMPAGNOLA EMILIA | 43               | 62.2       |
| M4812 | 13245.00        | DECRETI CONCESSIONE | 55           | CAMPAGNOLA EMILIA | 35               | 53.57      |

### Exploratory data analysis

The first step in data analysis is exploring and summarising the main characteristics of the dataset, to identify missing data, outliers and trends. An example of the original dataset is given in Table 1. The variables in the dataset are distinguished as categorical and numerical variables. The exploratory data analysis revealed the following information:

- There are 7999 rows and 34 columns in the raw dataset.
- Each row represents a specific building that was damaged in the earthquake and required construction.
- CostPerSQM (total repair cost/total floor area) was added to the dataset as a new variable by the authors as the dependent variable that aim to understand and predict.
- CONSTRUCTION REQUEST-YEAR has been added to the original dataset by the authors to reduce the number of values associated with the variable CONSTRUCTION REQUEST-DATE to a create a manageable independent variable.
- 35 columns representing independent variables, included 19 categorical and 16 numerical variables.
- Seven variables contained missing data, including Intervention Description, Construction State, House Number, Introduction Days, Intervention Type, Concession Date, and Intervention Cost

The original data was in Italian language. They were translated into English. Short names were used for the column names/variable names as handling long names could cause errors during modelling. The Table 2 explains the meanings of these variables.

Since the original dataset contains missing data and some features may carry repeated information, data pre-processing was performed to improve the quality of the data before feeding them to the ML models. Data pre-processing removed 12 variables that do not add value to the dependent variable CostPerSQM. For example, ID and CUP numbers are high cardinality features under which each building is assigned a unique ID number. Some variables contain similar or repeated information, for example, INTERVENTION-DESCRIPTION and categorised INTERVENTION TYPE. INTERVENTION TYPE was kept

while INTERVENTION-DESCRIPTION was dropped. Unique building attributes, such as ADDRESS, HOUSE NUMBER, CONCESSION-DATE, ADVANCE-PERCENTAGE, and CONTRIBUTION-REQUEST-DATE, were also removed, as these features are also high cardinality features, making it difficult for ML models to identify patterns or causal relationships between features and dependent variable. This paper intended to explore the causal relationships between building related features and the costs required or assigned to repair or restore or reconstruct the damaged buildings. The costs data include AMOUNT ASSIGNED, AMOUNT PAID, INTERVENTION COST, and CostPerSQM. Given that the reconstruction budget assigned, and the costs incurred vary with the floor area, the damage level, and many other features of the damaged buildings, the average cost per floor area, i.e. CostPerSQM, was chosen as the target variable, and examine how various building features impact this metric.

After eliminating the 13 features a total of 22 features were kept, including 8 categorical features and 14 numerical features (see Table 3). The rows that have missing values were dropped from the analysis, and the raw dataset was decreased from 7999 rows to 6025 rows.

### Characteristics of the target variable

The univariate distribution and box plot are depicted in Figure 1, to illustrate the density of the underlying distribution and the quartile groups of the dependent variable CostPerSQM. The average CostPerSQM of a building reconstruction is 938.83. According to Figure 1, the distribution is highly skewed and there are many outliers outside 1.5 times the interquartile range above the upper quartile ( $Q3 + 1.5 \cdot IQR$ ).

### Characteristics of categorical variables

Bar plots were depicted (see Appendixes) to help us understand the labels frequency for each single categorical variable. Most of buildings' PRACTICAL-STATUS is Decreti Concessione (Concession by Decree), and CONSTRUCTION-STATE is Cantiere Completato (Construction Complete). The variable AREA (the geographical area where the building is located) has a high number of labels. Mirandona area appears to

**Table 2.** Details of the dataset translated into the English language.

|    | Variable title            | Description                                                                                                                                                                                                                                                                                                                                                                                                           |
|----|---------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1  | ID                        | These are unique identification/reference numbers given to individual cases (applications) by the system.                                                                                                                                                                                                                                                                                                             |
| 2  | CUP                       |                                                                                                                                                                                                                                                                                                                                                                                                                       |
| 3  | INTERVENTION-DESCRIPTION  | A summary of the nature of damage and repair work needed to reinstate the building.                                                                                                                                                                                                                                                                                                                                   |
| 4  | AMOUNT-ASSIGNED           | The amount approved by the local authority. This will be the total amount that will be paid to the applicant in intervals based on the progression.                                                                                                                                                                                                                                                                   |
| 5  | AMOUNT-PAID               | Within the original database, this value is regularly updated based on stage payments made to the claimant. At the date of downloading the dataset, this is a snap value of the total amount paid towards an applicant.                                                                                                                                                                                               |
| 6  | ADDRESS                   | Address of the damaged building                                                                                                                                                                                                                                                                                                                                                                                       |
| 7  | HOUSE NUMBER              | House number of the damaged building                                                                                                                                                                                                                                                                                                                                                                                  |
| 8  | BUILDER                   | Name of the Builder undertaking the repair and reconstruction work.                                                                                                                                                                                                                                                                                                                                                   |
| 9  | BUILDER-ID                | The registered ID of the Builder.                                                                                                                                                                                                                                                                                                                                                                                     |
| 10 | DESIGNER                  | Name of the designer undertaking the design and supervision of repair and reconstruction work.                                                                                                                                                                                                                                                                                                                        |
| 11 | DESIGNER-ID               | The registered ID of the Designer.                                                                                                                                                                                                                                                                                                                                                                                    |
| 12 | CONCESSION-DATE           | Date of application for funding                                                                                                                                                                                                                                                                                                                                                                                       |
| 13 | ADVANCE-PERCENTAGE        | Initial payment % made to the applicant to start the repair work.                                                                                                                                                                                                                                                                                                                                                     |
| 14 | CONTRIBUTION-REQUEST-DATE | Date of the application for funding.                                                                                                                                                                                                                                                                                                                                                                                  |
| 15 | CONSTRUCTION REQUEST-YEAR | This is a category created by the authors to identify the year of the claim based on the information on the date of the claim/contribution request date from variable 13.                                                                                                                                                                                                                                             |
| 16 | INTERVENTION-COST         | This was the original cost of an intervention applied for funding before they were assessed and agreed on the assigned amount as in variable 3 above.                                                                                                                                                                                                                                                                 |
| 17 | PRACTICAL-STATUS          | Denotes the financing method for the repair works. Associated with two labels: Request In Preliminary Investigation; paid via the Concession decree.                                                                                                                                                                                                                                                                  |
| 18 | CONSTRUCTION-STATE        | The status of construction progress. Associated with two labels: Construction Complete, Construction in progress                                                                                                                                                                                                                                                                                                      |
| 19 | AREA                      | The geographical area where the building is located. The area is categorised based on the Commune categorisation for administration in Italy.                                                                                                                                                                                                                                                                         |
| 20 | INTERVENTION-TYPE         | This is a variable created by the authors to categorise the description of the work into a range of work categories: Repair and restoration, Repair and restoration with Seismic improvements, Repair, restoration and partial demolition and rebuilding, Demolition and rebuilding, Demolition and rebuilding with seismic improvement, purchase a new property.                                                     |
| 21 | DAMAGE-LEVEL              | A category assigned by the building inspectors following the earthquake as to the damage level. B/C minor damage; E0- Light Damage; E1/E2/E3 – Heavy damage                                                                                                                                                                                                                                                           |
| 22 | OPERATIONAL-LEVEL         | A category assigned by the building inspectors following the earthquake as to the operational level. Operational levels B-C and E0 are used for buildings with low damage and not needing immediate repair. Operational levels E1 and E2 indicate heavy building damage and need for repair before being housed. Operational level E3 is used for buildings so heavily damaged and needing demolition and rebuilding. |
| 23 | INSTRUCTION-DAYS          | Number of days since the instruction to go ahead with the work.                                                                                                                                                                                                                                                                                                                                                       |
| 24 | CONSTRUCTION-DAYS         | Number of days since the start of the construction or repair work.                                                                                                                                                                                                                                                                                                                                                    |
| 25 | NO-OF-HABITANTS           | Number of people living in the building.                                                                                                                                                                                                                                                                                                                                                                              |
| 26 | HOUSING-UNITS             | Number of housing units/flats within the building                                                                                                                                                                                                                                                                                                                                                                     |
| 27 | NON-PRINCIPAL(UNITS),     | Number of units used for purposes other than housing.                                                                                                                                                                                                                                                                                                                                                                 |
| 29 | PRODUCTION (UNITS)        | Number of units within the building used for production purposes                                                                                                                                                                                                                                                                                                                                                      |
| 30 | TRADE (UNITS)             | Number of units within the building used for Trade purposes                                                                                                                                                                                                                                                                                                                                                           |
| 31 | OFFICE (UNITS),           | Number of units within the building used as offices                                                                                                                                                                                                                                                                                                                                                                   |
| 32 | STORAGE(UNITS),           | Number of units within the building used for storage purposes                                                                                                                                                                                                                                                                                                                                                         |
| 33 | HOUSING_AREA              | Total floor area related to housing units in m2                                                                                                                                                                                                                                                                                                                                                                       |
| 34 | NON-HOUSING-AREA          | Total floor area related to non-housing units measured in m2. This is a variable created by adding floor areas related to non-housing units recorded in 5 consecutive columns.                                                                                                                                                                                                                                        |
| 35 | TOTAL-FLOOR-AREA          | Total floor area of the building (m2)                                                                                                                                                                                                                                                                                                                                                                                 |
| 36 | CostPerSQM                | This is the computed target variable. The total cost of the work/Total floor area (the value of the variable 3/ the value of the variable 32)                                                                                                                                                                                                                                                                         |

have the highest density of damaged buildings in the region. Most buildings require Repair and restoration, or repair and restoration with seismic improvement, or full/partial demolition and rebuilding. Most buildings are paid by Dretta and have a damage level of B/C.

### Characteristics of numerical variables

The density distributions of numerical variables are presented in Figure 2. Except for BUILDER-ID, and DESIGNER-ID, the other numerical variables have right-skewed distributions. Data skewness degrades statistics

models' (particularly distance-based models) ability to describe typical cases, as skewed regions may act as outliers and be handled as rare cases. Therefore, skewed numerical data tends to be transformed to approximate to a normal distribution, to ensure the prediction ability and accuracy of distance-based regression models, such as Linear Regression and Polynomial Regression. Since tree-based models are rule-based algorithms that do not make decisions based on distance, they are more robust enough to handle outliers (Lodhia *et al.* 2017). Hence decision tree ML models do not require data transformation to reduce skewness.

**Table 3.** Data pre-processing and feature selection.

| Dataset handling | Removed features                                                                                                                                                                                           | Kept features                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Pre-processing   | ID, CUP, INTERVENTION-DESCRIPTION, AMOUNT-ASSIGNED, AMOUNT-PAID, ADDRESS, HOUSE NUMBER, BUILDER, DESIGNER, CONCESSION-DATE, ADVANCE-PERCENTAGE, CONTRIBUTION-REQUEST-DATE, INTERVENTION-COST (13 features) | Categorical features: PRACTICAL-STATUS, CONSTRUCTION-STATE, AREA, INTERVENTION-TYPE, PAYER, OPERATIONAL-LEVEL, DAMAGE-LEVEL, CONSTRUCTION REQUEST-YEAR (8 features)<br>Numerical features: INSTRUCTION-DAYS, CONSTRUCTION-DAYS, BUILDER-ID, DESIGNER-ID, NO-OF-HABITANTS, HOUSING-UNITS, NON-PRINCIPAL(UNITS), PRODUCTION (UNITS), TRADE (UNITS), OFFICE (UNITS), STORAGE(UNITS), HOUSINGAREA, NON-HOUSING-AREA, TOTAL-FLOOR-AREA (14 features) |

**Table 4.** Feature selection for the final model.

| Dataset handling                      | Removed features                                                                                                                                            | Kept features                                                                                                                                                                                                                                                                                          |
|---------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Feature selection<br>During modelling | PRACTICAL-STATUS, INTERVENTION-TYPE, PAYER, BUILDER-ID, DESIGNER-ID, NO-OF-HABITANTS, HOUSING-UNITS, TRADE UNITS, STORAGE UNITS, HOUSING AREA (10 features) | Categorical features: CONSTRUCTION-STATE, OPERATIONAL-LEVEL, DAMAGE-LEVEL, (GEOGRAPHICAL)AREA, CONSTRUCTION REQUEST-YEAR (5 features)<br>Numerical features: INSTRUCTION-DAYS, CONSTRUCTION-DAYS, NON-PRINCIPAL UNITS, PRODUCTION UNITS, OFFICE UNITS, NON-HOUSING-AREA, TOTAL-FLOOR-AREA (7 features) |

### Feature selection for ML models

Having explored the characteristics of the dataset, the next step was to select a set of relevant features to build ML models and predict the target variable CostPerSQM. Feature selection is important and necessary in such a way that it reduces model complexity, overfitting, and computing power to train ML models. Removing noisy features and keeping only relevant features also enhances model interpretability.

Given that predicting CostPerSQM is a regression problem, It was logical to examine the relationships between features and the target variable using two linear regression models. The first regression model is F\_regression which runs univariate linear regression tests between each feature (regressor) and the target variable (CostPerSQM) and returns ANOVA F-statistic and p-values. Features (regressors) were ranked in the order of p-values, and the highest-ranked features were selected. The second model is RIDGE regularisation (see Hastie 2020) which solves a regression model where the loss function is the linear least square function, and regularisation is given by the L2-norm. This model is useful in mitigating the multicollinearity issues in linear regression.

A ranking of features determined by the above two models is depicted in Figure 3. The blue features are the features selected by both models, and the orange ones are chosen by F\_regression only, whereas the green features were selected by RIDGE regularisation.

Last 10 features in the graph above (PRACTICAL-STATUS, INTERVENTION-TYPE, PAYER, BUILDER-ID, DESIGNER-ID, NO-OF-HABITANTS, HOUSING-UNITS, TRADE, STORAGE, HOUSINGAREA) not selected by any model was therefore removed from ML.

### Machine learning models

Compared with traditional time series or regression-based prediction models, machine learning models have the ability to provide accurate and data driven predictions efficiently, even when historical data are severely skewed and do not follow normal distribution. Depending on the learning approach and the nature of the task, machine learning models can be broadly categorised into two main types, including supervised learning models and unsupervised learning models (Kong *et al.* 2019). Supervised learning models learn from labelled data, while unsupervised model learns from unlabelled data. Based on the characteristics of the data and the outcome variable to be predicted, supervised learning can be further classified into regression models and classification models. Regression models predict continuous numerical values, while classification models predict discrete categorical classes. In this study, the predicted variable CostPerSQM is a continuous numerical value and historical labelled data are available, hence, supervised regression models were developed.

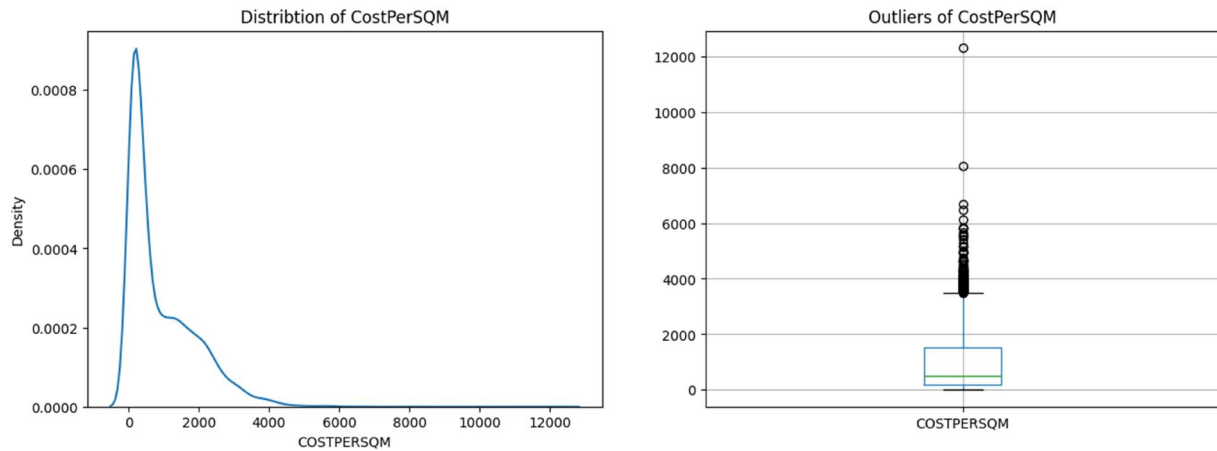


Figure 1. Distribution and box plot of the dependent variable CostPerSQM.

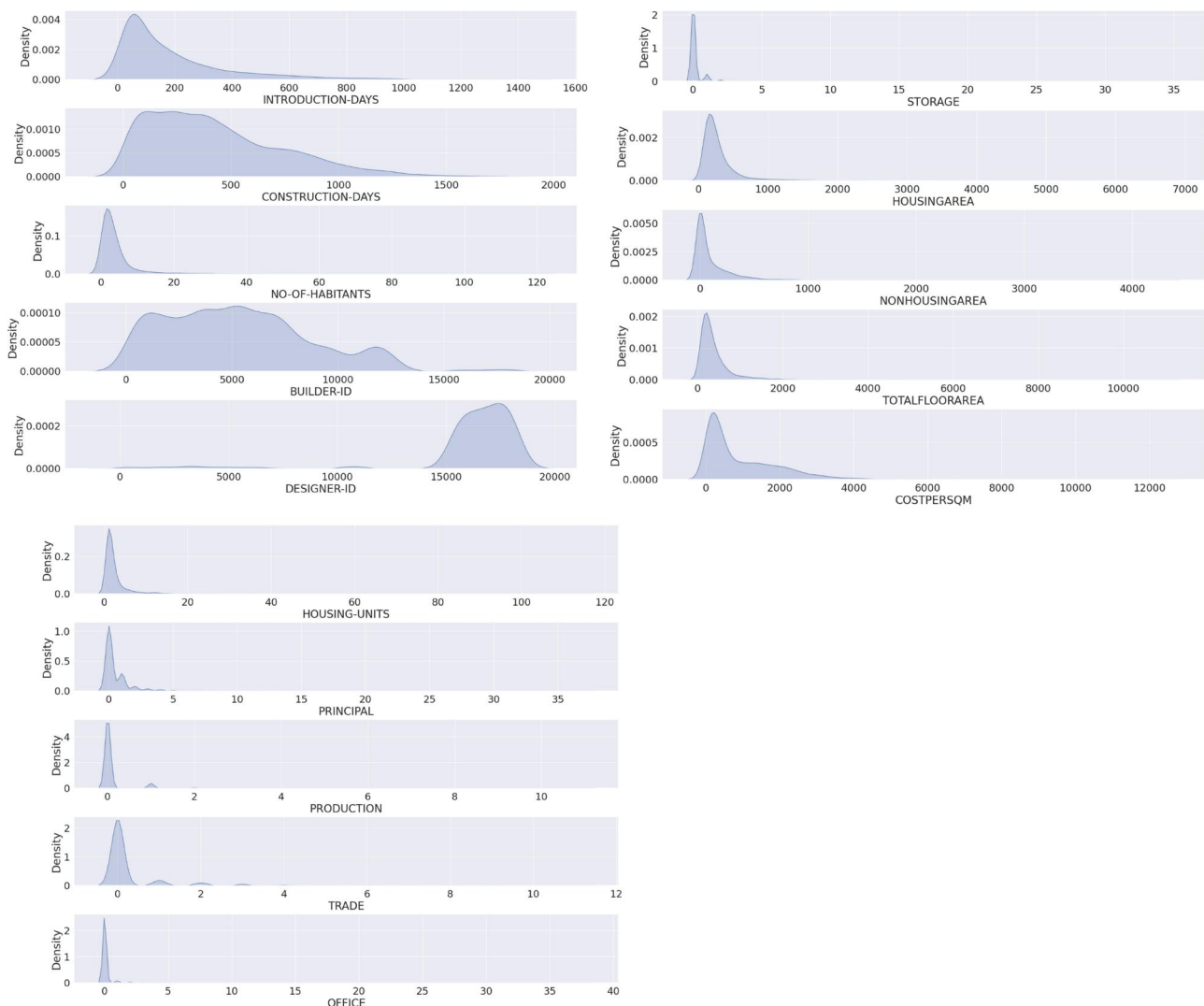


Figure 2. Kernel density estimations of the numerical variables.

There are several types of supervised regression models, each with its own characteristics and suitability for different types of data and tasks. Some common types of supervised regression models include

Linear Regression, Polynomial Regression, Support Vector Regression (SVR), Decision Tree Regression, Random Forest Regression and Gradient Boosting Regression. Decision Tree based models consistently

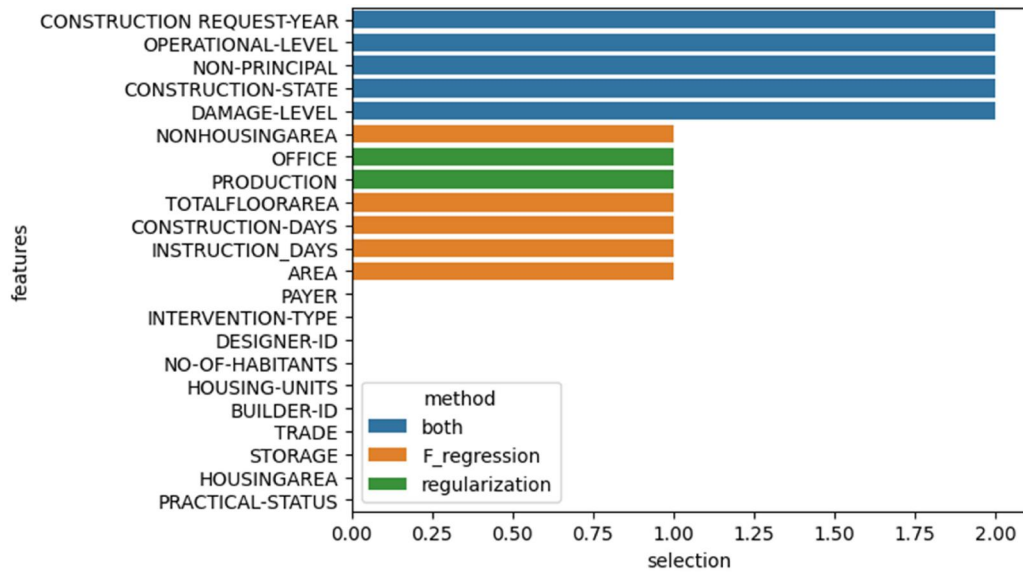


**Figure 3.** Feature selection by F\_regression and RIDGE regularization.

outperform traditional regression models on tabular-style datasets, where features are individually meaningful and lack strong multiscale temporal or spatial structures (Chen and Guestrin 2016, Lundberg *et al.* 2019). Given that the dataset in this study comprises tabular-style data, tree-based models were considered appropriate for making predictions. Random Forest, a bagging ensemble decision tree model introduced by Breiman (1996), targets variance reduction. In contrast, Gradient Boosting Regression, a boosting ensemble decision tree model proposed by Friedman (2001), sequentially trains models using residuals, mitigating both variance and bias.

Following the feature selection and feature engineering, machine learning based regression models were developed following these steps:

1. Features chosen by F\_regression and RIDGE, as shown in Figure 3, were used for the design, training, cross-validation and test of different machine learning-based regression algorithm models.
2. The dataset with selected features were split into training and test data (75:25 ratio of the original dataset). A 5-fold cross validation was performed in combination with grid search to improve and validate the performance of each machine learning model by tuning the model hyperparameters. Machine learning models with the tuned hyperparameters were used to predict the target variable CostPerSQM.
3. In total six machine learning models were developed to predict CostPerSQM (Alpaydin 2020, Trincherro and Canavero 2021), including three distance-based regression models (Linear Regression,

Polynomial Regression, Support Vector Regression), a Decision Tree Regression model and two Ensemble regression models (Random Forest Regression and Gradient Boosting Regression).

- Linear Regression: it assumes a linear relationship between the independent variable  $x$  and the dependent variable  $y$ .
- Polynomial Regression: the relationship between the independent variable  $x$  and the dependent variable  $y$  is modelled as an  $n^{\text{th}}$  degree polynomial in  $x$ .
- Support Vector Regression: it extends the Support Vector Machines, a nonparametric technique utilising kernel functions. Its objective is to fit as many instances as possible within a margin of tolerance, while constraining margin violations.
- Decision Tree Regression: it models complex nonlinear relationship between the independent variables  $x$  and the target variable  $y$  in a tree-like structure where each internal node represents a decision based on value of  $x$  and each leaf node represented the predicted value for  $y$ . Decision tree regression recursively partitions the data into subsets based on the values of the independent variable  $x$ , aiming to minimise variance or another specified criterion, and makes predictions by traversing the tree from the root to the appropriate leaf node.
- Random Forest Regression: it is an ensemble technique that builds multiple decision trees using subsets of the original dataset with replacement. The predictions of multiple

decision trees are averaged to reduce overfitting and improve generalisation.

- Gradient Boosting Regression: it works by building a series of weak decision tree models sequentially where each model tries to correct the errors of the previous model. Gradient boosting regression minimizes a loss function by iteratively fitting new models to the residuals of the previous predictions, gradually improving the overall predictive accuracy.
4. Transformation was applied to reduce the skewness and normalise the skewed numerical data before inputting them into the three distance-based regression models. Depending on values of numerical data, Square root, Box Cox Transformation, Log Transformation or Reciprocal Transformation could be applied.
  5. The performances of the machine learning models were evaluated using appropriate metrics, to gauge how well the models predict the target variable CostPerSQM. Since CostPerSQM is a single variable predicted using regression models, two commonly used evaluation metrics are:
    - i. Coefficient of Determination ( $R^2$ ): it measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1)$$

where  $y_i$  is the observed value and  $\hat{y}_i$  is predicted value;

$\bar{y}$  is the mean of the observed data;

$n$  is the number of observed (actual) values.

- ii. Root Mean Squared Error (RMSE): it measures the average difference between predicted values and the actual values. It gives a sense of how wrong the predictions are on average.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

### Machine learning interpretability

Sophisticated ML algorithms usually can produce accurate predictions, but their “black box” nature hinders the adoption of the ML model (Hassija *et al.* 2023). To address the limitations and challenges of interpreting traditional machine models, a wide variety of advanced model interpretation techniques were developed by researchers. Model-agnostic (Ribeiro *et al.* 2016) interpretation technique has been widely

used to explain feature importance in ML models. Feature importance indicates the amount of ‘importance’ of each feature using a single value. The single value represents the influence on the loss function (such as Gini impurity, Log Loss) by the presence of the feature. In ML models, there are different ways of ranking feature importance (Bommert *et al.* 2020, Effrosynidis and Arampatzis 2021), including:

- i. Decision tree-based feature importance can be measured as:
  - a. Coverage - the number of times a feature is used to split the data across all trees, weighted by the probability of reaching that node. The node probability can be calculated as the number of samples that reach the node, divided by the total number of samples.
  - b. Gain - the average training loss reduction (such as Gini impurity, Logloss, variance) or information gained when using a feature for splitting the data across the tree.
- ii. Permutation importance: Permutation importance is defined to be the decrease in a model score when a single feature value is randomly permuted (Breiman 2001). This technique breaks the relationship between the feature and the target; it reflects how important a feature is for a particular model rather than the intrinsic predictive value of a feature. This technique a model-agnostic approach and can be calculated many times with different permutations of the feature.
- iii. SHAP (SHapley Additive exPlanations) values interpret machine learning models based on aggregations of Shapley values (Lundberg and Lee 2017). Shapley value is the average marginal contribution of a feature value from coalitional game theory, and it tells us how to fairly distribute the model prediction among the features (or determinant factors) (Shapley 1953). SHAP is also a model-agnostic method. The core idea behind SHAP is to calculate contributions made by each feature a model’s output, namely the predicted CostPerSQM in this research, using cooperative game theory. The contributions made by each feature is computed as the SHAP value.

SHAP proved to be a good feature attribution method that is consistent and accurate (Lundberg and Lee 2017). Consistency means the attributed importance of a feature should not decrease/increase as a result of changing another feature that relies more/less on that feature. Accuracy requires the sum of all

the feature importance adds up to the total importance of the model (the importance of the model can be measured as the probabilities belonging to classes, or Log odds for classification problem, or  $R^2$  for regression problem).

## Results

### *Prediction performances of ML-based regression models*

To compare the prediction performances of the machine learning models, three sets of experiments were conducted by feeding different sets of features into the machine learning models.

1. In the first experiment, five features that were recommended by both feature selection methods were selected. They were CONSTRUCTION REQUEST-YEAR, OPERATIONAL-LEVEL, NON-PRINCIAL UNITS, CONSTRUCTION-STATE, AND DAMAGE-LEVEL. Since all the features are categorical variables, data transformation was not required.
2. In the second experiment, three more numerical features, namely, NON-HOUSING-AREA, OFFICE UNITS and PRODUCTION UNITS were added. These features contained zero values. Hence, Box Cox Transformation, Log Transformation or Reciprocal Transformation could not be used. Square root was applied to reduce skewness in these variables before they were fed into the three distance-based models.
3. In the third experiment, all the features selected by F\_regression and RIDGE were input to the ML models, which means the additional four features were added to the models built in the second experiment, including TOTAL-FLOOR-AREA, CONSTRUCTION-DAYS, INSTRUCTION-DAYS and (GEOGRAPHICAL) AREA. Square root transformation was also applied to these variables before they were input to the distance-based regression models to reduce the skewness.

The results show that performances of distance-based regression models worsen after square root transformation was applied, although the skewness of individual numerical variables is reduced. Therefore, as the next step the distance-based regression models was re-run by removing the transformation of the numerical variables and presented the results in Table 5. Table 5 compares the performances achieved by the six regression models in the three experiments, using the metrics of Root Mean Squared Error (RMSE)

and R squared Score ( $R^2$ ). It is apparent that additional input features in Experiments 2 and 3 improve the prediction performances of all the regression ML models except for the Polynomial Regression model. Therefore, the Polynomial Regression model is deemed unsuitable for the prediction. Among the other five regression models, Gradient Boosting Regression outperforms other models in terms of both RMSE and  $R^2$ . This is the case in every experiment. The Gradient Boosting Regression model explains 76% of the variance of the target variable CostPerSQM, and the RMSE value is 465.38.

### *Gradient boosting regression modelling*

Gradient Boosting Regression Modelling, was used to plot the predicted target variable CostPerSQM against the actual value in Figure 4. It is observed that most of the predicted values are close to the regressed diagonal line, having a relatively good fit. The biggest residual error on the test data, i.e. residual = actual test data – predicted test data, was over 3,000. Most of the residual errors lie between –1000 and 1000.

### *Interpretability using SHAP summary plots*

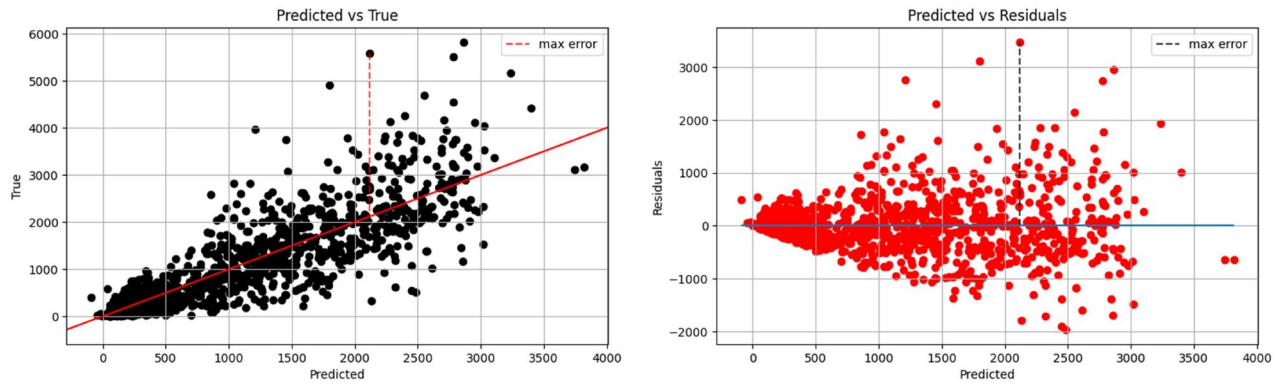
The following section, explored the importance of features in making the prediction and quantified their contributions to the prediction using the SHAP (SHapley Additive exPlanations) analysis. SHAP analysis reduces the limitations associated with black boxes of ML models by revealing the main independent variables that affect the dependent variable of a model (Grootendorst 2019). SHAP analysis produces a SHAP summary plot and SHAP dependence plots.

### *SHAP summary plot results*

SHAP summary plots combine feature importance with feature effects. It displays the magnitude, prevalence and direction of a feature's effect. The SHAP summary plot in Figure 5 sorts features for the target variable CostPerSQM by the sum of SHAP magnitudes (absolute of SHAP values) over all samples. The position on the y-axis is determined by the feature importance and the dot's position on the X-axis shows that impact that feature has on the model's prediction. Each point on a SHAP plot is one SHAP value for a prediction and feature. The colour of the dot represents the value of that feature from low to high. Red colour indicates a higher value of a feature, whilst Blue colour indicates a lower value of a feature.

**Table 5.** Evaluation of ML prediction performances.

| ML models                   | 1 <sup>st</sup> experiment |                | 2 <sup>nd</sup> experiment |                | 3 <sup>rd</sup> experiment |                |
|-----------------------------|----------------------------|----------------|----------------------------|----------------|----------------------------|----------------|
|                             | RMSE                       | R <sup>2</sup> | RMSE                       | R <sup>2</sup> | RMSE                       | R <sup>2</sup> |
| Linear regression           | 581.00                     | 0.64           | 571.59                     | 0.65           | 569.44                     | 0.66           |
| Polynomial regression       | 565.29 (n = 4)             | 0.66           | 769.69 (n = 2)             | 0.37           | 871.84 (n = 2)             | 0.20           |
| Support vector regression   | 589.50                     | 0.63           | 578.57                     | 0.65           | 577.05                     | 0.65           |
| Decision tree regression    | 551.40                     | 0.68           | 578.45                     | 0.65           | 578.45                     | 0.65           |
| Random forest regression    | 548.46                     | 0.68           | 525.36                     | 0.71           | 493.06                     | 0.74           |
| Gradient booting regression | 538.22                     | 0.69           | 504.62                     | 0.73           | 475.58                     | 0.76           |

**Figure 4.** Predict value vs True value and Predicted value vs Residuals.

In SHAP summary plots, the high value comes from the red colour in the vertical Y-axis, and the high probability is shown as the SHAP value on the horizontal X-axis. A few observations can be made based on the plot in Figure 5 related to feature importance, impact and correlation.

#### Feature importance ranking

OPERATIONAL-LEVEL, DAMAGE-LEVEL and NON-HOUSING-AREA are the top 3 important features when predicting CostPerSQM as they scored the highest feature values. These results partially tally with previous research where Damage levels to a building and the resulting operational level are used as the main predictors of resultant repair costs. All widely used simplified damage cost functions (such as EMS 98 damage scales applied to European buildings, and the Hazus MR4 loss assessment methodology) use damage level to convert physical damage to direct loss construed as the total repair costs (see Wanigarathna *et al.* 2022 for further details). However, the non-housing area (areas within a residential building that are used for purposes other than housing, e.g. small shops) has never been identified as one of the most important features related to damage repair cost prediction.

#### Impact and correlation

The plot in Figure 5 also reveals the direction of the effects and hence provides useful insights related to the impact and correlation of individual features. For

example, a lower value of OPERATIONAL-LEVEL is associated with a lower SHAP value, which implies certain OPERATIONAL-LEVELs can lower the predicted value of CostPerSQM. Higher DAMAGE-LEVELs or lower TOTAL-FLOOR-AREA are associated with higher CostPerSQM. Larger values of NON\_HOUSING\_AREA also decrease the CostPerSQM.

#### SHAP dependence plots results

A SHAP dependence plot is a scatter plot that shows the effect of a single feature on the predictions made by the model. The plots in Figure 6 show the SHAP feature dependence for each feature on CostPerSQM. These plots automatically include another feature that the chosen feature interacts most with and show the interaction effects.

These results confirm our understanding of certain features and their impact on building repair cost rates. Firstly, a clear interaction effect with OPERATIONAL-LEVEL is visible, where higher values of OPERATIONAL-LEVEL further increase the prediction of CostPerSQM 6(b). This observation is in line with our intuition, since higher OPERATIONAL-LEVEL indicates more severe damages, hence requiring higher repairing costs. This could be explained by the economies of scale and shared non-work related preliminaries costs (Ashworth and Perera 2015). Secondly, the plot in Figure 6(j) shows there is an approximately linear and positive trend between CONSTRUCTION-DAYS and the target

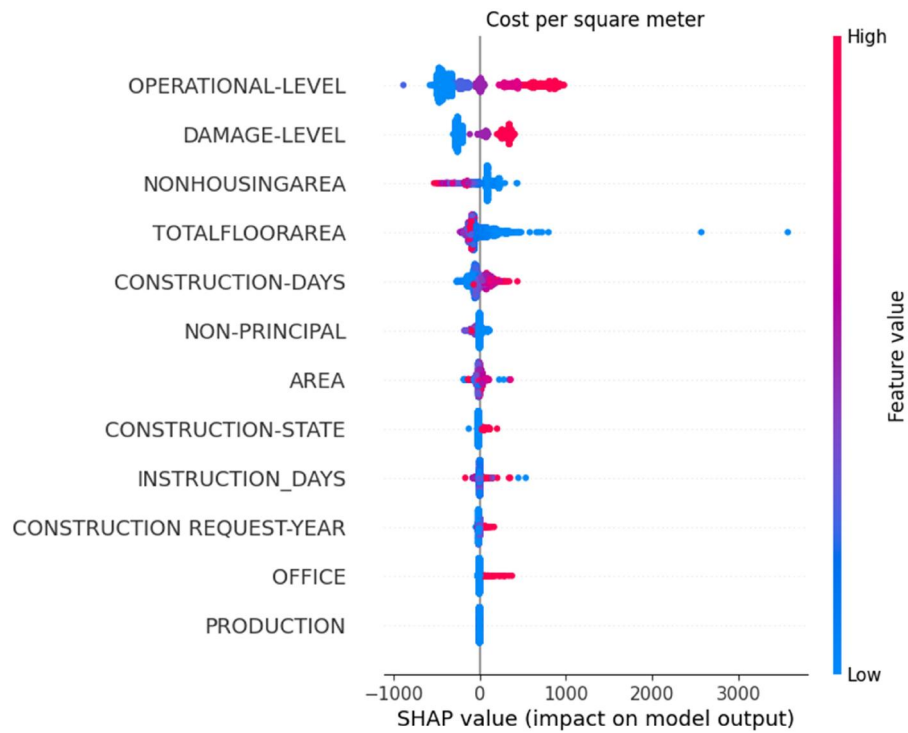


**Figure 5.** SHAP summary plot of the Gradient Boosting Regression model.

variable CostPerSQM, interacting with CONSTRUCTION-STATE. This suggests that the CostPerSQM increases with the number of CONSTRUCTION DAYS. This could be explained by the higher construction durations associated with complex repair projects requiring specialist repair works and the resultant increase in non-work related preliminaries as the duration increases. Results in plot 6(c) also support this explanation partially. CONSTRUCTION STATE of 0 (Construction Completed) decreases the prediction of the target variable when CONSTRUCTION-DAYS are less than 500 days, but it increases the prediction when CONSTRUCTION-DAYS are more than 500 days.

SHAP dependence plots also provide useful insights related to the impact of previously unsearched features on building damage repair cost rates. The plot in Figure 6(a) shows the prediction of CostPerSQM increases with the CONSTRUCTION REQUEST YEAR. This was not identified in the previous literature. This could be associated with the increased timing required to assess and design large and complex repair jobs, causing funding applications for such jobs to be late. Plot 6(k) showing the SHAP interaction value of INSTRUCTION-DAYS versus CONSTRUCTION-DAYS also supports this explanation. It shows that the impact of INSTRUCTION -DAYS on the prediction of the target variable remains almost the same when INSTRUCTION -DAYS are less than 600 days.

CONSTRUCTION DAYS is the feature that CONSTRUCTION REQUEST YEAR interacts most with. The impacts of CONSTRUCTION REQUEST YEAR change with the value of CONSTRUCTION DAYS, but without following an obvious trend.

Figure 6(f) shows a negative relationship between NON-HOUSING-AREA and the target variable CostPerSQM, indicating that higher values of NON-HOUSING-AREA decrease the prediction of CostPerSQM. This observation, not previously identified in the literature, is also evident in the SHAP summary plot in Figure 5. It could be associated with non-typical elements within other units, such as offices or production units requiring non-standard repairs. An interaction effect with OPERATIONAL-LEVEL is evident, where higher values of OPERATIONAL-LEVEL reduce the repair cost CostPerSQM by greater values if the non-housing area exceeds 100 square meters. This could be explained as: in cases of more severe damage, a higher OPERATIONAL LEVEL was assigned by building inspectors for repair, and it tends to incur significantly higher repair costs compared to those assigned a lower OPERATIONAL LEVEL. By leveraging economies of scale achieved in larger non-housing areas, where project bundling is implemented and non-work related preliminary costs are shared (Ashworth and Perera 2015), a greater reduction in CostPerSQM (absolute value of reduction) can be achieved when compared to buildings assigned a lower OPERATIONAL LEVEL. For example, when OPERATIONAL\_LEVEL = 0,

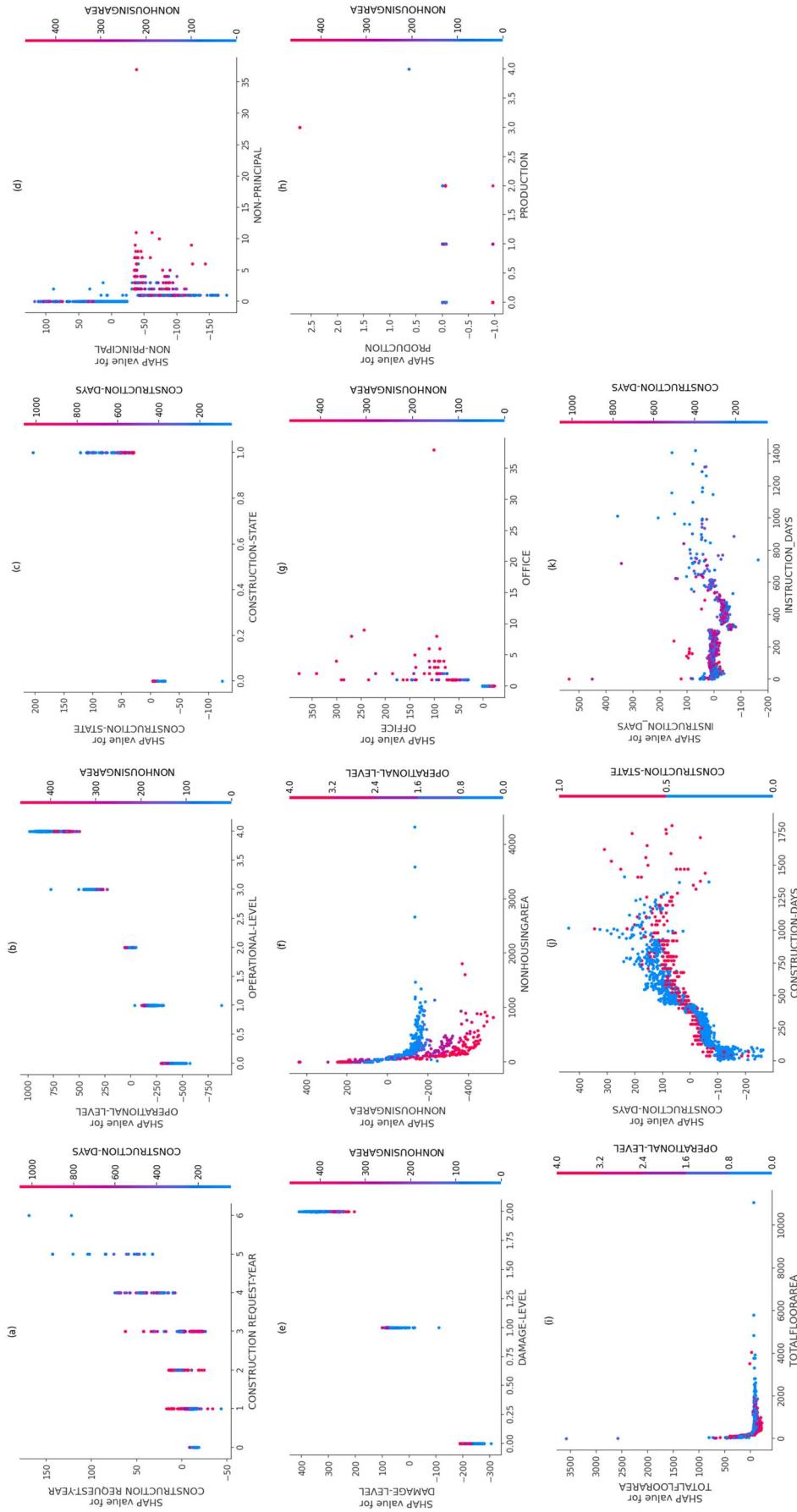


Figure 6. SHAP dependence plots for all the input features (with interaction visualisation).

increasing NONHOUSINGAREA from 977 to 1468 results a reduction of CostPerSQM from 135 to 58, yielding an absolute reduction of 77 sqm. Conversely, when OPERATIONAL\_LEVEL = 4, increasing NONHOUSINGAREA from 916 to 1584 leads to a reduction in CostPerSQM from 1587 to 1164, resulting an absolute reduction of 423 sqm.

## Discussion

The results of this research extend the existing knowledge of construction cost modelling in two ways. Firstly, it broadens the knowledge of antecedent disaster repair cost-estimating practices based on simplified damage-cost functions by establishing other key variables that could be considered during the cost-estimating process. Secondly, these results demonstrate how machine-learning techniques could be used to analyse micro cost data to develop cost models. This section will discuss these results further to provide insights into their potential application and will compare these results with previous research findings.

In contrast to the previous literature, our research findings encourage ML applications in construction cost modelling by exploring how state-of-the-art ML models that could eliminate limitations identified by previous researchers. Our approach has made advancements in three respects. Firstly, most of the previously used ML applications in cost modelling such as ANN, and FUZZY logic models, had a key limitation associated with their “black box” nature of learning which undermined the trust and limited the use in practice (Petch *et al.* 2022), or their limited learning abilities when data is incomplete (Guo *et al.* 2020). These limitations hindered the adoption of early AI-based cost models. Our research explored the suitability of six ML-based regression analysis methods to predict damage repair cost rates for residential buildings damaged by earthquake disasters. Three distance-based regression models (Linear Regression, Polynomial Regression, Support Vector Regression), a Decision Tree Regression model and two Ensemble regression models (Random Forest Regression and Gradient Boosting Regression) were tested within the research. Results confirmed that except for the Polynomial Regression model, all other modelling techniques were suitable for predicting the damage repair cost rates. Gradient Boosting Regression model outperformed other models. Gradient Boosting Regression is an ensemble regression model which combines weak learner models and predicts more accurate ensemble responses for new data by aggregating predictions from its weak learners (Alhamid 2022). The results of the Gradient

Boosting Regression model in predicting the damage repair cost rates is presented in this paper. Supplementary SHAP analysis plots eliminate the black box nature of previous ML applications by revealing the main independent variables that affect the dependent variable of a model, the direction of the impact and correlations between feature variables. Therefore, our model is a significant advancement compared to the previous AI-based models developed for construction cost modelling. Secondly, during traditional statistical analyses, skewed numerical data tends to be transformed to approximate a normal distribution to ensure the prediction ability and accuracy of distance-based regression models, such as Linear Regression and Polynomial Regression. Our research, tested the impact of the transformation of data to eliminate the skewness. The results showed that performances of distance-based regression models worsen after square root transformation is applied. The gradient Boost Regression model developed in this research uses rule-based algorithms that do not make decisions based on distance, hence, they are more robust enough to handle outliers. Thirdly, ML-based data analysis had other generic advantages compared to traditional statistical analysis. These include the speed of processing and analysis, ease of interpreting visual representation of data characteristics and results, the ability to handle and learn from large sets of microdata, and automated applications which reduce human errors in data processing. Therefore, these results could be welcomed confidently compared to previous traditional and ML based applications in cost modelling.

This research also advances the accuracy of antecedent earthquake loss modelling in two ways. Firstly, a more accurate state-of-the-art AI-based model to predict earthquake damage repair cost rates was presented as opposed to simplified damage-cost function based models (Wanigarathna *et al.* 2022). The repair cost of a building damaged by an earthquake could be affected by several factors (Almufti and Willford 2013, Wedawatta *et al.* 2018, Kahandawa *et al.* 2021, Ortiz *et al.* 2021). However, the current approach to the antecedent prediction of direct damage repair cost rates involves a simplified damage-cost function based methodology (Wanigarathna *et al.* 2022). In this approach, the only variables considered in estimating the damage repair cost rates of a building are the damage level and the building's replacement cost or the building's market value. Wanigarathna *et al.* (2022) found that, whilst this method could provide a reasonably accurate estimate in relation to the repair cost for a large building stock (e.g. at a regional level

estimate), significant inaccuracies exist at individual building level estimates.

The literature review conducted for this study is conclusive on the limited use of ML not only in antecedent cost prediction but also in cost management in general, therefore this study makes a compelling contribution to reducing the dearth. Traditional regression techniques such as linear and polynomial regression model which (for a simple dataset) can be computed using statistical software such as SPSS. The f-regression and ridge regression were used in training before more complex features were added. The Gradient Boost Regression model developed in this research considered ten features to predict the damage repair cost rate. The performance of the model is significantly accurate and covers about 76% of the cases ( $R^2 = 0.76$ ). This model is also considered methodologically superior to the previous traditional and ML-based models used in cost modelling. Missing values and outliers are renown causes of errors for large dataset collected from different geographical locations. The Gradient Boost Regression model helped to overcome such shortfalls, and by combining together several weak models it was able established nonlinear relationships with the target variable CostPerSQM thereby predicting more variables such as the NON-HOUSING UNITS that were previously missed by traditional statistical methods (Wanigarathna *et al.* 2022). The predictive accuracy of ML models is also considered generally very high (Gao *et al.* 2023). Use of ML for cost modelling involves sophisticated computations that requires specific expertise to run and interpret results. To overcome such restrictions, this research recommends use of the SHapley Additive exPlanations to bridge the gap between the ML black box and human interpretation of the predicted models.

The prediction process used in this research is robust and replicable by other researchers, and model developed and presented here could be utilised to make accurate and trustworthy antecedent earthquake reconstruction cost estimates. Secondly, results revealed the impact of a range of factors that affects earthquake damage repair cost rates, including some previously unknown factors. Damage level, resultant operational level and Non-housing floor area were identified as the three most influential factors for damage repair cost rate prediction. The number of NON-HOUSING units, TOTAL FLOOR AREA and THE CONSTRUCTION DAYS were identified as the next most significant features. Whilst construction economics researchers and practitioners have a reasonably

good understanding of the impact of TOTAL FLOOR AREA and the CONSTRUCTION DAYS on construction prices (see Ashworth and Perera 2015 for example), to the best of authors' knowledge, this is the first research to highlight the significant impact of the NON-HOUSING units and their respective floor areas on damage repair cost rate estimates. Knowledge on these factors could additionally be used to supplement cost estimates produced based on the simplified damage-cost function based methods. For example, the results showed that larger values of NON-HOUSING AREA decrease the CostPerSQM. A cost estimator could then use this knowledge to adjust (a % uplift) a damage-cost function based estimate by evaluating the level of NON-HOUSING AREA involved in the concerned building. Altogether, these results will enable accurate antecedent reconstruction cost estimates, and hence would provide vital information for effective disaster preparedness (Wanigarathna *et al.* 2022) and reduce decisions made based on inappropriate assumptions (Hill and Rossetto 2008, De Martino *et al.* 2017).

This research has one key limitation. The ML-based models used here only evaluated the importance of cost prediction features/variables available within the data set. During the pre-processing stage, it was necessary to remove some high cardinality variables, as they make it difficult for ML models to identify patterns or causal relationships between features and the dependent variable. Some of these variables could significantly impact damage repair cost rates. For example, work description is the most accurate variable in estimating damage repair cost rate. Yet, irrespective of their methodology, high-level cost models used in antecedent cost estimates will not consider work description as a variable, as precise work descriptions could not be developed in advance. However, some removed variables could be useful in antecedent cost modelling. For instance, the removed variable 'ADDRESS' of properties are linked to the nature of the location (urban or rural), which is a key determinant of construction costs. This model could not analyse the relative importance of this variable. In addition, there could be other variables that could significantly impact damage repair cost rates, which were not available for the researchers to model. These include building age, building height, structural details of foundations and other elements, and other design elements such as division of spaces, materials used and shape of the buildings (Ashworth and Perera 2015). GEM Building Taxonomy Version 2.0 (Brzev *et al.* 2013) provides a detailed list of

variables related to buildings that would affect the level of damage from an earthquake event.

## Conclusion

Machine Learning has made advancements in data driven decision making in many fields. However, use of machine learning for construction cost modelling steps in antecedent disaster loss assessments is limited. This research set out to explore the suitability of state-of-the-art ML approaches to develop a cost model to predict earthquake damage repair cost rate (CostPerSQM) for residential buildings. This research evaluated six widely used ML-based models for their ability to predict damage repair cost rates for residential buildings. Gradient Boost Regression model outperforms other models, hence was used to develop a ML model to predict earthquake damage repair cost rate. The model found that the Damage level, resultant Operational level and Non-housing floor area as the three most influential factors for damage repair cost prediction. It is widely known that damage level and operational level as key variable in predicting damage repair cost rates, and used as the key variable in simplified economic loss assessment models based on damage to cost ratios. However, this is the first research to establish the importance and the relationship between the damage repair cost rates and the non-housing area (such as offices, shops within a residential block) within the building. These results could be therefore be used to uplift the loss estimates for mixed-use residential buildings derived based on simple damage to cost ratios based models in order to increase the accuracy of such estimates.

This research also contributed to the scholarly knowledge of use of Machine Learning in disaster management. Research investigated the suitability of six Machine Learning techniques in Disaster loss assessment: three distance-based regression models (Linear Regression, Polynomial Regression, Support Vector Regression), a Decision Tree Regression model and two Ensemble regression models (Random Forest Regression and Gradient Boosting Regression). All models except Polynomial Regression model found to be accurate in predicting damage repair cost rate. The performances of the Gradient Boosting Regression model and Random Forest Regression were significantly accurate and covers about 76%, 74% of the cases ( $R^2 = 0.76$ ,  $R^2 = 0.76$ ) respectively. The research developed a Gradient Boosting Regression model to predict the damage repair cost rate. The model reduced the limitations associated with black boxes of

some other ML models by revealing the main independent variables that affect the dependent variable, impact direction and correlation between dependent variables. Therefore, these results could be welcomed confidently compared to previous traditional and AI-based applications in cost modelling and simplified damage-cost function based antecedent disaster loss assessment models. The proposed approach also proved to be cost effective alternative which unpick insights from secondary data and the learning could be applied to other regions similar building typologies to estimate potential earthquake induced losses to residential building stock.

One of the key limitations in ML based secondary data analysis is that they learn from the secondary data collected for other purposes, hence their learning is limited by data availability. Literature reveals other variables that impact damage repair cost rates. Yet these were not available within the secondary data set for the researchers to model. Further research could be used to explore the possibilities of incorporating multiple datasets collected in the aftermath of earthquake disasters to develop ML based loss estimation models to incorporate more variables.

## Acknowledgements

This work was supported by the European Union's Horizon 2020 under Grant No 700748. The authors would like to thank LIQUEFACT project partners who supported in identifying data sources.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Data availability statement

Data used for this research is publicly available in the following portal. <https://openricostruzione.regione.emilia-romagna.it/ricostruzione-privata>

## References

- Ahiaga-Dagbui, D.D., and Smith, S.D., 2014. Dealing with construction cost overruns using data mining. *Construction management and economics*, 32 (7-8), 682–694.
- Alhamid, M. (2022). Ensemble models A guide to learning ensemble techniques in a simple walkthrough. TowardsDataScience. <https://towardsdatascience.com/ensemble-models-5a62d4f4cb0c>
- Almufti, I., and Willford, M., 2013. *Resilience-based Earthquake Design Initiative (REDi) for the Next Generation of Buildings*. San Francisco, CA: Arup

- Alpaydin, E., 2020. *Introduction to machine learning* [online]. MIT press. Available at: <https://mitpress.ubliish.com/ebook/introduction-to-machine-learning-4-preview/9935/Cover>.
- Ansal, A., 2015. *Perspectives on European Earthquake Engineering and Seismology: Volume 1*. Cham: Springer Open. Available at: <https://doi.org/10.1007/978-3-319-07118-3>.
- Ashworth, A., and Perera, S., 2015. *Cost studies of buildings*. London: Routledge.
- Baker, J.W., Cremen, G., Giovinazzi, S. and Seville, E., eds., 2016. *Proceedings of the 2016 annual conference of the New Zealand society for earthquake engineering*, Christchurch, New Zealand [on-line].
- Bird, J.F., and Bommer, J.J., 2004. Earthquake losses due to ground failure. *Engineering geology*, 75 (2), 147–179.
- Bommert, A., et al., 2020. Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational statistics & data analysis*, 143, 106839.
- Breiman, L., 1996. Bagging predictors. *Machine learning*, 26 (1996), 123–140.
- Breiman, L., 2001. Random forest. *Machine learning*, 45 (1), 5–32.
- Brzev, S., et al., 2013. GEM building taxonomy version 2.0, GEM Technical Report 2013-02 V1.0.0, 188 pp., GEM Foundation, Pavia, Italy.
- Canesi, R., and Marella, G., 2017. Residential construction costs: An Italian case study. *International Journal of Applied Engineering Research*, 12 (10), 2623–2634.
- Chen, W., and Zhang, L., 2022. An automated machine learning approach for earthquake casualty rate and economic loss prediction. *Reliability Engineering & System Safety*, 225 (2022), 108645.
- Chen, T and C. Guestrin, (2016). Xgboost: a scalable tree boosting system. In: *Proceedings 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, ACM. 785–794.
- Cremen, G., and Baker, J.W., 2019. A methodology for evaluating component-level loss predictions of the FEMA P-58 seismic performance assessment procedure. *Earthquake Spectra*, 35 (1), 193–210.
- Cunningham, T., 2017. Cost control during the construction phase of the building project: The consultant quantity surveyor's perspective. Technological University Dublin.
- Daniell, J.E., Schaefer, A.M., and Wenzel, F., 2017. Losses associated with secondary effects in earthquakes. *Frontiers in built environment*, 3, 30.
- De Groeve, T., Corbane, C., and Ehrlich, D., 2015. *Guidance for Recording and Sharing Disaster Damage and Loss Data: Towards the development of operational indicators to translate the Sendai Framework into action*. EUR 27192. Luxembourg (Luxembourg): Publications Office of the European Union; 2015. JRC95505.
- De Martino, G., Di Ludovico, M., Prota, A., Moroni, C., Manfredi, G. and Dolce, M, eds., 2017. *Proceedings of 6th International Conference on Computational Methods in Structural Dynamics and Earthquake Engineering*, At Rodi (Greece) [on-line].
- Del Vecchio, C., et al., 2018. Repair costs of existing RC buildings damaged by the L'Aquila earthquake and comparison with FEMA P-58 predictions. *Earthquake spectra*, 34 (1), 237–263.
- Del Vecchio, C., DI Ludovico, M., Pampanin, S. and Prota, A., (2017). Actual versus predicted repair costs: case studies on RC buildings damaged by L'Aquila earthquake. In: 16th World Conference on Earthquake Engineering. Santiago, Chile during January 09-13, 2017.
- Del Vecchio, C., Ludovico, M.D., and Prota, A., 2020. Repair costs of reinforced concrete building components: from actual data analysis to calibrated consequence functions. *Earthquake spectra*, 36 (1), 353–377.
- Effrosynidis, D., and Arampatzis, A., 2021. An evaluation of feature selection methods for environmental data. *Ecological informatics*, 61, 101224.
- Emsley, M.W., et al., 2002. Data modelling and the application of a neural network approach to the prediction of total construction costs. *Construction management & economics*, 20 (6), 465–472.
- Ferguson, M., and Martin, A., 2015. Earthquake-induced structural damage classification algorithm, CS 229 ML – final report. Dept. of Civil and Env. Engineering, Stanford University.
- Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 29 (5), 1189–1232.
- Gao, B., et al., 2024. Prediction and driving factors of forest fire occurrence in Jilin Province. *China journal of forest research*, 35, 21.
- Geekiyana, D., and Ramachandra, T., 2020. Nexus between running costs and building characteristics of commercial buildings: hedonic regression modelling. Built environment project and asset management. *Built environment project and asset management*, 10 (3), 389–406.
- Grootendorst, M. (2019). *Opening Black Boxes: How to leverage Explainable ML. Using PDP, LIME, and SHAP to create interpretable decisions that create value for your stakeholders*. TowardsDataScience. <https://towardsdatascience.com/opening-black-boxes-how-to-leverage-explainable-machine-learning-dd4ab439998e>
- Guo, C., et al., 2020. Deep learning technique for process fault detection and diagnosis in the presence of incomplete data. *Chinese journal of chemical engineering*, 28 (9), 2358–2367.
- Hashemi, S., Ebadati, O.M., and Kaur, H., 2020. Cost estimation and prediction in construction projects: a systematic review on ML techniques. *SN applied sciences*, 2 (10), 1–27.
- Hassija, V., et al., 2023. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive computation*, 16 (1), 45–74.
- Hastie, T., 2020. Ridge regularization: An essential concept in data science. *Technometrics*, 62 (4), 426–433.
- Hill, M., and Rossetto, T., 2008. Comparison of building damage scales and damage descriptions for use in earthquake loss modelling in Europe. *Bulletin of earthquake engineering*, 6 (2), 335–365.
- Iglesias, J., et al., 2024. Predictive modeling of severe weather impact on individuals and populations using Machine Learning. *International journal of disaster risk reduction*, 105, 104398.
- Ji, S., et al., 2019. Cost estimation model using modified parameters for construction projects. *Advances in civil engineering*, 2019 (Art. no. 8290935). <https://doi.org/10.1155/2019/8290935>

- Kahandawa, R., et al., 2021. Earthquake damage repair loss estimation in New Zealand: what other variables are essential based on experts' opinions? *Buildings*, 11 (9), 385.
- Kameshwar, S., Misra, S., and Padgett, J.E., 2020. Decision tree based bridge restoration models for extreme event performance assessment of regional road networks. *Structure and infrastructure engineering*, 16 (3), 431–451.
- Kim, J., et al., 2018. Development of a maintenance and repair cost estimation model for educational buildings using regression analysis. *Journal of Asian architecture and building engineering*, 17 (2), 307–312.
- Kim, J.M., et al., 2016. Estimating the Texas Windstorm insurance association claim payout of commercial buildings from Hurricane Ike. *Natural hazards*, 84, 405–424.
- Kong, Q., et al., 2019. Machine learning in seismology: Turning data into insights. *Seismological research letters*, 90 (1), 3–14.
- Kong, Z., et al., 2020. Condition monitoring of wind turbines based on spatio-temporal fusion of SCADA data by convolutional neural networks and gated recurrent units. *Renewable energy*, 146, 760–768.
- Kousky, C., and Michel-Kerjan, E., 2017. Examining flood insurance claims in the United States: Six key findings. *Journal of risk and insurance*, 84 (3), 819–850.
- Li, H., and Love, P.E., 1999. Combining rule-based expert systems and artificial neural networks for mark-up estimation. *Construction management & economics*, 17 (2), 169–176.
- Linardos, V., et al., 2022. Machine learning in disaster management: Recent developments in methods and applications. *Machine learning and knowledge extraction*, 4 (2), 446–473.
- Lin, A., et al., 2020. A big data-driven dynamic estimation model of relief supplies demand in urban flood disaster. *International journal of disaster risk reduction*, 49, 101682.
- Liu, Z., et al., 2019. Accuracy analyses and model comparison of machine learning adopted in building energy consumption prediction. *Energy exploration & exploitation*, 37 (4), 1426–1451.
- Lodhia, Z., Rasool, A., and Hajela, G., 2017. A survey on ML and outlier detection techniques. *IJCSNS international journal of computer science and network security*, 17 (5), 271–276.
- Lü, X., et al., 2015. Modeling and forecasting energy consumption for heterogeneous buildings using a physical-statistical approach. *Applied energy*, 144, 261–275.
- Lundberg S. M., and Lee S., 2017. A unified approach to interpreting model predictions. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
- Lundberg, S. M., Erion, G. G., and Lee, S.-I., 2019. *Consistent individualized feature attribution for tree ensembles*. arXiv preprint arXiv:1802.03888.
- Ortiz, D., Reinoso, E., and Villalobos, J.A., 2021. Assessment of business interruption time due to direct and indirect effects of the Chiapas earthquake on September 7th 2017. *Natural hazards*, 108(3), 2813–2833.
- Parameswaran, T., Jayawickrama, T. S., and Melagoda, D. G., 2019. Analyzing the impact of location factors on building construction cost in Sri Lanka. In: Proceedings of the International Conference on Industrial Engineering and Operations Management. Bangkok, Thailand.
- Petch, J., Di, S., and Nelson, W., 2022. Opening the black box: the promise and limitations of explainable ML in cardiology. *Canadian journal of cardiology*, 38 (2), 204–213.
- Polese, M., et al., 2015. Assessing reparability: simple tools for estimation of costs and performance loss of earthquake damaged reinforced concrete buildings. *Earthquake engineering & structural dynamics*, 44 (10), 1539–1557.
- Rafiei, M.H., and Adeli, H., 2018. Novel ML model for construction cost estimation taking into account economic variables and indices. *Journal of construction engineering and management*, 144 (12), 04018106.
- Ribeiro, M. T., Singh, S., and Guestrin, C., 2016. *Model-agnostic interpretability of ML*. In: ICML Workshop on Human Interpretability in ML (WHI 2016), New York, NY. [https://arxiv.org/pdf/1606.05386.pdf?source=post\\_page](https://arxiv.org/pdf/1606.05386.pdf?source=post_page)
- Ruddock, L., et al., 2010. Post-tsunami reconstruction in Sri Lanka: Assessing the economic impact. *International journal of strategic property management*, 14 (3), 217–230.
- Seeley, I. H., 1997. *Quantity surveying practice*. 2nd ed. Basingstoke: MacMillan Press Ltd.
- Shapley, L.S., 1953. A value for n-person games. *Contributions to the theory of games*, 2 (28), 307–317.
- Spence, R., et al., 2003. Comparing loss estimation with observed damage: a study of the 1999 Kocaeli earthquake in Turkey. *Bulletin of earthquake engineering*, 1 (1), 83–113.
- Shirzadi, A., et al., 2017. A comparative study between popular statistical and machine learning methods for simulating volume of landslides. *Catena*, 157, 213–226.
- Stojadinovic, Z., Kovacevic, M., Marinkovic, D., and Stojadinovic, B, eds. (2017). Proceedings of the 16th world conference on earthquake engineering (16WCEE). SANTIAGO-CHILI.
- Sun, W., Bocchini, P., and Davison, B.D., 2020. Applications of artificial intelligence for disaster management. *Natural hazards*, 103 (3), 2631–2689.
- Sutton, J., and Tierney, K., 2006. *Disaster preparedness: Concepts, guidance, and research*. Colorado USA: University of Colorado.
- Tocchi, G., et al., 2023. The use of machine-learning methods for post-earthquake building usability assessment: A predictive model for seismic-risk impact analyses. *International journal of disaster risk reduction*, 97, 104033.
- Trinchero, R., and Canavero, F., 2021. Machine learning regression techniques for the modeling of complex systems: An overview. *IEEE electromagnetic compatibility magazine*, 10 (4), 71–79.
- UNISDR 2017. *UNISDR strategic framework 2016-2021*. [https://www.unisdr.org/files/51557\\_unisdrstrategicframework20162021pri.pdf](https://www.unisdr.org/files/51557_unisdrstrategicframework20162021pri.pdf)
- Vona, M., et al., 2018. An optimized procedure to estimate the economic seismic losses of existing reinforced concrete buildings due to seismic damage. *Buildings*, 8 (10), 14 4.
- Wang, W., Li, L., and Qu, Z., 2023. Machine learning-based collapse prediction for post-earthquake damaged RC columns under subsequent earthquakes. *Soil dynamics and earthquake engineering*, 172, 108036.
- Wanigarathna N, Henjewe, C., Jones K, Morga M, Pascale F (2022). Modelling earthquake damage repair costs: Improving accuracy for preparedness decision making. In: 38th Association of Researchers in Construction Management conference (ARCOM), Glasgow, UK.

- Wanigarathna, N., Yarovaya, L., Pascale, P., Morga, M., and Jones, K. (2018). *Deliverable D5.3 Community resilience and cost/benefit modelling: Socio-technical-economic impact on stakeholder and wider community*. <http://www.liquifact.eu/wp-content/uploads/2018/12/D5.3-Community-resilience-and-costbenefit-modelling.pdf>
- Wanous, M., Boussabaine, H.A., and Lewis, J., 2003. A neural network bid/no bid model: the case for contractors in Syria. *Construction management and economics*, 21 (7), 737–744.
- Wedawatta, G., Ingirige, B. and Sugathapala, K. 2018. Long-term sustainability and performance of post-disaster housing projects. Chartered Institute of Building (CIOB) Bowen Jenkins Legacy Research Fund Research Report. Aston University, Birmingham, UK. [https://research.aston.ac.uk/portal/files/25966154/CIOB\\_report\\_Wedawatta\\_et\\_al\\_Final.pdf](https://research.aston.ac.uk/portal/files/25966154/CIOB_report_Wedawatta_et_al_Final.pdf)
- Xie, Y., et al., 2020. The promise of implementing machine learning in earthquake engineering: A state-of-the-art review. *Earthquake spectra*, 36 (4), 1769–1801.
- Yu, S., et al., 2015. Quantitative assessment of disaster resilience: An empirical study on the importance of post-disaster recovery costs. *Reliability engineering & system safety*, 137, 6–17.
- Zhao, J., et al., 2024. Multi-source driven estimation of earthquake economic losses: A comprehensive and interpretable ensemble machine learning model. *International journal of disaster risk reduction*, 106, 104377.