Earthquake Damage Prediction of Buildings in Nepal using Machine Learning tools.

Subash Ghimire* (1), Philippe Gueguen (1), Danijel Schorlemmer (2)

1. *Subash Ghimire, PhD scholar, ISTerre, University of Grenoble Alpes, France, subash.ghimire@univ-grenoble-alpes.fr
1. Philippe Gueguen, ISterre, University of Grenoble Alpes, philippe.gueguen@univ-grenoble-alpes.fr
2. DanijelSchorlemmer, GFZ German Research Centre for Geosciences, Potsdam, Germany, ds@gfz-potsdam.de

Abstract

Decision-makers and stakeholders need rapid assessment of the potential damage following earthquake events to develop and execute disaster risk reduction strategies and to systematically respond to the emerging situation in post-disaster situations. Classical risk assessment methods are resource- and time-consuming. In this study, the Mw 7.8 Gorkha, 2015 Nepal Earthquake crowd-sourced building damage data is used to explore the efficiency of various machine-learning techniques in rapid earthquake-induced building damage assessment. The Random Forest Regressor showed the best performance among several machine learning methods considered in this study. For rapid seismic damage assessment in Nepal, for a given earthquake scenario, the building features data collected from the existing built-up environment can be used as an input to this model and the output will help decision-makers to take appropriate decisions.

Key words

Machine learning, seismic risk assessment, building damage portfolios, building damage prediction

1. Introduction

Earthquakes are less frequent in occurrences but contribute significantly to physical and social consequences. On average, since 1990-2017, annually, earthquakes result around USD 34.7 billion losses globally (OECD, 2018) and USD 5 billion losses in Nepal (UNDRR, 2019). It is crucial for decision-makers and stakeholders to have rapid assessments of potential damage due to earthquake events (Bommer & Crowley, 2006). For a successful emergency response planning before and after an earthquake, the spatial distribution of damage over the built environment is required (Earle et al., 2010; Ranf et al., 2007). Various classical methods exist for estimating earthquake-induced building damage based on ground shaking. These methods require a lot of information on building portfolios and earthquake ground motion. This makes seismic risk assessment at regional/urban scale quite challenging because the collection of building information and application of damage assessment methods is time and resource consuming.

For the last decade, the progress in artificial intelligence (AI) tools and their application in various domains has increased. Yet, there is only a very limited number of applications of AI for rapid seismic risk assessment. Riedel et al.(2014, 2018) showed the ability of the Support Vector Machine for seismic vulnerability assessment at urban or regional scales. Mangalathu et al. (2020) showed an application of the machine learning technique in rapid seismic risk assessment using an earthquake damage data portfolio of the 2014 South Napa earthquake. They concluded that the use of the rapidly growing machine learning technique in the field of rapid seismic risk assessment provides a reliable estimate of the earthquake-induced potential building damage. To assure the use of AI technique in seismic risk assessment, investigation on the efficiency and relevancy of AI technique in seismic damage assessment at regional scale is required.

Moreover, building-damage portfolios of earthquake events are starting to become openly accessible. For example, the National Planning Commission of Nepal (http://eq2015.npc.gov.np/) shared a massive household data survey of the damaged buildings after the Mw 7.8 2015 Gorkha Nepal earthquake. The objective of this paper is to test the effectiveness and relevancy of several AI methods for predicting spatially distributed seismic damage. This article presents the results on the performance of various machine learning models in rapid damage earthquake assessment using the Nepal earthquake damage portfolio.

2. Description of the Damage Database

On 25 April 2015, a devastating earthquake of M_w 7.8 hit the central Nepal with an epicentral about 80km NW from Kathmandu, hypocentre depth of 8.2 km, and 120 km rupture length towards the east. Thousands of households were damaged, around 8 million people were affected (8,790 fatalities and 22,300 injuries). The 2015 Nepal earthquake building-damage database consists of 762,106 building datasets collected in eleven districts of Nepal (Fig. 1). The severity of damage is grouped into five grades observed by visual inspection. Similarly, the information about each building feature: number of stories, age of the building, height, plinth area, construction material, ground slope condition, building position with respect to another building, and roof type were also assigned during visual observation. The detailed description of these five grades and building features is available on the same website (http://eq2015.npc.gov.np/docs/#/faqs/faqs). The geo-localization of buildings was provided in the ward level, ward is the smaller administrative unit. In addition, the ground motion data is added to the database from the ShakeMap tool from the United States Geological Survey. In this study, macroseismic intensities (MSI) map from the ShakeMap is considered as an input ground motion (Fig. 2) and assigned to all the buildings located in the same ward.

In the database, number of story ranges from 1-9 storey (Fig. 3a), age ranges from 1-200 years (Fig. 3b), plinth area ranges between 70 to 5000 sq. ft. (Fig. 3c), height ranges between 6-97 ft. (Fig. 3d). The MSI value ranges from 5.30 to 8.30 (Fig. 3e). Likewise, 82.89 (%) /13.86 (%) / 3.24 (%) of the buildings were located in, respectively, flat/moderate/steep slope, (Fig. 3g), 28.05 (%) / 66.10 (%) / 7.85 (%) buildings were associated with heavy / light/ RC roofing-system, respectively (Fig. 3h). Similarly, 79.31 (%) / 16.98 (%) / 3.53 (%) / 0.17 (%) of buildings were stand-alone / one-side-attached / two-side-attached / three-side-attached to another building (Fig. 3i). The distribution of the buildings according to damage grades (DG) in the database is: 10.34 (%) in DG1, 11.45 (%) in DG2, 17.90 (%) in DG3, 24.12 (%) in DG4, and 36.19 (%) in DG5 (Fig. 3f).

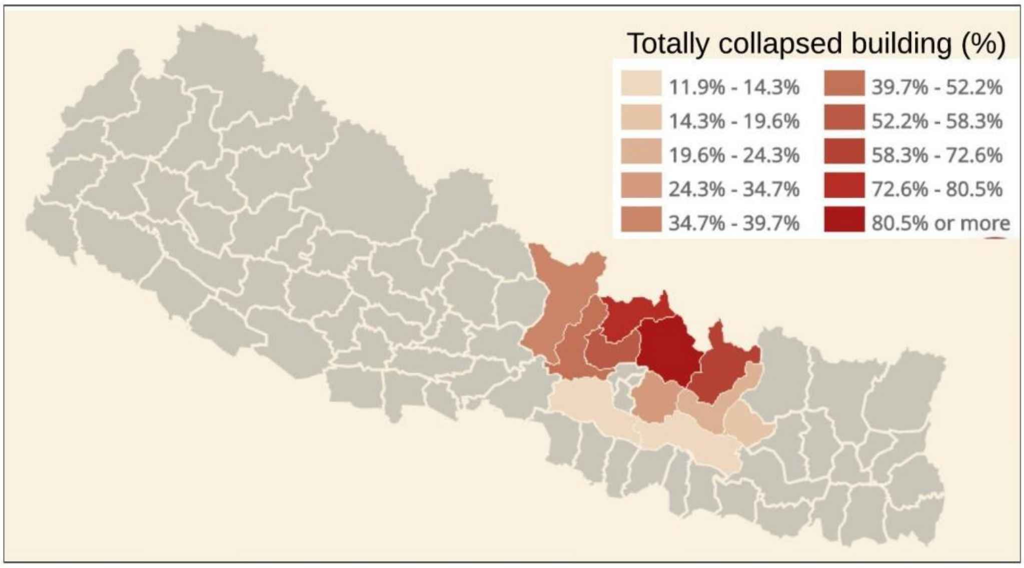

Figure 1. Location of 11 districts where the 2015 Nepal earthquake building damage data are available. It also illustrates the severity of the earthquake effect in each district in terms of the collapsed buildings. (Source: http://eq2015.npc.gov.np/#/compare).

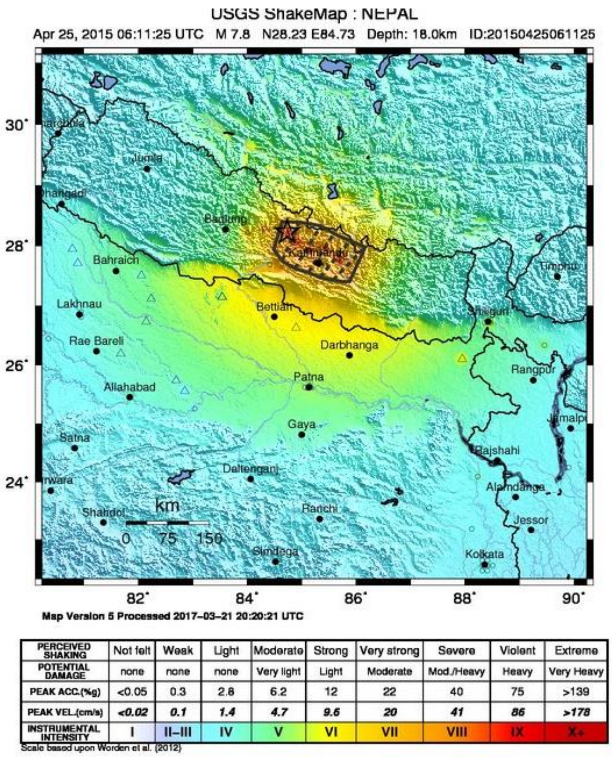

Figure 2. Spatial distribution of 2015 Nepal earthquake ground motion intensity. (Source: https://earthquake.usgs.gov/earthquakes/eventpage/us20002926/shakemap/intensity).

3. Method

This study assessed the efficiency of Linear Regression (LR), Support Vector Regressor (SVR), Gradient Boosting Regression (GBR), Random Forest Regression (RFR), Gradient Boosting Classification (GBC) and Random Forest Classification (GBC) in damage prediction. A brief description of these methods is provided in the annex. Interested readers are suggested to refer to Friedman et al. (2001) and scikit-learn machine learning in Python (Pedregosa et al., 2011) for detailed information on these machine-learning methods. 0.48% of the dataset was observed with missing values. The missing data points associated with categorical variables (damage grades, ground slope, material, roof type and position) were removed and the outliers associated with the numerical variables (number of storeys, age, the height of buildings) were replaced by their respective mean value. The entire dataset is randomly divided into training and testing subsets. Following the recommendation of Friedman et al. (2001), 70% of the data is used as a training set and 30% is used as a testing set. The training set is used to train the machine learning model and the testing set is used to observe the predictive performance of the machine learning model. For each machine-learning model, the features of buildings (number of storeys, height, age, plinth area, ground slope condition, position, roof material, construction material), as well as the intensity of ground motion, are defined as input features and damage grades as response variables. The performance of each machine learning model is evaluated through the coefficient of determination (R^2 scores) and Root Mean Square Error (RMSE) scores for regression and accuracy scores for classification problems. Higher the value of R^2, accuracy score and lower the RMSE value, better is the performance of the model.

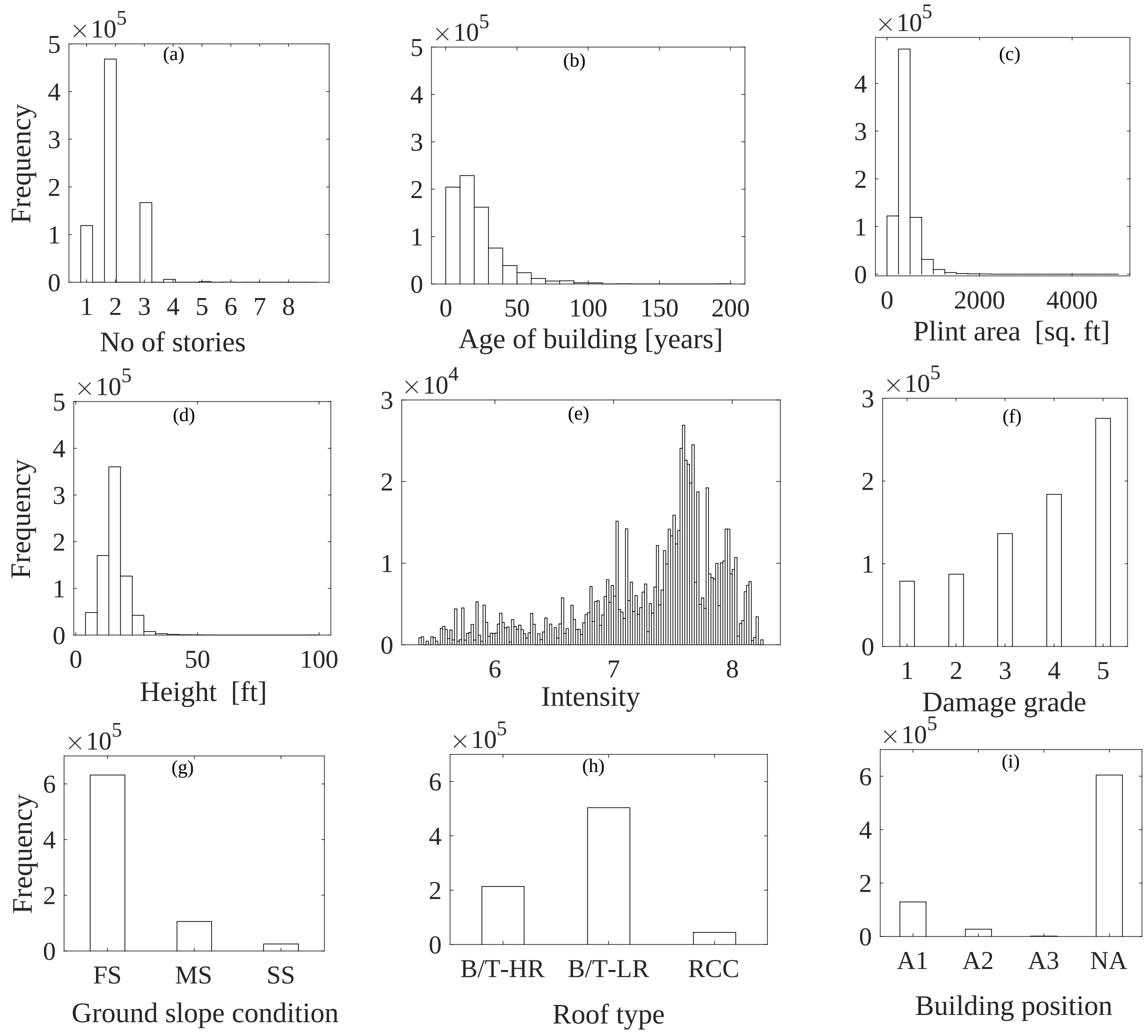

Figure 3. Distribution of different features in the dataset. The y-axis is the frequency and the x-axis in frame is (a) number of story, (b) age of the building, (c) plinth area of building, (d) height of the building (e) macroseismcic intensity, (f) damage grade, (g) ground slope condition at building location (h) type of construction material used in roof, and (i) position of building with respect to another building. In frame (g) FS/MS/SS represent flat/mild/steep slope, respectively. In frame (h) B/T-HR, B/T-LR, represent bamboo/timber-heavy-roof, bamboo/timber- light-roof and RCC represents reinforced cement concrete. In frame (i) A1/A2/A3 and NA represent attached with one/two/three sides and not attached, respectively.

4. Results and Discussion

The LR and SVR are observed to have the values of R^2 score equal to 0.41 and 0.38 and RMSE score equal to 1.06 and 1.08, respectively. The lowest R^2 value and the highest RMSE value for LR and SVR methods prove less suitable for this dataset. They oversimplified the complex non-linear interaction among the features present in the dataset. Similarly, the GBC and RFC methods are observed to have an accuracy score of 0.33 and 0.55, respectively. GBC and RFC are also unable to classify the true damage grade with high accuracy. The highest values of R^2 score are 0.58 and 0.56, and the lowest RMSE values are 0.88 and 0.87 are observed for GBR and RFR, respectively. These methods give higher efficiency in the damage prediction. GBR and RFR can reproduce the stronger non-linear interaction that exists among different features present in the dataset.

The performance, effectiveness, and computational time of these methods are very sensitive to the value of model parameters (hyperparameters). The GBR method requires careful tuning of a greater number of hyperparameters as compared to RFR. Thus, RFR is observed to be the most efficient method in building-damage prediction.

Fig. 4 shows the results of the RFR method in the test dataset. Few misclassifications are pointed out both by considering the frequency of correctly assessed DGs i.e. predicted damage is within one step from the observed value and the median value of assessed DGs that deviate from the classification provided in the field surveys. This illustrates the high strength of RFR method in damage prediction, which is very crucial from the perspective of seismic risk assessment. Thus, using RFR model, the spatial distribution of seismic damage can be predicted using the basic features of buildings and building-damage information from the existing post-disaster survey and vulnerability assessment with a reasonable level of accuracy.

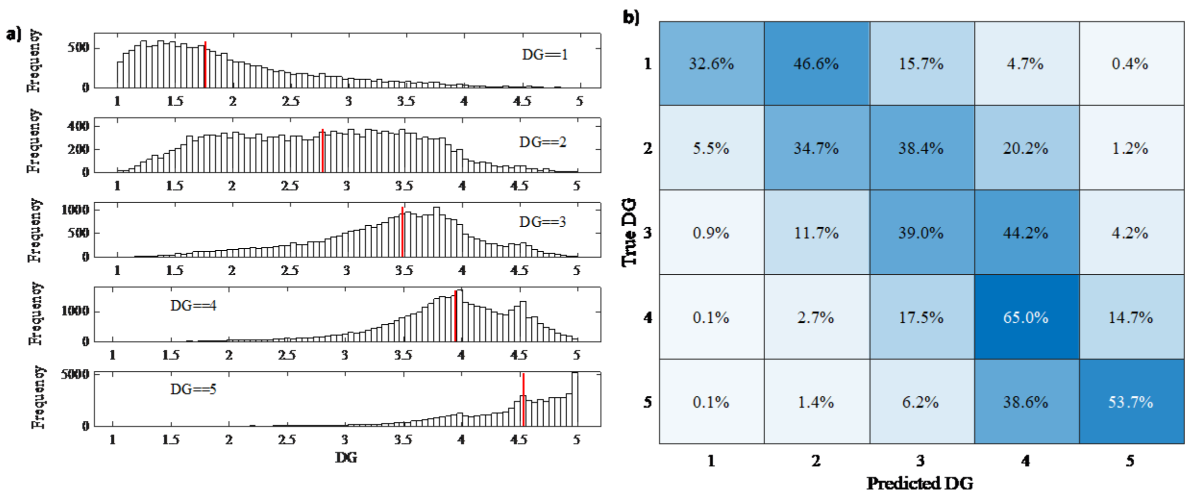

Figure 4. Graphical representation of the predictive performance of the RFR model on the test dataset. In frame (a) the x-axis is the predicated damage grade (DG) and the y-axis is the frequency. The red

vertical line represents the median value. The true damage grade is noted in the same subplot. In frame (b) the x-axis is the predicted DG and the y-axis is the true DG.

5. Conclusion

The efficiency and relevancy of machine learning techniques in rapid seismic risk assessment is studied using the 2015 earthquake building damage data from Nepal. Performance of Linear Regression, Support Vector Regression, Gradient Boosting Regression, Random Forest Regression, Gradient Boosting Classification, and Random Forest Classification in building-damage prediction using basic features of building was tested. The Random Forest Regression is observed to be the most efficient in damage prediction. A reasonable estimate of the damage at a given level of the ground motion is possible using basic features of building and RFR model, resolving the time and resource consumption issues.

The 2015 Nepal earthquake building-damage portfolio and the RFR model can be used for the site specific or global rapid seismic risk assessment in Nepal i.e. using the RFR model trained on the 2015 Nepal earthquake building-damage dataset, we can predict potential damage for a given earthquake scenario by considering the same input features data collected from the existing built-up environment. The output of such assessment model may assist stakeholders and decision-makers in rapid seismic risk assessment in order to formulate and implement new plans and policies in earthquake disaster risk reduction.

The 2015 Nepal earthquake building-damage dataset can be used as a powerful tool for seismic risk assessment in Nepal. The building-damage database is associated with significant amount of noise. Fine refinement of the existing dataset including all available post-disaster building damage data is recommended. Similarly, the development of national building database collecting key information of building is necessary to facilitate seismic risk assessment in Nepal.

As a future perspective, further investigation in rapid seismic risk assessment should be carried out by considering the key building features (number of storeys, plinth area, age, height etc.) that are easily accessible and could be used as a good proxy to predict building damage using the most suitable machine learning technique. Investigation of the applicability of the machine learning model with other open-data platforms like OpenStreetMap (OSM) should be investigated for rapid seismic risk assessment.

6. Acknowledgement

This work is part of the URBASIS program led by P.G at ISTerre/Université de Grenoble Alpes. We thank LabEx OSUG@2020 (Investissements d'avenir-ANR10LABX56 and the ITN-MSCA URBASIS project, a project funded by the EU Horizon 2020 program under Grant Agreement Number 813137. Part of this work was supported by the Real-time earthquake rIsk reduction for a reSilient Europe (RISE) project, funded by the EU Horizon 2020 program under Grant Agreement Number 821115. S.G. would like to thank Kathmandu Living Labs for their most valuable assistance.

7. References

Bommer, J. J., & Crowley, H. (2006). The influence of ground-motion variability in earthquake loss modelling. *Bulletin of Earthquake Engineering*, *4*(3), 231–248. https://doi.org/10.1007/s10518-006-9008-z

Earle, P. S., Wald, D. J., Jaiswal, K. S., Allen, T. I., Hearne, M. G., Marano, K. D., Hotovec, A. J., & Fee, J. M. (2010). Prompt assessment of global earthquakes for response (pager): A system for rapidly determining the impact of earthquakes worldwide. *Earthquake Research: Background and Select Reports*, 31–46.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* ((Vol. 1, N). New York: Springer series in statistics.

Mangalathu, S., Sun, H., Nweke, C. C., Yi, Z., & Burton, H. V. (2020). Classifying earthquake damage to buildings using machine learning. *Earthquake Spectra*, *36*(1), 183–208.

https://doi.org/10.1177/8755293019878137

OECD. (2018). *Financial Management of Earthquake Risk*. 108. http://www.oecd.org/finance/insurance/Financial-management-of-earthquake-risk.pdf

Pedregosa, F., Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., & Mueller, A. (2011). Scikit-learn. *GetMobile: Mobile Computing and Communications*, *19*(1), 29–33. https://doi.org/10.1145/2786984.2786995

Ranf, R. T., Eberhard, M. O., & Malone, S. (2007). Post earthquake prioritization of bridge inspections. *Earthquake Spectra*, *23*(1), 131–146. https://doi.org/10.1193/1.2428313

Riedel, I., Guéguen, P., Dunand, F., & Cottaz, S. (2014). Macroscale vulnerability assessment of cities using association rule learning. *Seismological Research Letters*, *85*(2), 295–305. https://doi.org/10.1785/0220130148

Riedel, Ismaël, & Guéguen, P. (2018). Modeling of damage-related earthquake losses in a moderate seismic-prone country and cost–benefit evaluation of retrofit investments: application to France. *Natural Hazards*, *90*(2), 639–662. https://doi.org/10.1007/s11069-017-3061-6

UNDRR. (2019). *Disaster Risk Reduction in Nepal: Status Report 2019*. 1–30.

Annex

Linear Regressor

Linear Regression (LR) explains the relationship between target variables through a linear combination of input (predictors) variables. The functional form of the LR is given below as:

$$Y = \sum_{i=0}^{n} w_i x_i = w^{\mathrm{T}} x$$

Here, the weight w_0 represents the y-axis intercept and w_i is the weight coefficient of the input variable, and Y is the target variable. The LR fits a linear model with coefficients $w = (w_1, \ldots, w_p)$ to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation. The LR has simple analytical and computational properties. They provide an adequate interpretable description of how the input affects the output. This method is computationally efficient. The weight associated with each input variable helps in features importance identification. The LR is oversimplified (unable to capture the complexity of the problem), and is very sensitive to outliers. The LR assume that data are linearly separable, special attention should be paid with multicollinearity issues, not very efficient to nonlinear data (https://scikit-learn.org/stable/modules/linear_model.html).

Support Vector Regressor

Support vector machines (SVM) is a set of supervised learning methods used for classification, regression, and outlier detection. In SVM, the input features are transformed into a higher-dimensional space where two classes can be linearly separated by a high dimensional space called a hyperplane. The SVM was originally used for classification problems and then extend to regression problems called Support Vector Regression (SVR). SVR maintains all features of SVM. The model produced by SVR depends only on the subsets of the training dataset because the cost function ignores samples whose prediction is close to their target. Three types of implementation are possible for SVR: SVR, Nu-SVR, and Linear SVR. SVM is effective in high dimensional spaces, memory efficient, versatility in kernel functions. This method is more suitable when the number of features in more than the number of data. SVM is less suitable when the number of data points is so large, they do not provide direct probability estimate, overfitting could be an issue when the number of features is larger than the of data points (https://scikit-learn.org/stable/modules/svm.html).

Gradient Boosting

Gradient Boosting (GB) is a generalization of boosting to the arbitrary differentiable loss function. The GB is based on an ensemble of several decision trees. A decision tree represents a set of conditions or restrictions that are hierarchically organized and successively applied from a root to a lead of the tree. The GB is an accurate and effective procedure that can be used for both regression and classification. It is shown that both the approximation accuracy and execution speed of the GB can be substantially improved by incorporating randomization into the procedure. Specifically, at each iteration, a subsample of the training data is drawn at random (without replacement) from the full training data set. This randomly selected subsample is then used in place of the full sample to the base learner and compute the model update for the current iteration. This randomized approach also increases robustness against the overcapacity of the base learner. The GB has lots of flexibility in terms of the loss function. They can easily handle missing data, often works great with categorical and numerical data. This is sometimes computationally expensive, requires careful tuning of hyperparameters (model input parameters). (https://scikit-learn.org/stable/modules/ensemble.html#gradient-boosting).

Random Forest

Random Forest (RF) ensemble the performance of several decision trees to classify or predict the value of variables, which is based on bagging. Decision trees are trained by using a random subset of the original features. The RF can model complex relationships in the data and account for non-linear relationships between predictor and response variables by the adaptive nature of the decision rules. The RF has better generalization performance, less sensitive to outliers, does not require tuning of many hyperparameters. It works with continuous and also categorical predictors and also can handle missing data (https://scikit-learn.org/stable/modules/ensemble.html).