Modelling Earthquake Damage

Shaima Shoukat
0642224

Big Data Analytics Program, Trent University
Peterborough, ON
shaimashoukat@trentu.ca

Abstract – This paper illustrates the application of machine learning to predict the earthquake damage grade of buildings. The project focussed on feature engineering techniques that addressed categorical features with high cardinality. Two techniques, namely target encoding and entity encoding were adopted. Altogether, the results of three models namely, Random Forests, Neural Networks and Logistic Regression; were compared. It is a multi-class classification problem wherein the model predicts the damage severity that could be low, medium or high; based on the location and construction of the buildings.

I. INTRODUCTION

The application of Machine Learning techniques to disaster prediction and response has been gaining momentum in recent years. It facilitates decision makers to respond to and adopt disaster mitigation measures. Various ML approaches aid in answering questions like 'When is a disaster likely to occur', 'What would be its severity?', 'What is the extent of damage', etc. This project attempts to answer one such question, 'What is the damage grade of buildings after an earthquake based on its location and structural specifications?'. The National Planning Commission of Nepal originally collected the earthquake data. Later, a subset of it was curated and made available as a part of an online competition on [1].

The 2015 Earthquake of Nepal, measuring 7.8 on the Richter's scale, caused widespread loss of lives and damage to property. Owing to Nepal's underdeveloped network to measure seismic activity, neighbouring China sent eight teams of scientists and experts to study the earthquake after-effects and evaluate the extent of destruction to property. In addition to this, lack of proper implementation of building code combined with old construction practices led to increased failure of structures [2]. The scientists, therefore studied how distance from the epicentre and different materials used in construction responded to the earthquake. This is especially important in Nepal because it lies on the fault line between the Indian and Eurasian plates, thereby being prone to frequent high intensity earthquakes. The Government then conducted a survey to estimate the extent of damage to each structure along with other socio-economic factors. The goal was to grant compensation to individuals based on the damage to their homes and property [1].

II. LITERATURE REVIEW

Buildings in Nepal mostly fall in the following four categories based on foundation type: Mud-bonded masonry, cement-bonded masonry, wooden structures and RC frames [2]. Majority of the buildings in Nepal are made of mud-bonded masonry due to its low-cost, especially in rural areas. This type of construction makes use mud to hold together bricks but has low load bearing capacity [2]. Over 50% of such houses were severely damaged to beyond repair and only 3 % survived the after shocks with little cracks [2]. Cement-bonded masonry faired much than mud-bonded masonry. This uses cement to hold together bricks but is not widely used there. 30% of such houses were severely damaged to beyond repair and 17 % had very little damage to them [2]. Wooden frames made of timber have high load bearing capacity. Such kind houses account up to 25% of the total houses in Nepal. However, timber frames on masonry or with infill walls lower the aseismic performance and increase the collapse rate of buildings. Reinforced Concrete type of foundation is popular in cities and gave the best aseismic performance. Only 10 % of such buildings were severely damaged to beyond repair and nearly 40 % suffered little to no damage at all [2]. In addition to the foundation type, other factors were responsible for collapsed structures. Roofs often buckled due to the heavy weight of roof tiles or RC slabs used. Foundations were not built on levelled terrain causing the buildings to crumble [3]. Masonry constructions with shared walls that were not connected firmly with mud mortar gave away [3]. In addition to this, older masonry structures

suffered extensive damage than new ones. Mud mortar used in masonry units failed to bind the bricks and stones effectively [3]. In some areas, not taking account the soil type and corresponding foundation design guidelines, led to foundations sinking in, buildings cracking and falling completely [3]. All these practices including structural and non-structural deficiencies greatly increased the magnitude of loss.

The effects of natural disasters involve a range of complex factors. Understanding these factors and coming up with mitigating solutions is not always easy due to their quantity and complexity. In such cases, Machine learning can be effectively used to procure solutions. One such application is to use remote sensing data to analyse post-earthquake damage.

Authors in [4] employ multilayer feed forward neural networks, radial basis NNs and RFs to analyse earthquake damage that occurred in Haiti in 2010.They use high-resolution remote sensing data to obtain pre and post disaster images from Digital Globe Foundation [4]. Nearly 1.5 million pixels were used to train and 900 buildings were used for testing to classify them on a scale of 1 to 5, % being the worst. Structural spatial information like Laplacian of Gaussian was extracted to improve accuracy. Dissimilarity and dimensionality reduction were used extract textural information from the images [4]. Data with these features was converted into 14-D arrays. 9 input features trained an ANN, with two hidden layers, using back propagation method. The MLA called radial basis function was built using basis functions [4]. The third algorithm used was Random Forests. The results showed that ANN performed the best with accuracy of 74% and a false negative error of 38% [4].

The authors in [5] compared the performance of three different models to assess the damage from the 2014 South Napa earthquake:

First, Random Forests were used to classify damage levels after earthquake. Features like earthquake intensity, building construction properties and distance from the epicentre are modelled to predict the damage class [5]. Out of the 2227 samples, 75% were used for training and the rest for testing. Second, Natural Language Based Prediction Model was deployed on the post earthquake survey data. Out of the 3243 records, 75% were used for training and the rest for testing. The Third model was based on convoluted neural networks that used images of the damaged buildings. This model was built using 227 images of buildings. The second model based on Natural Language Processing fared the best with 87% accuracy. The first and third were able to predict accurately only 63 % and 64% of the time respectively.

Author in [6] proposes a novel method to assess structural damage to tall because they are built using different procedures. The author makes use of Non Linear Response History analysis to extract the response patterns of the building. These response patterns refer to the drifts and damage to structural components like load-bearing stud wall. Damage simulation is carried out to extract damage patterns of primary structural components that can be identified visually as being damaged. The above two patterns can be obtained by simulating motions like the earthquake. One motion generates one response pattern and multiple damage patterns corresponding to each key structural component. Next, Incremental Dynamic Analysis (IDA) is performed to assess the collapse capacity of the damaged structure under sequential motions. The collapse capacity gives safety capacity of the damaged building. Third, IDA is performed the collapse capacity of the intact building under sequential motions. This collapse capacity gives the safety capacity of the intact structure. A ratio of the collapse capacities of the intact and damaged buildings is compared to a threshold to decide on the building's safety. The author then makes use of Machine Learning algorithms like CART and Random Forests to build a model that uses these responses and damage patterns as predictor variables, and the corresponding collapse safety ratio compared with the threshold as the response variable, to predict the safety level of the building. A classification and Regression Tree was developed with an accuracy of 90% and Random Forests performed slightly well than CART [6].

While the application of Machine Learning for natural disaster damage prediction is still in its early phase, there are a couple of instances where ML algorithms have been used in similar but not exactly the same circumstances. For example, authors in [7] use MLA to classify damages caused by heavy rainfall. They propose two approaches wherein same-day weather data is used and historic weather data is used respectively. They compared the performance of bagging, boosting and random forests. However, since there is class imbalance the major class of 'no severe damage' is under sampled to match the number of instances of the class ' severe damage'. 10-fold cross validation was implemented. Meteorological data from the Korean Weather Open Data was collected for different areas in Korea. The Predictor variables included features like Temperature, Precipitation, Fog, Wind, Regional characteristics, etc. Approach 2 of using historical data instead of same-day data is picked on account of it being more practical. Boosting model is fit using historical data producing an accuracy of 95.8%.

Random Forests tended to over fit and also demonstrated a lower accuracy score. As a result Boosting was picked as a final model [7].

Another example of machine learning application in disaster damage prediction is proposed in [8]. Authors in [8] collected data from the NOAA Storm database with around 50 features. Because a lot of these features could be interacting and non linear to the response variable, artificial neural networks are used to model damage class [8]. The tornado data was combined with other data such as land cover of the tornado, socioeconomic data, etc. All the variables were normalised or standardised and some were further log transformed to deal with high variance. Data was portioned into training, validation and test sets. NNs are first used to obtain the best hyper parameters. Cross Validation set is used to pick the best model and test set is used to measure the mean squared error. Wide, deep and descending neural networks were tested and exhibited over fitting. To counter this dropout regularization was implemented. Wide models were also used to determine if a tornado would or would not cause damage as a binary classification problem [8]. With respect to conditional models, wide models with a MSE=0.0935 fared the best. With respect to binary occurrence models, a wide NN produced an accuracy of 87% on the test set [8].

Based on the literature, Random Forests, Neural Networks have been used often to predict earthquake damage with good results. I would like to illustrate and compare the performance of three models namely: RF, NN and Log Reg with the earthquake dataset. Also, studies conducted after the earthquake showed that some construction practices were more affected by the earthquake than others. I would like to perform some exploratory analysis and illustrate this through plots.

III. Proposed Approach

As proposed in the objective, model performances of Random Forests, ANNs and Regression will be compared.

The following steps will be carried out for each of the Machine Learning Algorithms:

1. Data Preparation –
   - Missing values – the missing values will either be removed or imputed.
   - Class imbalance – The data consist of three damage classes. If the distribution of these classes in the data is not uniform, Over-sampling will be performed.
   - Transforming the features to numerical form suitable for algorithms.
   - For categorical variables one-hot encoding will be done and for ordinal variables, dummy variables will be chosen.
   - Feature Engineering – There are 36 variables in the dataset. Some of them are categorical variables with high cardinality. As KNN algorithm is not likely to perform well high dimensionality, Chi-square test and Principal Component Analysis will be explored to select the most important and relevant features of the data.
   - Normalization – Once the variables have been chosen, they will be normalised to values between -1 and 1. This is important especially for neural networks as they perform better when the features have been standardised.
2. Model Generation –
   - Random Forests – The model is fine-tuned using Random Search Cross Validation [9]. A number of tree values (75,100, 200) will be compared.
   - Neural Networks – Neural network with three hidden layers and softmax activation function to deal with the three classes.
   - Regression – Multinomial Regression model will be tested with the best features.
3. Deployment – The final model will be submitted to the Competition.

IV. Dataset

The dataset consists of 38 features along with an index that represents the building number. There are three different kinds of features: 1) Features that contain the structural information of the building like age, no. Of floors, foundation type, roof type, material used in construction (for e.g. mud stone, cement, bamboo, timber, etc.). 2) The second type of features contain legal ownership information like geographical location of the building, who it is owned by and the purpose for which the building was being used (for example, whether it was a school, an office, police station, etc.). 3) The third type of feature contains the damage grade of the building after it was assessed. Most of the features are categorical. Only three features like age area percentage of the building, height percentage of the building are continuous. There are nearly 270000 training records and about 90000 test records.

V. Methodology

The first step after loading the data was to check for missing values. The data does not contain any missing values. Next, some

exploratory analysis was performed. I plotted data to check for class imbalance (figure 1). Also, another plot was

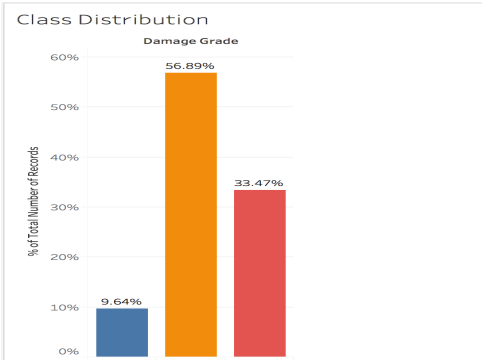


Figure 1 Class imbalance

generated to see how the foundation type of buildings was related with damage in each of the 31 districts. A third plot illustrating the damage with respect to land condition in each of the 31 districts was also plotted. The variables geo_code (30 factors), geo_code2(1500 factors) and geo_code3 (12000 factors) were left as numerical values initially because of their high cardinality. These represent the district, sub-region inside the district of the building. Only the feature representing the district was selected. Other categorical features were one -hot encoded, I used the chi-sq. test to identify the best 20 features out of the 36 features in the entire data. the test was able to rank the best features in the decreasing order of importance. I decided to pick 23 variables, normalized their values using the Scalar() function and over sampling the least damage and severely damage classes to match the moderately damage class. The model was then fit using these features with 100 trees. This gave a micro averaged F1 score of 0.94. This was indicative of the fact that the model was over-fitting.

One-hot encoding the district variable to improve was not a good approach, as it resulted in too many features. In addition, Principal Component Analysis did not perform well when compared to the chi-sq. test in reducing the number of features. Hence, chi-test was used for feature selection. However, to deal with high cardinality categorical variables, I decided to implement target encoding with Random Forests and Logistic Regression.

Target encoding basically replaces the categorical level with its average occurrence for the specific target class in that record. In addition, to prevent information leak and subsequent over-fitting, k-fold target encoding was performed for the categorical variables [10]. Here, the average occurrence of a categorical level from all the k-1 folds is calculated for its occurrence in the kth fold. Figure 1 provides a good description of this method. K-fold target encoding was used with both Random Forests and Logistic Regression.

Once encoded, the class distribution was analysed. Out of the three classes –low, medium and high, 57% of the records were moderately damaged, 10% were of low damage and 33 % of high damage. Under sampling and over sampling of class imbalance did not seem to do much for the model score. Hence, Synthetic Minority Over sampling technique, also called SMOTE, was adopted. This essentially generates synthetic examples of the minority class based on the nearest neighbour approach of the minority classes. Finally, the data was scaled using min-max normalization and fed to Random Forests and Logistic Regression.

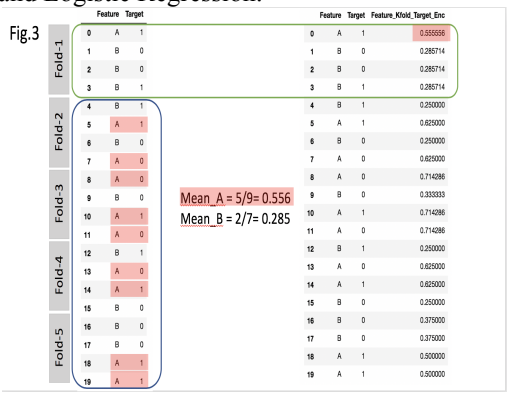


Figure 2 5 fold Target encoding [10]

With Neural networks, neither one-hot encoding nor target encoding seemed to generate good scores. This is because with one-hot encoding, the neural network is unable to learn that the columns with the different levels of the features are related. It fails to identify the underlying relationships of the one-hot encoded columns. Therefore, another categorical encoding technique called entity embedding was adopted. It creates a (m/d) vector for each categorical feature in the dataset [11]. 'm' is the number of levels in the categorical variable and 'd' can be 'm/2' to begin with. In the figure below, a categorical feature District with seven levels in considered.

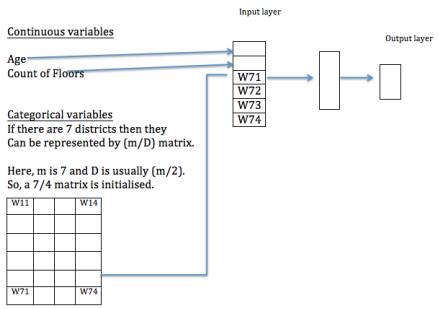
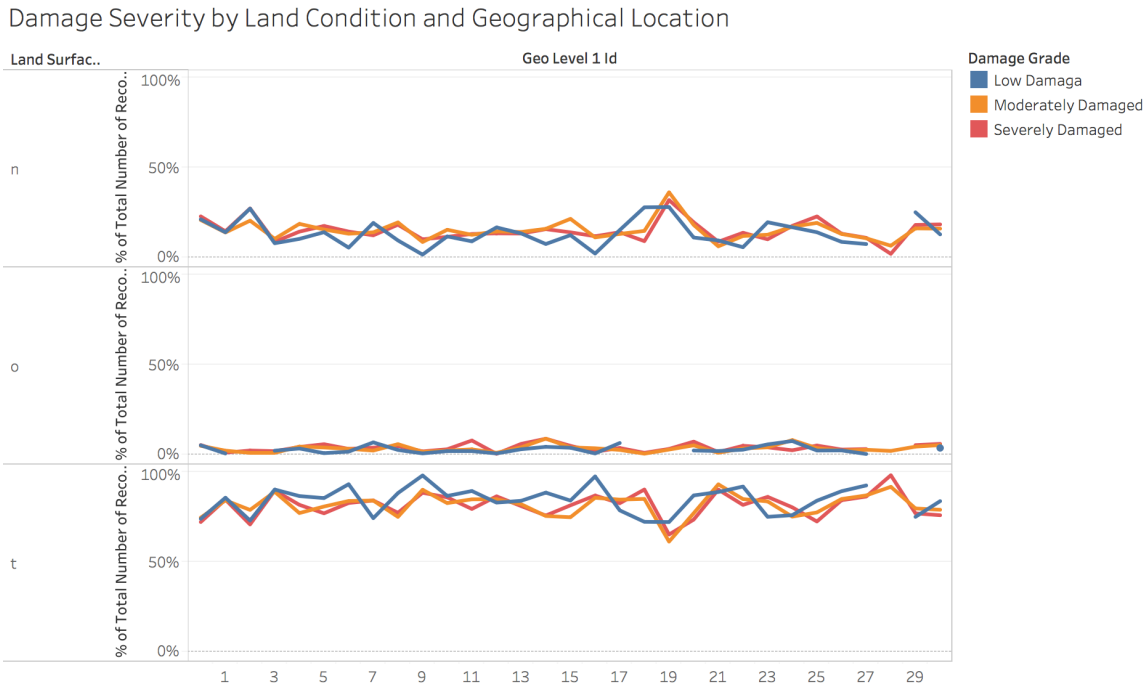
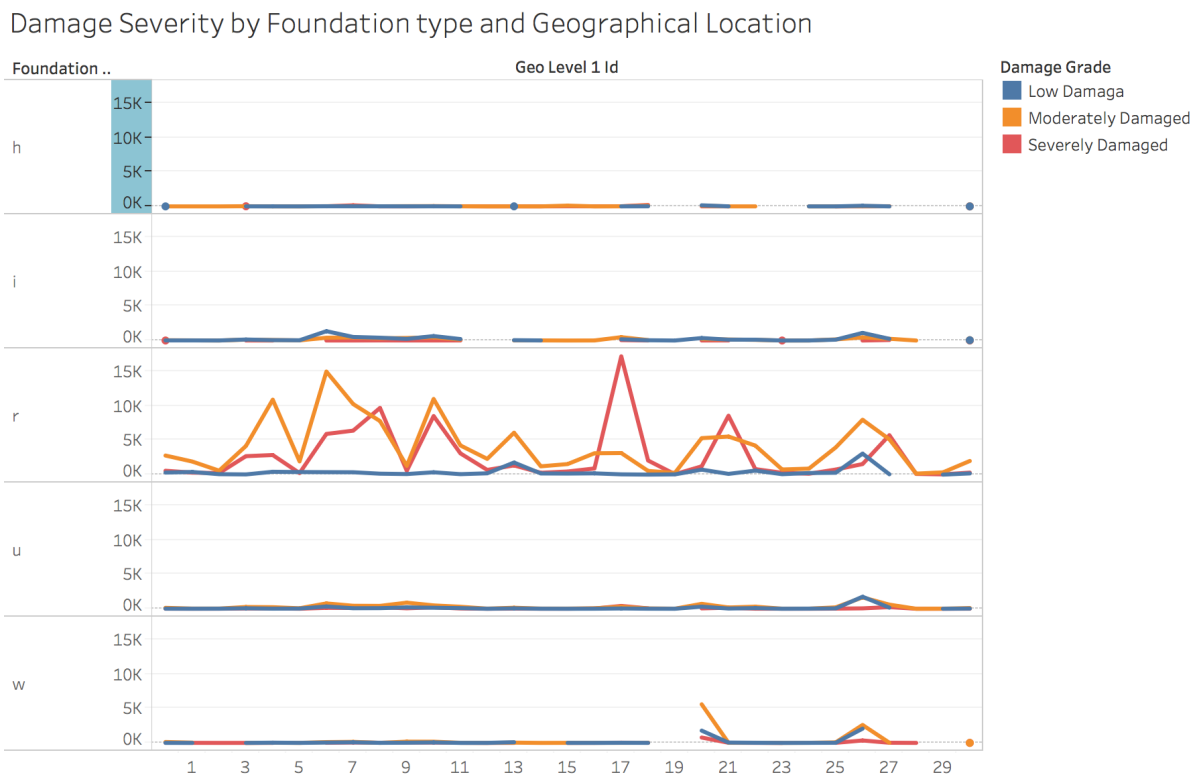


Figure 3 Entity Embedding



The trend of % of Total Number of Records for Geo Level 1 Id broken down by Land Surface Condition. Color shows details about Damage Grade.

Figure 4 Damage by land condition and district



The trend of sum of Number of Records for Geo Level 1 Id broken down by Foundation Type. Color shows details about Damage Grade.

Figure 5 Damage by foundation type and district

A corresponding 7/4 matrix is generated which is then concatenated with the other features in the dataset and fed to the neural network as shown in figure 3.

VI. Results & Conclusion

Exploratory analysis- A plot of the damage severity by land condition and district is shown in figure 4. It can be seen that 70% percentage of buildings with land condition 't' had some form of damage. It can be inferred that buildings built on certain land conditions are more impacted by the earthquake than other. Figure 5 shows that buildings with foundation type 'r' were more damaged than those built with other foundations. It can be concluded from this that the government can take measures to discourage people from building practices that are more vulnerable to earthquake damage. It can enforce a strict building code to reduce future damage.

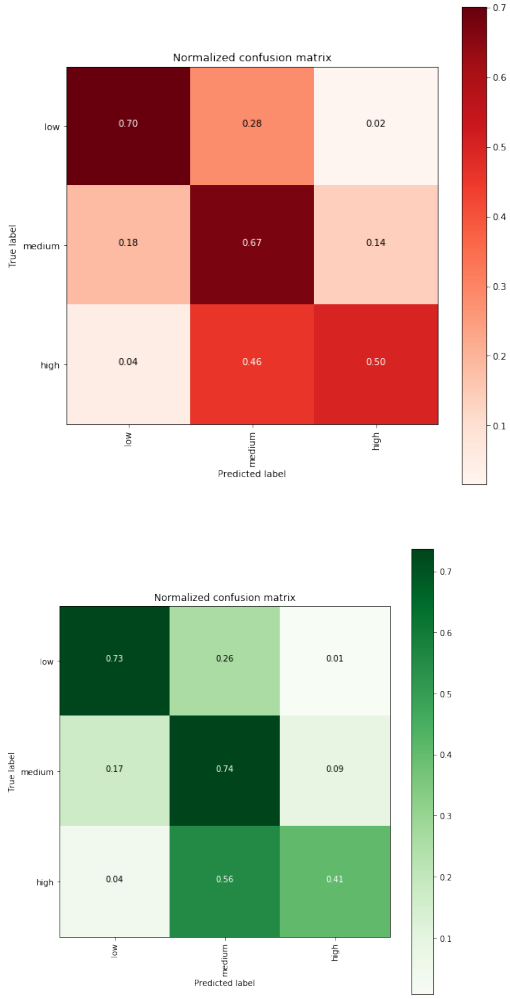
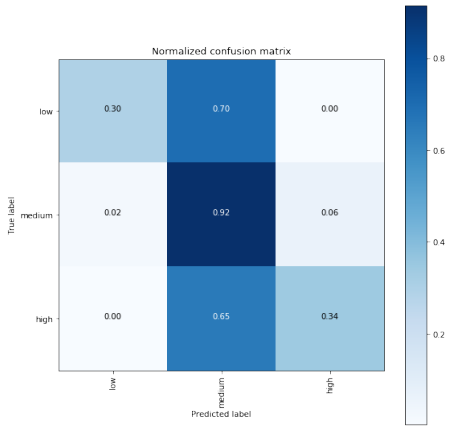


Figure 6 & 7 Confusion matrix for Random Forests and LogR

The performance metric chosen is micro averaged f1-score. This is used as it best presents the unbalanced classes.

The Final Random Forest and Logistic Regression models were trained using a combination of one hot encoding and target encoding. The best features were selected using chi-sq. test and SMOTE was implemented for class imbalance. Both models performed comparably with f1-scores of 0.6656 and 0.663 respectively. Figures 6 & 7 show the normalized confusion matrices of the above two models. It can be seen that RFs were slightly better than LogR in classifying the severely damaged buildings. On the other hand LogR was better at classifying the low and moderately damaged buildings.



Neural networks were trained by apply entity embedding and selecting all features in the final model. Softmax activation function was used and the model was penalized to deal with the imbalanced class weights. The f1-score was the highest at 0.67. Although, it can be seen from the confusion matrix that the model seems to be accurately classifying 92% of the moderately damaged buildings but seemed to do very poorly on the other two classes. This can be seen in figure 8. To conclude, I think Neural Networks can be our best although the class weights need to be further fine tuned for it to perform better with the other two classes as well.

VII. Future Work

A lot of studies have been using Satellite images of damage buildings to predict the damage grade [4]. The dataset could also include details like earthquake intensity and distance from the epicentre to improve the model performance. Also, models trained with light GBMS, LSTM and stacked GBMS have shown promising results [12].

REFERENCES

[1] [Online]. Available: https://www.drivendata.org/competitions/57/nepal-earthquake/

[2] Chen, Hao & Xie, Quancai & Li, Zhiqiang & Xue, Wen & Liu, Kang. (2016). Seismic Damage to Structures in the 2015 Nepal Earthquake Sequences. Journal of Earthquake Engineering. 10.1080/13632469.2016.1185055.

[3] Gautam, D., Rodrigues, H., Bhetwal, K.K et al. Innov. Infrastructure. Solut. (2016) 1:1

[4] Austin T.Cooner, Yang Shao & James B. Campbell (2016). Detection of Urban Damage Using Remote Sensing and Machine Learning Algorithms.

[5] Mangalathu, Sujith & Burton, Henry. (2018). Machine-learning based earthquake damage detection of residential buildings. 10.13140/RG.2.2.14516.71047.

[6] Zhang, Y. (2019). Post-Earthquake Performance Assessment and Decision-Making for Tall Buildings: Integrating Statistical Modeling, Machine Learning, Stochastic Simulation and Optimization. *UCLA*. ProQuest ID: Zhang_ucla_0031D_17995. Merritt ID: ark:/13030/m5mh2nk

[7] Changhyun Choi, Jeonghwan Kim, Jongsung Kim, Donghyun Kim, Younghye Bae, and Hung Soo Kim, "Development of Heavy Rain Damage Prediction Model Using Machine Learning Based on Big Data," Advances in Meteorology, vol. 2018, Article ID 5024930, 11 pages, 2018.

[8] Jeremy Diaz, Maxwell B. Joseph, "Predicting property damage from tornadoes with zero-inflated neural networks", Weather and Climate Extremes, Vol 25,2019,100216, ISSN 2212-0947.

[9] [Online]. Available: https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

[10] [Online]. Available: https://medium.com/@pouryaayria/k-fold-target-encoding-dfe9a594874b

[11] [Online].Available: https://towardsdatascience.com/decoded-entity-embeddings-of-categorical-variables-in-neural-networks-1d2468311635

[12] [Online].Available: https://github.com/arpan65/Earthquake-Damage-Modelling/blob/master/EQ%20Prediction%20Model.ipynb