**Project: Richter's Predictor: Modeling Earthquake Damage**

Web site: <https://www.drivendata.org/competitions/57/nepal-earthquake/>

**Introduction**

This project is based on the competition Driven Data published related to an earthquake [1] that took place at Nepal on April 2015 with a magnitude of 7.8 Mw according to a US Geological Survey [2], Its epicenter was east of Gorkha. It killed nearly 9,000 people and injured nearly 22,000[3].

Data has been collected on the damage suffered by various buildings, it had been identified the degree of damage on a scale from one to three, where three represents the greatest damage. The purpose of the project seems to be to develop predictive models to help identifying buildings prone to damage in other regions, in the event of another possible earthquake.

This document analyses the elements that eventually affect the degree of damage that a natural disaster, such as an earthquake, can cause. Then predictive models had been built in order to approximate a precise solution, which is measured with the performance metric required for the problem, finally, the conclusions of the study are raised.

The result of the project was uploaded in order to score the predictions that we got, which generated a response that is within 6% of the global participant results in this competition.

**Dataset description**

The data was collected by Kathmandu Living Labs and the Central Bureau of Statistics, using mobile technology for identifying the impact it had. The initial purpose was to identify beneficiaries for government assistance, but it also became helpful for studying the impact damage related to the household characteristics.

The data is composed of three files, train, test and submission files, in the case of the training file, it has 260,601 rows, and the test file has 86,868 rows. The description of every feature is in the Description Section of the web site [1]. It has 39 features and a target that is the damage grade the earthquake caused to buildings in Nepal

**Exploratory Data Analysis**

The exploratory data analysis was performed by analyzing the context of the problem and the characteristics of the variables as observed in the directory of the project, in this case “Exploration.ipynb” in the notebooks directory of the project.

Starting with the feature label corresponding to “damage\_grade”, in which it was observed that it has three possible values according to the degree of intensity of the destruction of the buildings in the place where the earthquake occurred, they are the following:

* + - 1 represents low damage
    - 2 represents a medium amount of damage
    - 3 represents almost complete destruction

It is observed that they are not balanced data, the number of occurrences of the first value of "1" is much lower than the other two remaining. In addition, it is worth mentioning that this feature is ordinal type and has a specific order associated with the degree of intensity of the earthquake.

There are 39 features that have been analyzed differentiating between numerical and categorical attributes. In the case of numerical features:

* count\_floors\_pre\_eq
* age
* area\_percentage
* height\_percentage

in which a evident bias was identified, practically in all these variables, which indicates the existence of outlier values that can harm the prediction process. In this case, processes were applied that reduce the intensity of this observed bias.

In the particular case of the feature count\_families, it was not transformed and was used to perform the Feature Engineering.

In the case of the features related to the geographical region, it was identified that these features from their concept are of the categorical type since they represent regions in Nepal.

* geolevel\_1\_id
* geolevel\_2\_id
* geolevel\_3\_id

The problem in this case was that for each feature there is a very high number of categorical values, so in this case it was decided to transform these features by applying the conditional probability related to the target, which generated nine new features.

The data set has the following features that are categorical, they describe, in general, types of materials used in the type of land constructions, and some of the position tasks, among others. All these encoded,

* land\_surface\_condition
* foundation\_type
* roof\_type
* ground\_floor\_type
* other\_floor\_type
* position
* plan\_configuration
* legal\_ownership\_status

What is observed is first that there is a coincidence between specific values of some of these features regarding the distribution in relation to the feature target. This indicates that there is a set of common characteristics in various buildings. In addition to the distribution in relation to the target, it is very similar in the values mentioned.

These following features are binary in the data, in this case it was determined to keep them as is, because they are binary valued:

* has\_superstructure\_adobe\_mud
* has\_superstructure\_mud\_mortar\_stone
* has\_superstructure\_stone\_flag
* has\_superstructure\_cement\_mortar\_stone
* has\_superstructure\_mud\_mortar\_brick
* has\_superstructure\_cement\_mortar\_brick
* has\_superstructure\_timber
* has\_superstructure\_adobe\_bamboo
* has\_superstructure\_rc\_non\_engineered
* has\_superstructure\_rc\_engineered
* has\_superstructure\_other

The same as the features before, the ones following are binary, in this case it was determined to keep them as is, with the exception of the feature “has\_secondary\_use” which was redundant and was removed.

* has\_secondary\_use\_agriculture
* has\_secondary\_use\_hotel
* has\_secondary\_use\_rental
* has\_secondary\_use\_institution
* has\_secondary\_use\_school
* has\_secondary\_use\_industry
* has\_secondary\_use\_health\_post
* has\_secondary\_use\_gov\_office
* has\_secondary\_use\_use\_police
* has\_secondary\_use\_other

It had been created a set of features, that were of great help to obtain higher prediction results, which are the following,

['CntFloorAge'] = ['count\_floors\_pre\_eq']/(['age']+0.1)

['CntFloorsArea'] = ['count\_floors\_pre\_eq']/['area\_percentage']

['CntFloorsHeight'] = ['count\_floors\_pre\_eq']/['height\_percentage']

['AreaPerAge'] = ['area\_percentage']/(['age']+0.1)

['HeightPerAge'] = ['height\_percentage']/(['age']+0.1)

['AreaPerHeight'] = ['area\_percentage']/['height\_percentage']

['CntFamFloors'] = ['count\_families']/['count\_floors\_pre\_eq']

['CntFamArea'] = ['count\_families']/['area\_percentage']

['CntFamHeight'] = ['count\_families']/['height\_percentage']

**Performance metrics**

To measure the performance of our algorithms, the site mention the F1 score which balances the precision and recall of a classifier. Traditionally, the F1 score is used to evaluate performance on a binary classifier, but since we have three possible labels we will use a variant called the micro averaged F1 score.

**Results**

This is the target variable for predicting the level of impact that earthquake caused, for housing reconstruction. The project had been built in python using vscode®, and submitted to Data Driven®, getting a result of 0.74444 that is among 6% of the ranking.

**Bibliography**

[1] Driven Data. (2018). Richter's Predictor: Modeling Earthquake Damage. 01-01-2018, de Driven Data Sitio web:

<https://www.drivendata.org/competitions/57/nepal-earthquake/page/134/>

[2] USGS. (2018). M 7.8 - 36km E of Khudi, Nepal. 01-01-2018, de USGS Sitio web: <https://earthquake.usgs.gov/earthquakes/eventpage/us20002926/executive>

[3] Wikipedia. (2020). April 2015 Nepal earthquake. 01-01-2015, de Wikipedia Sitio web: <https://en.wikipedia.org/wiki/April_2015_Nepal_earthquake#:~:text=The%20April%202015%20Nepal%20earthquake,Intensity%20of%20VIII%20(Severe)>

.