Richter's Predictor: Modeling Earthquake Damage*

First Author[1][FrancescoSalerno], Second Author[2,3][EdoardoNardi], Third Author[2,3][MalickJobe], and Fourth Author[3][AndreiSauca]

[1] University of Pisa
[2] Master's degree in Data sciecne and Business Informatics

Abstract. The aim of this report is to highlights the important criteria from the Richter's Predictor: Modeling Earthquake Damage dataset, that will be used in order to try to predict the damage level of an earthquake.

1 Data Understanding

1.1 Data Semantics

We want to emphasize that we are still at the first step. In fact the use of the variables involved in the model creation might change during studying.

At a first look we can see that the Dataset is built by 260601 row and 39 columns, all of them appear as not null.

To study better variables information, statistics and their behaviour we decide to understand to which type every variable belongs. and we identify three different variable types **Qualitative - Discrete**, **Quantitative - Binary** and **Quantitative - Nominal**.

Qualitative - Discrete Into our data-set we identify 9 variable belonging to the Quantitative - Nominal type, in order to check if the semantic domain does not differ from the syntactic one we calculate from the data-set maximum and minimum value that every single variable can assume. All this computation is summarized in the Table 1 below.

* University of Pisa

col_name	attribute_type	domain	domain_size
building_id	Qualitative - Discrete	[4 - 1052934]	260601
geo_level_1_id	Qualitative - Discrete	[0 - 30]	31
geo_level_2_id	Qualitative - Discrete	[0 - 1427]	1414
geo_level_3_id	Qualitative - Discrete	[0 - 12567]	11595
count_floors_pre_eq	Qualitative - Discrete	[1, 2, 3, 4, 5, 6, 7, 8, 9]	9
age	Qualitative - Discrete	[0 - 995]	42
area_percentage	Qualitative - Discrete	[1 - 100]	84
height_percentage	Qualitative - Discrete	[2 - 32]	27
count_families	Qualitative - Discrete	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]	10

Table 1: Qualitative - Discrete variables

Regarding this 9 variables the syntactical and semantic domain coincide, we assuming that semantic domain when not specified coincide with any syntactical one.

The Qualitative - Discrete variables that was significant for usat this unsupervised learning phase are **count_floors_pre_eq**, **age**, **count_families**, **height_percentage**. For these variables we provide some statistical information in the Figure 1 below.

	count_floors_pre_eq	age	count_families	height_percentage
count	260601.000000	260601.000000	260601.000000	260601.000000
mean	2.129723	26.535029	0.983949	5.434365
std	0.727665	73.565937	0.418389	1.918418
min	1.000000	0.000000	0.000000	2.000000
25%	2.000000	10.000000	1.000000	4.000000
50%	2.000000	15.000000	1.000000	5.000000
75%	2.000000	30.000000	1.000000	6.000000
max	9.000000	995.000000	9.000000	32.000000

Fig. 1: Statistical Values Description

Taking for example the **age** variable we can immediatly notice that has a very high variance, this is immediatly explained to the fact that the higher value 995 is very far from the mean, the same for the nth percentiles.

We can see better this fact exploring the Figure 2 that show with a boxplot how far is the 995 value from the others.

We have more than one thousand (1390) record assuming this value. The higher close to this amount is two hundred. This let us thing the 995 value could be a null value or anyway it is related to building which couldn't be classified or aged. It seems to be a good explanation for us. In fact we decided in the following task to replace it with mean.

Others important feature we discovered was an higher correlation between **count_floors_pre_eq** and **height_percentage** variables, we try to combine

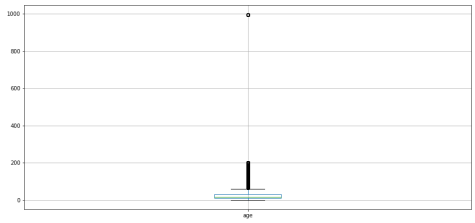

Fig. 2: Age boxplot

them into a single ratio, but when we passed to the supervised phase we notice that this kind of computation does not help.

In this phase we take in account also a possible interaction between **area_percentage** and **count_families** but we weren't be able to find some type of correlation.

Other security check was performed to **geo_level_1_id**, **geo_level_2_id** and **geo_level_3_id** to verify that every lowest level belongs to at most one single father level.

All the variable belonging to this group have been plotted using bar chart plot types.

Quantitative - Nominal Also for Quantitative - Nominal we discover that the syntactical domain, that you can see from Table 3, are the same that the semantic ones.

level_0	index	col_name	attribute_type	domain	domain_size
8	8	land_surface_condition	Quantitative - Nominal	[n, o, t]	3
9	9	foundation_type	Quantitative - Nominal	[h, i, r, u, w]	5
10	10	roof_type	Quantitative - Nominal	[n, q, x]	3
11	11	ground_floor_type	Quantitative - Nominal	[f, m, v, x, z]	5
12	12	other_floor_type	Quantitative - Nominal	[j, q, s, x]	4
13	13	position	Quantitative - Nominal	[j, o, s, t]	4
14	14	plan_configuration	Quantitative - Nominal	[a, c, d, f, m, n, o, q, s, u]	10
26	26	legal_ownership_status	Quantitative - Nominal	[a, r, v, w]	4

Table 2: Quantitative - Nominal variables

To plot these variable we mostly used pie chart plot, and in case they have too long domain we decide or use bar chart plot or to subdivide the plot in two separate plot, the first with the main class and "Others" and the second with all the others classes.

This kind of subdivision can be seen from Figure: 3.

We don't discover any important significant behaviour to analyze for Quantitative - Nominal variables.

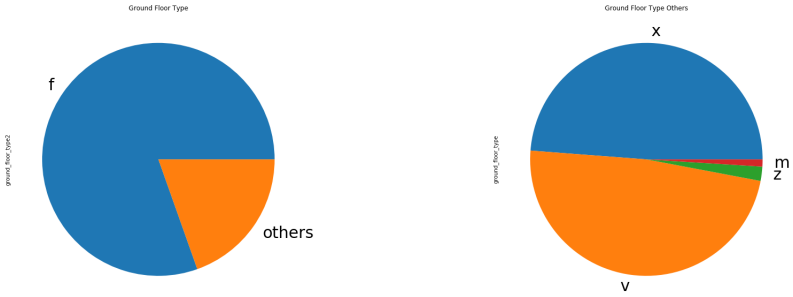

Fig. 3: Ground floor Type

Quantitative - Binary Also for Quantitative - Binary variable we perform a domain check, this is quite trivial since the variables has been labeled as binary due to the fact that they can assume only 0 an 1 values. Below a in Table 3 you can see all the variable that belong to Quantitative - Binary type.

	col_name	attribute_type	domain	domain_size
15	has_superstructure_adobe_mud	Quantitative - Binary	[0, 1]	2
16	has_superstructure_mud_mortar_stone	Quantitative - Binary	[0, 1]	2
17	has_superstructure_stone_flag	Quantitative - Binary	[0, 1]	2
18	has_superstructure_cement_mortar_stone	Quantitative - Binary	[0, 1]	2
19	has_superstructure_mud_mortar_brick	Quantitative - Binary	[0, 1]	2
20	has_superstructure_cement_mortar_brick	Quantitative - Binary	[0, 1]	2
21	has_superstructure_timber	Quantitative - Binary	[0, 1]	2
22	has_superstructure_bamboo	Quantitative - Binary	[0, 1]	2
23	has_superstructure_rc_non_engineered	Quantitative - Binary	[0, 1]	2
24	has_superstructure_rc_engineered	Quantitative - Binary	[0, 1]	2
25	has_superstructure_other	Quantitative - Binary	[0, 1]	2
28	has_secondary_use	Quantitative - Binary	[0, 1]	2
29	has_secondary_use_agriculture	Quantitative - Binary	[0, 1]	2
30	has_secondary_use_hotel	Quantitative - Binary	[0, 1]	2
31	has_secondary_use_rental	Quantitative - Binary	[0, 1]	2
32	has_secondary_use_institution	Quantitative - Binary	[0, 1]	2
33	has_secondary_use_school	Quantitative - Binary	[0, 1]	2
34	has_secondary_use_industry	Quantitative - Binary	[0, 1]	2
35	has_secondary_use_health_post	Quantitative - Binary	[0, 1]	2
36	has_secondary_use_gov_office	Quantitative - Binary	[0, 1]	2
37	has_secondary_use_use_police	Quantitative - Binary	[0, 1]	2
38	has_secondary_use_other	Quantitative - Binary	[0, 1]	2

Table 3: Quantitative - Binary variables

For binary variables statistical information are less significantly, as for nominal ones, so we decide to plot them as pie chart.

As you can see from 3 all the variable in this group belongs to two main group **has_superstructure** and **has_secondary_use**

Our idea was to try to combine them in a single variable to decrease the number of considered variable, as you will see we reach the goal for **has_secondary_use** variables but not for **has_superstructure**.

Another important consideration we find analyzing these variables was that all variables that has more then one secondary uses have exactly 2 secondary uses and the second is **has_secondary_use_other**.

1.2 Correlation with target variable damage_grade

The predictable variable **damage_grade** is the aim of the research this represent the damage grade caused to a building and it can assume 3 different values low, medium and high.

Here we are going to show the correlations we found from variable and target.

Before do that we need to exclude the trivial result that building close to the center of the earthquake have been damaged more then others.

We discover from correlation matrix that the **geo_level_1_id**1 is the most correlated one for geographical level variables, so we start analyze, region per region which is the mean of damage grade, and, as you can see from Figure 4, the mean is quite similar from a region to each others.

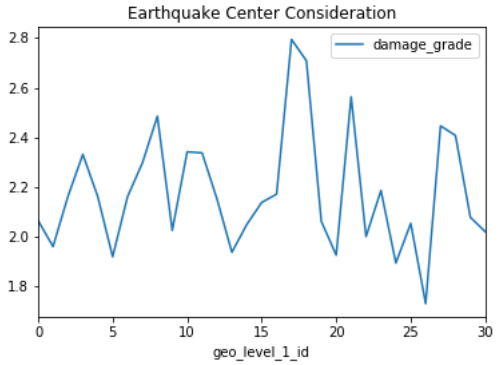

Fig. 4: Earthquake mean by Geo level 1

So we feel safe to exclude Geo levels from this correlation phase.

Before starting seeing the correlation matrix with target table we tried to transform all the consideration taken into account in the Data Semantic task into a reliable result.

First we work on **count_floors_pre_eq** and **height_percentage** correlation and we tried to create a new variable **floor_height_ratio** as the ratio between these two variable.

This does not help because we passed from 0.076 for **height_percentage** with **damage_grade** to 0.016 for **floor_height_ratio** with **damage_grade**.

Quite the same situation trying to zip all the **has_superstructure** variables in a single one, we passed from a correlation between **has_superstructure_adobe_mud** and **damage_grade** of 0.29 to a correlation of 0.086, very bad result.

We perform the same for **has_secondary_use** variables and we obtain a good result passing from 0.098 correlation of **has_secondary_use_hotel** with **damage_grade** to 0.1 with the zipped variable **secondary_use**. This is not a good result itself, only 0.02 more correlation, but is a good result in term of decrease feature number to correlate with **damage_grade**

Once we finished the transformation variable task we cant take a look to the final correlation matrix in Figure 5.

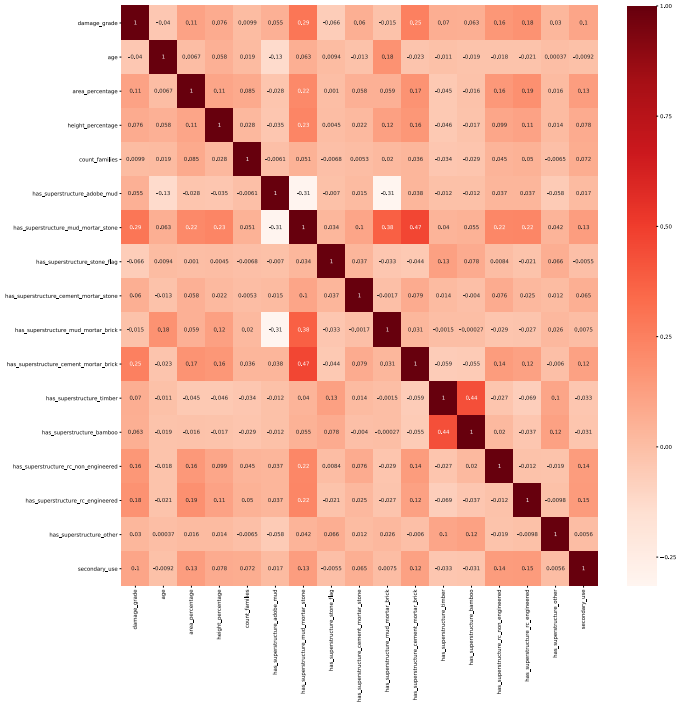

Fig. 5: Correlation Matrix

From this matrix we can conclude that all the variables that has a correlation value bigger then the threshold of 0.1 will be chosen for next model construction part that will be taken in consideration in the following tasks.

So for the feature selection part the following variables
area_percentage (0.11),
has_superstructure_mud_mortar_stone (0.29),
has_superstructure_cement_mortar_brick (0.25),
has_superstructure_rc_non_engineered (0.16),

, **has_superstructure_rc_engineered** (0.18)
secondary_use (0.1).