Machine Learning Mini Project: Earthquakes from Turkey: A Beginner's Prediction With AutoML

Esra Kantarcı

February 5, 2021

1 Introduction

The dataset from Kaggle consists of the earthquakes from Turkey between the years 1910 and 2017. The columns are date, time, lat (latitude), long (longitude), country (for the center), city, area, direction (of seismic waves), dist (distance), xm (biggest value of md,mw,ms and mb), md(mean magnitude of the duration), richter, mw(moment magnitude), ms(surface wave), mb(body wave).

This dataset from Kaggle is not really a good one to evaulate, since when you take a look at the distribution between the years, you will see that the 1980-2000 part of the data is actually missing. The clever step could be using http://www.koeri.boun.edu.tr/sismo/2/earthquake-catalog/ to get a new dataset file to process or to use World Earthquake Dataset and filter the values by latitude and longtitude columns. However, since I want to submit this to Kaggle after this much effort, I want to use this dataset (and remember that our data is actually biased while processing).

Personally, I picked this dataset, because I am currently concerned with the anxiety of earthquakes. Since I am living in an apartment with 11 floors and it was built before 1999, I am a little bit alarmed. So I wanted to check out the fault line in Mediterranean Sea when I saw the dataset in Kaggle. Sadly, after the quick evaluation, I wanted to move out from my apartment as soon as possible, because Aegean-Mediterranean part is active and the depth in Antalya's fault-line is high: which you will see from the correlation chart, it has positive correlation between magnitudes and depths. As you can search from the Internet, bigger depthed fault lines actually points the supra-subduction zones, which are highly tectonic and due to subduction of the earth-plates that creates Earth's continental crust.

2 Tools Used

- ML and Visualization Libraries: Auto-sklearn, Basemap, Seaborn, Sklearn - RandomForestRegressor

- Kaggle Dataset: Earthquakes in 1910-2017, Turkey
  https://www.kaggle.com/caganseval/earthquake
- Google Colab: You can find out the full .ipnyb version with outputs in attachment.

3 Data Preprocessing

There are a big part of null values in "city", "area", "direction", "distance" and "mw" values in the dataset. We do not need city-area-direction since we have better and more specific columns such as longitudes and latitudes. Therefore, we can always drop these values without any issue. But "mw" is actually important. As you can see from the correlation table, in general, "xm" has correlation with "mw". So, for the further investigations I filled NaN mw values with corresponding xm values, which is a better approach than using means for preprocessing part.

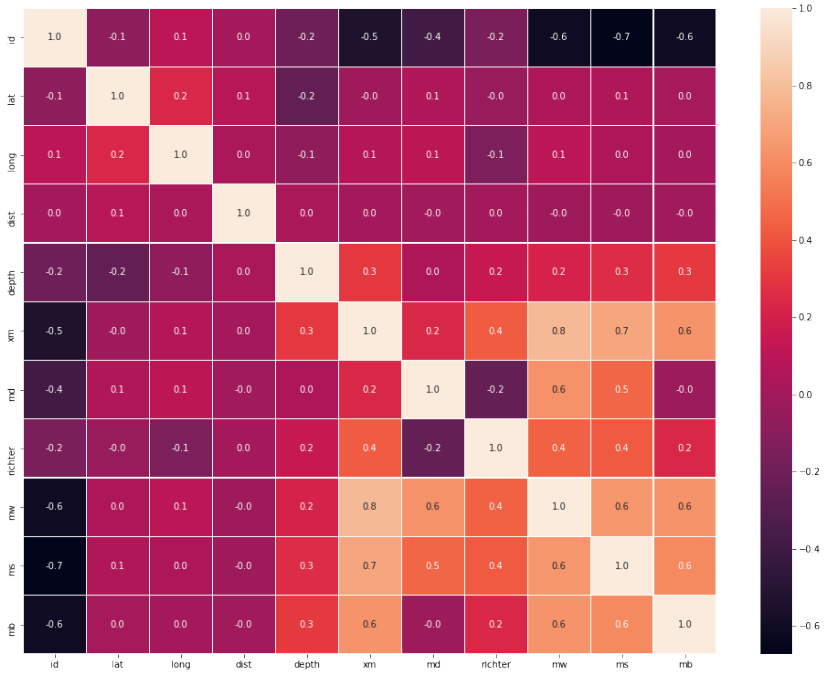

Figure 1: Correlation Heatmap for Integer Valued Columns

From the researches, mw and md values are the ones which are the most important during the earthquake analysis. So, I decided to use these values for y(output).

Also, the “date” and “time” columns are out of format. Therefore in preprocessing part, we also need to parse the values in datetime and integer versions. Therefore, I had used userdefined function to parse “date” column into “yeardate”, “monthdate”, “daydate” and the combination of “date”-“time” as datetime type object in “period” column, for future usages.

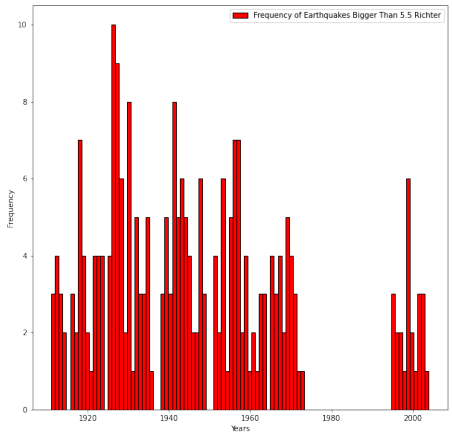

Figure 2: Missing Data from 1970 to 2000

```
df.isna().sum()
```

id	0
date	0
time	0
lat	0
long	0
country	0
city	12253
area	11030
direction	13945
dist	13945
depth	0
xm	0
md	0
richter	0
mw	19004
ms	0
mb	0

dtype: int64

Figure 3: Missing Values in Columns

I did not use normalization methods, because the values were handy as they were, and it was good for the readability purposes. However, if I wanted to make analysis from city or direction, I would be using label encoding for utilities.

And before the training model, I decided to use longitude, latitude and year values for input and ms and depth values for the output. It is because longitude and latitude covers most of the values in the data and ms is more reliable than the mw values in this dataset. I added depth because in different fault lines which are in the same area, there are different kind of magnitude-generation-potentials. Using yeardate for the input is also a little problem, the period value could be even better to evaluate, but in the automl results the R2 score with period was too weak. The best predictions were done based on these column values.

Also, using xm could be a better choice for output, since it gets the biggest magnitude of all. However, there is a problem: Since it depends on other column values as well, which ends up after the earthquake itself, it was just going to increase complexity of model. But I did try out the xm for output, R2 score was again below 0.3, which is considered weak.

Then I realized, even if there is no NaN value in md column, there are many 0.0 values which was not calculated due duration being too short. These could affect the model, therefore I changed 0.0 values into 0.01, which also increased the R2 score eventually.

Also, after plotting the earthquakes by md and year, I realized that the earthquakes between 1970-2000 are also missing, as you can see from the figures

above.

In order to split data to test-train, I used sklearn library and splitted by 0.2 factor, shuffled as a last step of data preprocessing before the training.

4 Methodology

First of all, using Seaborn library, I visualized the correlation table to check the columns and their effects. As explained before, because of the reasons md column will be chosen as output(y) value. For the further visualization purposes to find out the fault lines with big md values, I used BaseMap library.

After plotting the earthquakes by year, I had found out that earthquakes between 1970-2000 are actually not available in the dataset as you can see in Figure-2. After using the dataprocessing techniques explained above, Random-Forest model was used. In this kind of datasets, Random Forest approach is acceptable due to its regression capabilities and accuracy due to cross validation handling between trees.

As soon as the regression by random fores algorithm's score turned out low(0.57), I wanted to check out auto-ml solutions for model selection and prediction. Surprassing the expectations, on auto selected model on latitudes-longitudes-year values as inputs and md values as outputs, R2 score was higher than 0.83.

Then I tried something different, which are not in this version of code segment, because it was not useful. But I wanted to state why this did not work out:

In 2020 we encountered big and destructive earthquakes:

- Elazığ Earthquake: 24 Ocak 2020, 6.5 -Latitude: 38.3593 N, Longitude: 39.0630 E
- İzmir Earthquake: 30 Ekim 2020. 6.9 -Latitude: 37.8881 N. Longitude: 26.7770 E

Using these values as test values we end up with an array with 0.0099 and 0.0099 outputs, which means they were not predicted, the probability of earthquake in 2020 simply do not exist and even if it exists, the value is lower than 0.001 which is even lower than the minimum value of md. This is because we do not have data for year 2020 to train our model. Also, for future prediction purposes, we could generate the values from previous set.

5 Experiment Results

5.1 Overview

5.1.1 Earthquake distribution by magnitudes:

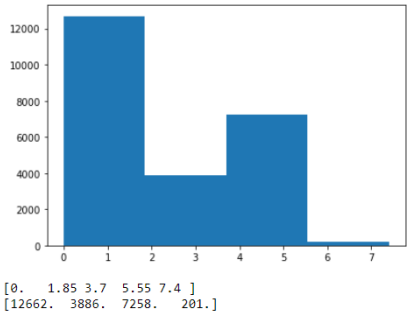

Figure 4: Earthquake Distribution by Magnitudes

- 12662 earthquakes were between 0 and 1,85 in magnitude
- 3886 were between a 1,85 and 3,7
- 7258 earthquakes between a 3,7 and a 5,55
- 201 earthquakes were greater than 5.55

5.1.2 The cities encountered the earthquakes magnitude higher than 6.0 richter

Alphabetically sorted 20 unique cities: Adana, Afyonkarahisar, Antalya, Aydın, Balıkesir, Bartın, Bingöl, Burdur, Çanakkale, Çankırı, Denizli, Düzce, Erzincan, Erzurum, Eskişehir, Gazimagusa, Hakkari, İstanbul, İzmir, Karabük, Kırşehir, Kocaeli, Muğla, Muş, Sakarya, Tekirdağ, Tokat, Tunceli

5.1.3 Correlations

- The values in xm depends on md, mb, mw and ms (which is pretty obvious by definition), but it also has 0.3 positive correlation with depth. This means the magnitude of the earthquake will be higher if the depth is bigger.
- Depth values have negative correlation with latitudes and longitudes, this means when we go to east or south the depth of the fault line will be bigger.
- Distance values almost do not have any dependency of any other values.

5.1.4 Fault-Line Visualization

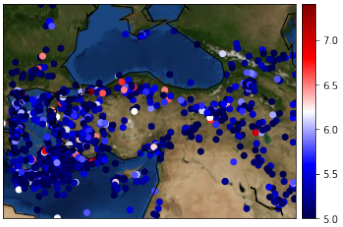

Figure 5: Earthquakes Bigger than 5.0 Richter Magnitude

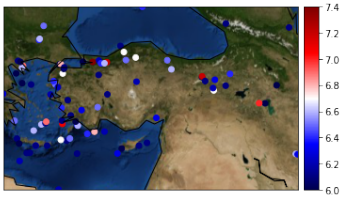

Figure 6: Earthquakes Bigger than 6.0 Richter Magnitude

- As it can be interpreted from the map visualization, the areas with bigger magnitudes are on North Anatolian Fault Line and gets bigger magnitudes as it goes to East part of the fault line.
- East Anatolian Fault Line and Hellenic Arc also actively encounter big magnitude earthquakes frequently.

5.2 About Features

- Latitudes and Longitudes are the ones to evaluate and predict the earthquakes. They are also the values which area, city, country like columns are dependent.
- Fault-lines and the earthquake activities are frequently located at the same areas with previously collected data, which indicates latitudes and longitudes are also good features to evaluate.
- Depth and magnitude values are related to each other. But using them at the same auto-ml model lowers the R2 score.
- Using ms, mb, xm values for output element ends up with weak R2 values.

5.3 About Model Selection

- At the last step of data preprocessing, the train-test values from the dataset was splitted using sklearn split method by taking randomly 0.2 of the initial data as test.
- Since we need to use regression, for developed decision tree-like algorithm, Random Forest Algorithm was selected. It ends up with low score.
- Using Auto-ML from scikit library, Extra-Trees, AdaBoost and Gradient-Boosting models are trained. Automatically tuned the model and R2 value of the predictions were above 0.8, which is pretty strong.

5.4 About Predictions and Score

- As you can see, Random Forest got 0.57 score where Auto-ML reached 0.83 R2 value.

Figure 7: AutoML R2 Value

- Predictions have 0.70 MSE compared to actual set.
- Predictions have 0.84 RMSE compared to actual set.
- Cannot predict future values, due to lack of data (since it needs inputs from the year column)
- Using depth as feature lowers the prediction score
- Using depth as output value lowers the prediction score
- Using mb, mw or xm as output value lowers the prediction score
- Using md as output value ends up with highest prediction score
- Filling the md values of 0.0 with 0.01 increases the prediction score

6 How to Improve?

- We can generate the similar data for future timestamps and try to fit model again using these values, so we can get future-time predictions.
- We can use different datasets and join them, because this dataset lacks many important values, such as 1999 Düzce earthquake and the timestamp is not dependable.
- We can use Bogazici University Kandilli Observatory's Earthquake catalog to get newer dataset.
- For further evaluation, we can use "period" column to check if there is a similar behaviour on the same area before the big magnitude earthquake happens and we can use prediction model with this approach.

7 Conclusion

In conclusion, Turkey is the active fault-line area and should be ready for earthquakes bigger than 5.5 richter magnitudes. The most dangerous ones are in the North-Anatolian Fault Line and East-Anatolian Fault Line. The most active region which encounters 5.5 and bigger magnitudes are in Hellenic Arc area: Mediterranean and Aegean part of Turkey.

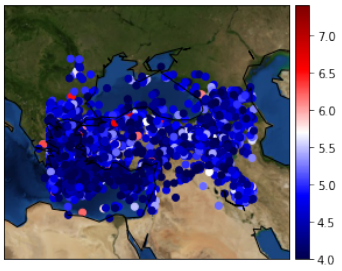

Figure 8: All Earthquakes in Dataset, Indicates that Turkey is in Earthquake Zone

	prediction	actual	difference
0	0.013979	0.01	0.003979
1	0.010843	0.01	0.000843
2	4.744133	4.80	0.055867
3	4.247197	4.00	0.247197
4	1.049433	0.01	1.039433
...	...	...	...
5997	0.011415	0.01	0.001415
5998	2.829980	0.01	2.819980
5999	3.836044	3.80	0.036044
6000	0.404014	0.01	0.394014
6001	4.811301	5.80	0.988699

Figure 9: Predictions Using Auto-Sklearn Model

Bigger depth of earthquake often indicates bigger magnitudes. The best features to use in our model was year-latitude-longitude combination.

Regression RandomForest Algorithm did not result good. However, using Auto-Sklearn library and its AutoML utilities (Extra-Trees, AdaBoost and Gradient-Boosting models combined), we end up with R2 score higher than 0.8 which indicates strong effect.

References

[1] Kaggle - LANL Earthquake EDA and Prediction
https://www.kaggle.com/gpreda/lanl-earthquake-eda-and-prediction

[2] Kaggle - Earthquake Data Distribution Of Turkey
https://www.kaggle.com/caganseval/earthquake-data-distribution-of-turkey

[3] Kaggle - Earthquakes in Turkey (1910 - 2017) - Basic EDA
https://www.kaggle.com/berkesun/earthquakes-in-turkey-1910-2017-basic-eda

[4] Kaggle - Earthquake Prediction - Data from Significant Earthquakes, 1965-2016
https://www.kaggle.com/mahadevmm9/earthquake-prediction

[5] Kaggle - Earthquake Analysis on World Map
https://www.kaggle.com/mriduleecs/earthquake-analysis-on-world-map

[6] Kaggle - Quick Charts on Earthquakes
https://www.kaggle.com/ddmngml/quick-charts-on-earthquakes

[7] KDNuggets - Easy Quide to Data Preprocessing in Python
https://www.kdnuggets.com/2020/07/easy-guide-data-preprocessing-python.html

[8] Data Flair Training - Data Preprocessing, Analysis and Visualization
https://data-flair.training/blogs/python-ml-data-preprocessing/

[9] Jim Frost - How to Interpret R-Squared in Regression Analysis
https://statisticsbyjim.com/regression/interpret-r-squared-regression/