A machine learning approach for predicting volcanic eruptions, one segment at a time

Radboud University

Sebastiaan Ram (S1063000)
Steffen de Jong (S1065975)
NWI-IBI008
Project type: "Problem"
GitHub project: https://github.com/PeaceDucko/dm-volcanic-eruption-project

11th January, 2021

Abstract
Currently, volcanic eruptions are 'predicted' by analyzing tremors from seismic signals. Unfortunately, potential patterns are hard to interpret. This paper presents a take on predicting volcanic eruptions with the help of machine learning. Different types of regression models were trained on a data set provided by the INGV, with the goal of finding a regression model with a minimal mean absolute error based on the time to eruption. The data set originally contained roughly 5.4 billion readings, which were reduced to one million using data and dimensionality reduction techniques and algorithms, like data aggregation. Fitting models on this reduced data set resulted in the optimal model being a K-Nearest Neighbor Regressor with a mean absolute error of $3.973 \cdot 10^6$, showing that a volcanic eruption can be predicted by the model with an average error of 21 hours. These results are far from perfect, but do give an incentive to further improve on the techniques and methods applied in this study. However, the results do indicate an increase in efficiency and reliability of data-driven models opposed to manual predictions. Showing that, despite some limitations in our approach, further research is worth exploring.

Contents

1 Introduction

Volcanic eruptions belong amongst one the most frequent natural disasters, any unforeseen eruption can result in tens of thousands of lives lost. Scientists currently 'predict' the time to eruption by surveying volcanic tremors from seismic signals. Unfortunately, observed seismic patterns are hard to interpret.

One way to improve the quality of these interpretations is with the help of Artificial Intelligence and satellites, as showed by Witze [2019], or by taking additional factors into account, like elastic-brittle crust or the amount of energy loss (Kilburn, 2018). However, these approaches are expensive and little attention has been paid to improving the quality of the already acquired signals.

This paper presents a take on interpreting basic seismic signals with solely the use of machine learning. It is expected that the time to eruption can be accurately predicted using regression algorithms. The paper describes the resources used to prepare the environment and what techniques and algorithms were implemented. The combinations of several of these implementations form a solid foundation for current and future seismic prediction.

2 Methods

The current experiment is based on the data set provided by Italy's Istituto Nazionale di Geofisica e Vulcanologia (INGV). The data consist of approximately 4.400 training and 4.500 testing segments, each in a separate file. Furthermore, each segment contains 60.001 normalized readings from 10 different sensors over a 10-minute time span. Additionally, segments belonging to the train set include their recorded time to eruption, which served as the key feature to be predicted (National Institute of Geophysics and Volcanology, 2020). The time to eruption was measured in 1/100th of a second.

As a result of the immense size of the data set, data and dimensionality reduction was necessary. The experiment was performed using Python 3.8 running on an Anaconda distribution. Different reduction methods like Principal Component Analysis, data aggregation and Continuous Wavelet Transformation, as described by Lapinsa [2020] and Feike et al. [2020], were explored. It was quickly found that only 10.000 of the 60.001 sensor readings were enough to accurately replicate each segment's seismic activity, resulting in a great reduction of the size of the data set and computation time for further aggregation.

Furthermore, sensor data from each segment was aggregated to form the basis on which the model was trained. Dimensionality reduction was applied using the sum, mean, standard deviation, minimum, maximum and 10th, 25th, 50th, 75th and 90th percentile on both the original and absolute data over each of the sensors, as proposed by Isaienkov [2020]. This approach resulted in a reduction from 5.4 billion to only one million readings.

The time to eruption can be categorized as a continuous variable, thus as a regression problem. As a result, both linear and non-linear models were fitted, including Linear Regression, Random Forest and K-Nearest Neighbor. Additionally, 10-Fold cross validation was applied to each model to detect overfitting and use the available segments efficiently (Bronshtein, 2017).

Finally, 10-Fold Grid Search was applied to a K-Nearest Neighbor Regressor to optimize model performance. Amongst the optimized hyperparameters, the number of neighbors, metric and weights were chosen. Furthermore, the number of neighbors was based on the first eight numbers of the Fibonacci sequence, starting at 1. No Grid Search was applied to either Random Forest or Linear Regression, mainly because results of these models were not significant enough to keep exploring. The best performing model was eventually trained on all the data.

3 Results

In our study, the time to eruption of various volcanic segments was analysed and interpreted using various regression-based machine learning algorithms like Linear Regression, Random Forest and K-Nearest Neighbor. Also, data and dimensionality reduction was necessary to obtain meaningful results, it was found that only 1/6th of the data points were needed to accurately represent the seismic activity, as can been seen in Figure 1.

Furthermore, aggregation was applied using various statistical measurements over the sensor data to achieve dimensionality reduction, this resulted in a data set of 4431 rows and 110 columns. In addition, Principal Component Analysis was added to the aggregated data set, though this approach did not yield any benefits.

First, Linear Regression was attempted to be fitted, but failed because no immediate linear correlations in both the data from the original and absolute data set were present, as can be seen in Figure 2.

Random Forest gave hopeful results on the original, aggregated data, but performed better on the absolute data with an R2 score of 0.682 and a mean absolute error of $5.46 \cdot 10^6$ with a max depth of 30 using 10-Fold cross validation (Fig. 3). However, K-Nearest Neighbor proved to be more efficient on absolute values, reaching an R2 score of 0.677 and a MAE of $3.97 \cdot 10^6$ (Fig. 4), a significant improvement in the MAE opposed to the Random Forest Regressor.

Finally, the best K-Nearest Neighbor model was optimized using 10-Fold Grid Search. This resulted in slightly better scores with $3.95 \cdot 10^6$ for the mean absolute error and 0.707 as R2 score.

On the aggregated, absolute data, the best performing K-Nearest Neighbor model was trained on all the training data and evaluated on a test set. The final predictions were sent to the submission page of the competition for review and returned a mean squared error of $7.43 \cdot 10^6$ which, according to the measurement of the time to eruption, is equal to 74.340 seconds, or, 21 hours.

Various regression models were trained on aggregated high-dimensional volcanic sensor data to predict the time to eruption. Based on this approach, detection of certain patterns of the data may have slightly lost their accuracy. Nevertheless, these results provide an interesting foundation for a switch to a more simplistic, data-driven prediction approach, in which high-dimensional data can be interpreted faster and more reliable.

4 Discussion

In this study we tested to which extend it is possible to accurately predict volcanic eruptions using machine learning based on simple seismic data. We found that results were most promising using a K-Nearest Neighbor regressor, achieving a significantly lower mean absolute error than it's other regression competitors.

The findings of this study complement those of Kilburn [2018], confirming that, even with minimal data, volcanic eruptions can be detected in time for people to be evacuated to safety. In addition, the methods applied in our study make the data accurate and representable, regardless of the size of the data.

The decision for using absolute data was made because the seismic nature of the original data would result in a mean of zero. Also, big seismic spikes in the data could be more easily detected and had a greater impact on the shape of the seismic pattern, making segments more distinguishable.

This study indicates that data-driven models are faster and more reliable in interpreting seismic signals and predicting the time to eruption than their manual counter-approach. However, some limitations are worth noting. Due to time- and memory constraints, certain methods and techniques could not be fully explored. One of our key observations was the increasing number of features emerging from the aggregation step, which could have negatively impacted model performance. Future work could improve on the efficiency of the data by inspecting and eliminating columns which are not or barely contributing to the model's decision making (Florian, 2020). Benefits of this approach can show an increasing contribution of Principal Component Analysis and Continuous Wavelet Transformation, techniques that were originally overshadowed by the sheer amount of aggregation features. Finally, future research should also consider the use of Neural Networks as a technique to improve final results.

References

Bronshtein, A. (2017). *Train/test split and cross validation in python*. *Towards data science*. https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6

Feike, S. (2020). *Multiple time series classification by using continuous wavelet transformation*. *Towards data science*. https://towardsdatascience.com/multiple-time-series-classification-by-using-continuous-wavelet-transformation-d29df97c0442

Florian. (2020). *Linear regression, trees and neural networks*. *Kaggle*. https://www.kaggle.com/florian12/linear-regression-trees-and-neural-networks

Isaienkov, K. (2020). *Ingv - volcanic eruption prediction. eda. modeling*. *Kaggle*. https://www.kaggle.com/isaienkov/ingv-volcanic-eruption-prediction-eda-modeling

Kilburn, C. R. J. (2018). *Forecasting volcanic eruptions: Beyond the failure forecast method*. *Frontiers in earth science*. https://www.frontiersin.org/articles/10.3389/feart.2018.00133/full

Lapinsa, S., C.Roman, D., Jonathan, Rougier, Angelis, S. D., Cashman, K. V., & Kendall, J.-M. (2020). *An examination of the continuous wavelet transform for volcano-seismic spectral analysis*. *Sciencedirect*. https://www.sciencedirect.com/science/article/pii/S0377027319303051

National Institute of Geophysics and Volcanology. (2020). *Ingv - volcanic eruption prediction*. *Kaggle*. https://www.kaggle.com/c/predict-volcanic-eruptions-ingv-oe/notebooks

Witze, A. (2019). *How ai and satellites could help predict volcanic eruptions*. *Nature*. https://www.nature.com/articles/d41586-019-00752-3

Appendices

Sensor 1 seismic activity comparison between original and sliced data

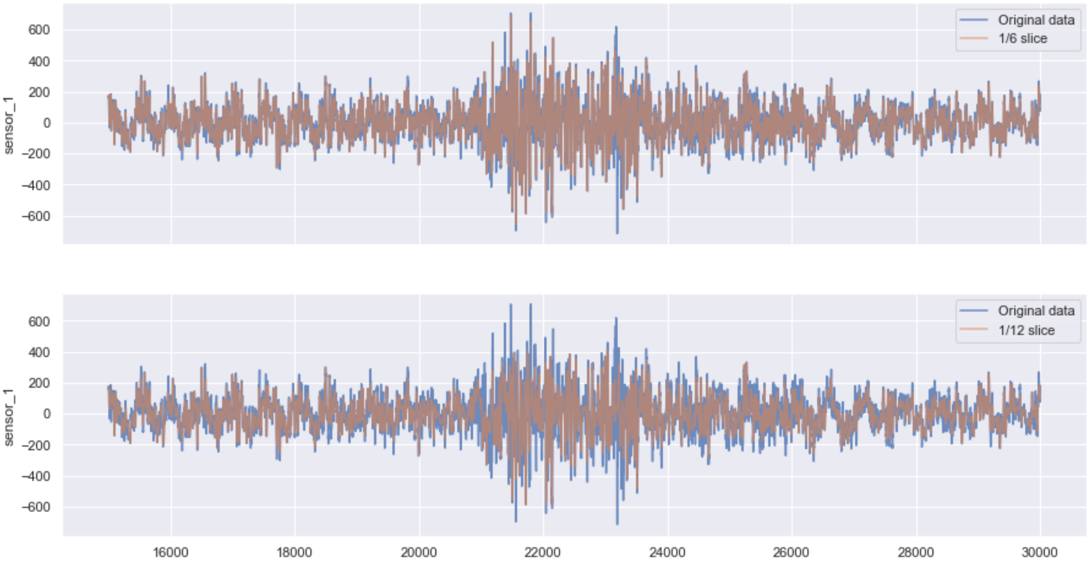

Fig. 1. The figure shows the original data as blue and sliced data as orange. The top row shows a 1/6 slice of the data plotted on top of the original data, the bottom figure shows 1/12th.

No linear correlations are present in the absolute aggregated data set

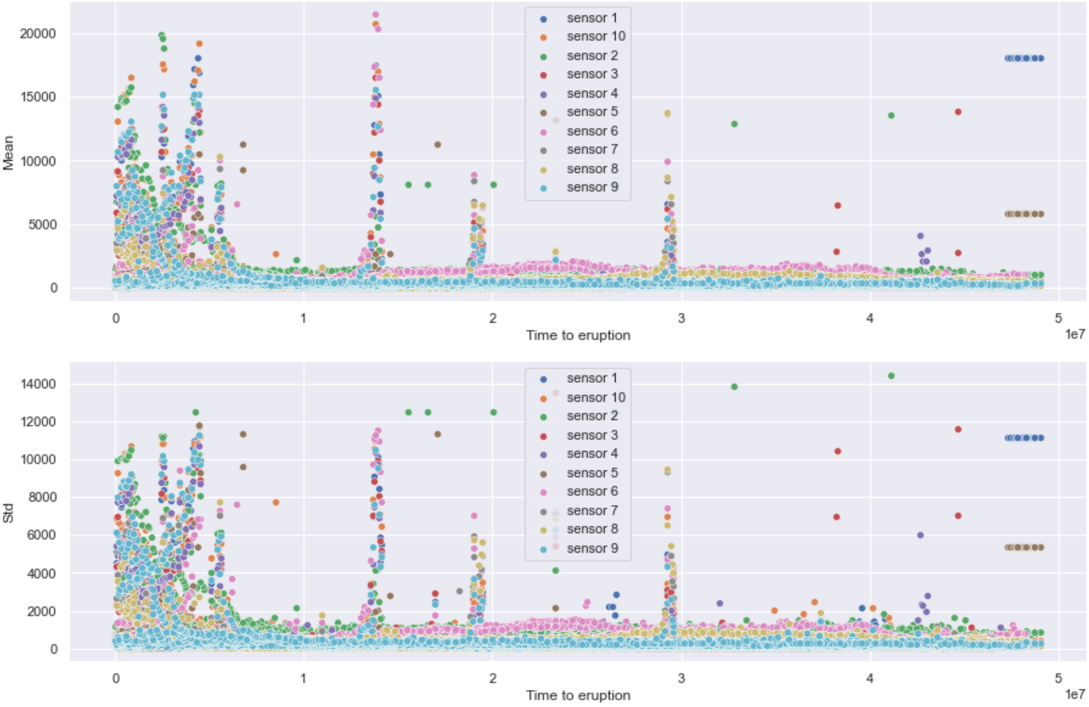

Fig. 2. The plot shows two measure: mean and std, of all the 10 sensors with the time to eruption on the x-axis.

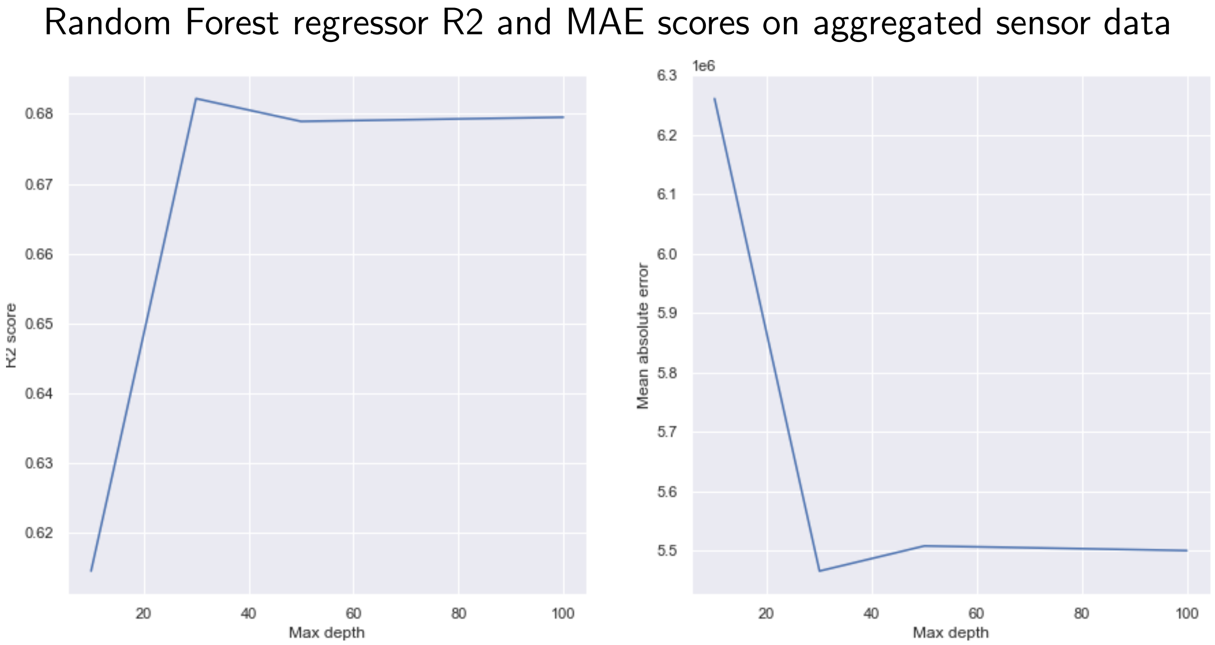

Fig. 3. The figure shows the R2 score on the left and the mean absolute error on the right of a Random Forest regressor. The x-axis shows how the scores change with the *max_depth*.

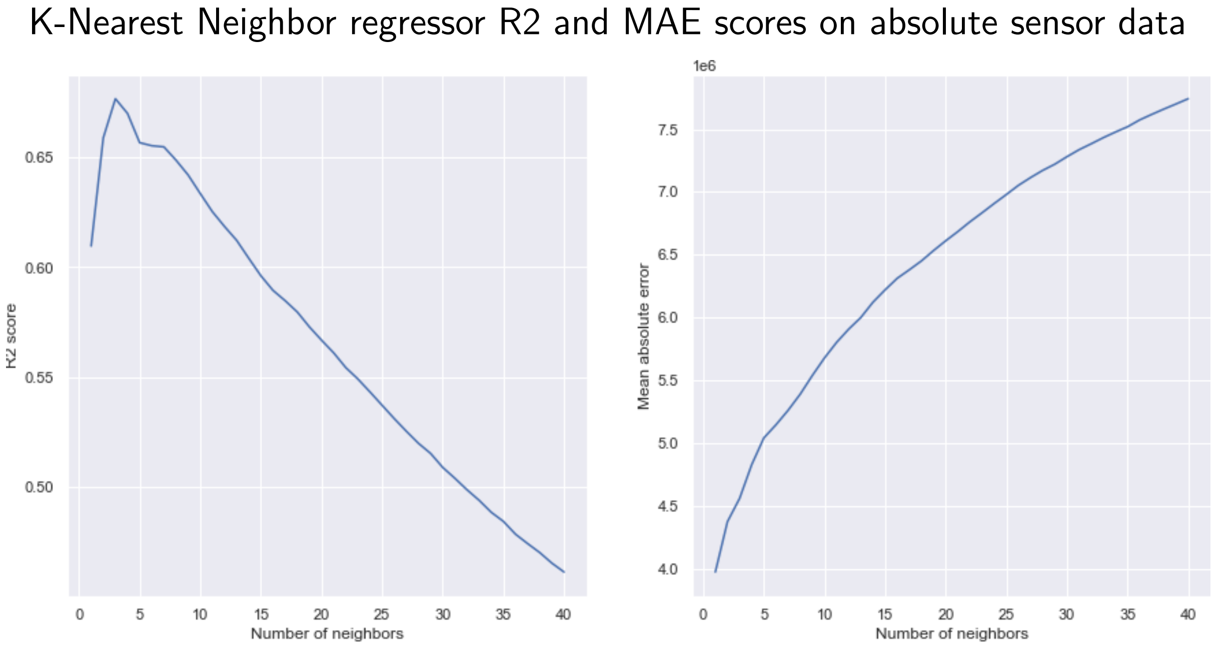

Fig. 4. The figure shows the R2 score on the left and the mean absolute error on the right of a K-Nearest Neighbor regressor. The x-axis shows how the scores change with an n amount of neighbors.