Comparing logistic regression and random forrest machine learning classifiers for landslide hazard in Vaud, Switzerland

Vikeså Håkon[1]

Abstract

Landslides costs billions of dollars each year (Klose et al. 2014). The mechanism are often complex and the implentation of machine learning can lead to a better understanding, mointoring and mitigation of damages.

1. Introduction

Machine learning is a discipline within computer science that has taken huge leaps within the last decades (Chollet, 2017). By using machine learning techniques the computer can look at input data and find the rules itself instead of being hard coded by a human. This enables the computer to be much more flexible and to work with much larger and varied datasets enabling the use of efficient data analysis in new fields of study.

A landslide is a downward movement of rock, debris, earth or soil (Meng, 2022). This natural hazard causes billions of dollars of damages each year (Klose et al. 2014). The knowledge of how landslide works and which factors influences it may lead to preventing both fatalities and other damages to the society. Machine learning can play a significant role for its ability to analyse large amount of data and find patterns in previously difficult to process datasets.

In this paper two machine learning algorithms will be trained, tuned and compared on a landslide dataset. The best model will then be used to create a landslide probability map of the canton of Vaud in Switzerland. The two models being logistic regression classifier and random forest classifier.

2. Data

The study area for this scientific paper is the canton of Vaud located in the western part of Switzerland.

The dataset is split into two sections. Both sections consists of eight categories where six of the categories are continues scales and two are categorical values. The six continuous categories are distance to road, a digital elevation model (DEM), topographical water index (TWI), plan curvature and profile curvature. The two categorical categories are the geology and the landcover. In addition the first sections contains a binary values for wheather or not the point is a landslide or not.

The first section of the dataset is used to train, tune and test the model. For comparing the performance and feature importance of the random forest and logistic regression algorithms an answer is needed. The second part of the dataset consist of eight maps the canton of Vaud, one for each of the eight features. These maps are the basis for the risk map of the canton of Vaud.

The part 1 of the dataset contains 5186 to train, tune and test the models. Part two of the dataset contains 4 520 223 points to make predictions of the psudo hazard map.

In the part two of the dataset there is one more value of geology compared to part one. The two categorical values is encoded using one hot encoders. A one hot encoder makes new columns in the dataframe for every unique value in the categorical value. Meanwhile the classifiers requires the dictionary to make a prediction. To come around this the geology value not featured in the training set is removed in the risk map. This regards 200 points out of 4 520 223.

The dataset is a proprietary dataset who's access is regulated by Tom Beuchler $< tom.beucler@unil.ch >$.

3. Methodology

Logistic regression was chosen as a baseline and random forest was used to compare to. The best performing algorithm would be used to create the risk map of Vaud. The data was already in large part processed and cleaned. All the steps are defined under and were done in python.

(1) The first step was to split the data into train, validation and a test set. The ratio was set to (73/12/15) for train/validation/test split. This was done to have a large dataset for the algorithm to train on, a validation set to tune hyperparameters on and have a sufficient test set to achieve consistent results. A one hot encoder was implemented for

[1] University of Lausanne, Lausanne, Vaud, Switzerland. Correspondence to: Håkon Vikeså <haakonjensen.vikesaa@unil.ch>.

Table 1. Performance of random forrest and logistic regression before optimazation

	LOGISTIC REGRESSION	RANDOM FORREST
TRUE PRED.	81%	84%
FALSE PRED.	81%	86%

the categorical parameters of geology and landcover.

(2) Secondly the base models were trained on the training set and then tested on the test set. These results can be seen in tab. 1.

(3) Thirdly the models hyperparameters were then tuned on the validation set and then tested on the test set. For the tuning an exhaustive grid search was implemented using $HalvingShearchCV$ for both classifiers. For the tuning of the logistic regression the number of features was tested with a range of 1,2,3,4 and 5. In addition the number of estimators was tested. The number of estimators was tested between 100 and 500 with a step of 50. For the logistic regression the solver was tested, for newton-cg, lbfgs, liblinear, the penalty was set to l2, the C was tested for 100, 10, 1.0, 0.1 and 0.01, and lastly the max iter was tested for 2000, 5000, 10000 and 20000.

(4) Lastly after having a performance evaluator and having the most suitable model prediction could be made over the entire region of Vaud. The images where imported and every pixel was iterated trough and processed through the model and a psudo landslide risk was outputted.

4. Results

The performance of both the both the default randomized forrest and the logistic regression can be seen in tab. 1

After a echaustive grid shearch was implemented using $HalvingGridSearchCV$ for both classifiers. The best parameters for random forest was max features = 5 and n estimators = 450. For logistic regression the best parametrs C = 1.0, max iter = 10000, solver = liblinear. From **??** the increse in the performance in both classifiers can be seen.

The optimized random forest classifier showed an increase in accuracy of $0,7\%$ while the optimized logistic regression showed an increase of 0.2%.

From seeing that the optimized random forrest classifier preformed slighlty better than both the logistic regression and the unoptimized random forest, the optimized random forest was chosen to make the risk map of Vaud.

In fig. 2 you can see the feature unportnace of the optimized random forrest cliassifier.

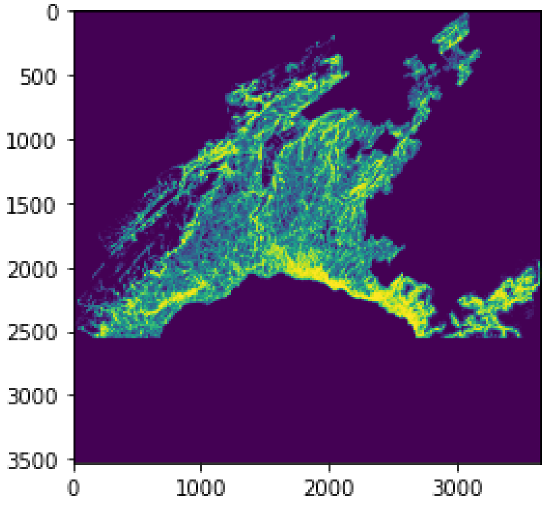

Figure 1. The prabability that each pixel has a landslide. The model is an optimized random forrest clissifier trained on a dataset of landslides in the canton of Vaud. The map is not finished since the processing timed out or crashed every time.

Fig. 3 showcases the feature importance of the differant geology and landcover found in the study area. Fig. 3 shows that glacial deposits and metamorphic rock are importand for detemining the the landslide risk. In addition no vegetation, shrub vegetation and forst are important landcovers.

5. Discussion

Fig.1 is not a ture risk map. This map calculates the propablility of there being a landslide right now, no temporal feature to the dataset. This is beacause in the training set that teh algorithm trained on there is no time parameter an therfore it all happens at once. There this is not a true risk that quantifies the excact risk of one place compared to another but as a relative map that showcases a regions with similar features as other places where landslides has occured. As seen the fig. 1 is not complete. This is due to not having the time to process the whole image. A better method would have been to use a reducer and reducing the resolution of the images. However this was not done in this paper.

For encoding the categorical values a one hot endocer was used. This required a value from the geology caterory to be removed. The use of a differant encoder may have solved this problem and kept every catergorical value. In addition during the calculations of the hazard map every pixel was iterated through. To make sure all inputs into the calssifier would remain the same size a mock dictionary was made inside the loop to make it the exact same format as the

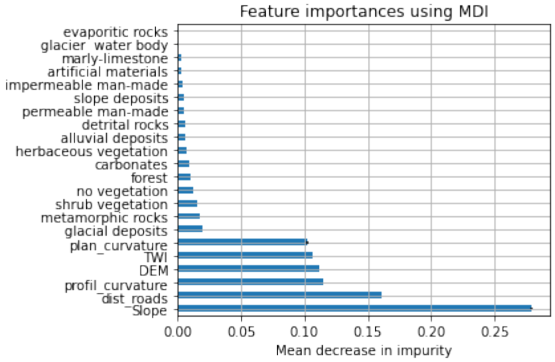

Figure 2. This is the feature importnace of the Rnadom forrest classifier.

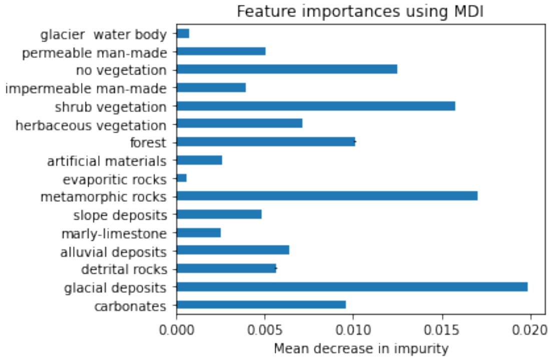

Figure 3. The feature importance of each geology and landcover

training dataset. This made the calcualitions take a long time. A better chosen encoder of catergorical values may have made an more efficient calculation.

By tuning the hyper paramters there is only a margianl imporovement. This begs the question if for this scientific question it is necessary to tune the hyperparamters given the extra time it takes both to code but also to run the code.

When looking at feature importance in fig. 2 the most important feature is the slope. This is reasnable since landslides are gravity driven processen an the steeper the terrain the higher the chances of a landslide is. However the second most important feature is the distance to road. This is slightly more supprising given the fact that natrully the distance to road does not have anything in commen with roads. The roads may however play into both the slope stability bot also there might be a higher probability of a landslide registed and mapped if it is located closer to a road.

6. Refrances

Chollet, F. (2017). Deep learning with python, *Manning publication Co.*

Klose, Martin and Highland, Lynn and Damm, Bodo and Terhorst, B. (2014) Estimation of Direct Landslide Costs in Industrialized Countries: Challenges, Concepts, and Case Study, *In Landslide Science for a Safer Geoenvironment*

Meng, X. (2022, October 15). landslide. Encyclopedia Britannica. https://www.britannica.com/science/landslide

7. Appendix

The code used for this project can be found here.