Machine Learning Project
Predicting Snow Line Elevation and Activity of Glaciers

Margaux Hofmann

Abstract

This project aims to determine the accuracy with which it is possible to predict the snow elevation line and the glacier activity (retreat or advance), using Machine Learning algorithms. These two parameters are indeed important for current glacier melt, and yet they are among the least frequently reported parameters in glacier inventories, as they are difficult to measure remotely. A Multiple Linear Regression is applied to predict Snow Line Elevation, and a Random Forest is used in comparison with a logistic regression and a simple Decision Tree Classifier to predict activity. The prediction of Snow Line Elevation give really good result, while the prediction for the activity is not optimal, even with Random Forest.

1. Introduction

Monitoring glaciers to know their characteristic and activity is complicated, and takes a lot of time if the measurements are made in situ (Rabatel et al., 2017). It's however important to have information about glaciers, especially in the context of climate change. Most of the glacier in the world are retreating (Sommer et al., 2020). Only a few glaciers in the world are stationary and even fewer are advancing (Lodolo et al., 2020). It's a major issue, particularly for water management but also for mountain stability (Jouvet et al., 2011) It would therefore be useful to be able to collect information about glaciers remotely.

This was done by the National Snow and Ice Data Center, which created an inventory of the world's glaciers, based on aerial images and maps, the World Glacier Inventory. But some parameters are still difficult to determine remotely, even with the availability and the good resolution of satellite images. Therefore, many values are missing, especially for Snow Line Elevation and Glacier Activity, where there is no value for the majority of the glacier.

The Snow Line Elevation is important because it can be used to approximate the ELA, the Equilibrium Line Altitude (Racoviteanu et al.(2019), Rabatel et al.(2012)). The ELA separates the ablation zone (area down the glacier where the ice melt) and accumulation zone (upper of the glacier, where the snow is transformed in ice) over one year (Benn & Lehmkuhl, 2000). It is an key factor to take into account to understand the activity, melting or advance of a glacier (Braithwaite & Raper, 2009).

The goal of this project is to be able to determine the Snow Line Elevation and the Glacier Activity, based on the other information of the dataset, where the value are not missing. This project will comport two different steps. The first one is to predict the Snow Line Elevation with a very simple Multiple Linear Regression. The second step is to predict the activity of the glacier, comparing three different algorithm : Logistic Regression, Decision Tree Classifier and Random Forest.

2. Data

The data used was found on Kaggle. It's the World Glacier Inventory, provided by the National Snow and Ice Data Center (NSIDC), based in the USA. This inventory, with data from 2000 but updated in 2012, includes almost all the glaciers of the world, with their position (longitude and latitude). A lot of other parameters are available, covering areas such as elevation, orientation, activity, length, width, area and snow line elevation. The categorical parameters are already encoded into numerical value (from 0 to 9 for class and form, from 0 to 8 for activity, where 0 represent uncertainty). All those parameters are measure remotely. The number of glaciers covered by a parameter is different for each parameter. The shape of the dataset is 132890 row (glaciers) and 39 columns. The variables we will try to predict are first the continuous variable Snow Line Elevation, which correspond to an altitude, and secondly the activity of the glacier, which is a categorical variable ranging from 0 to 8, 0 and describing different states of the glacier (eg: slight advance, marked retreat, stationary ...)

3. Methodology

This project is done using Python in Google Colab, mainly with the Scikit-Learn library.

3.1. Snow Line Elevation - Multiple Linear Regression

As the Snow Line Elevation is a numerical variable, a simple Multiple Linear Regression is used. The variables choose to try to predict are "Latitude", "Longitude", "Glacier Area", "Mean Length", "Mean Elevation","Minimum Elevation", "Maximum Elevation", "Primary Class", "Glacier Form". Rows containing Nan have been removed from the dataset, so the number of row is now 5700. The dataset has been split into training, validation and test set, with 70% of the data for the training set, 15% for the validation set and 15% for the test set. As those parameters have different scale, a standard transformation is first applied over the predictors. After the construction of the model, the contribution of each predictor has been extract to determine the feature importance.

Two metrics have been chose to evaluate the model. The coefficient of determination (r2 score), which indicate how good the predictors explain the distribution of the target values. The closer the value is to 1, the better the model is. The second metric is the Mean Absolute Error, which indicate the average of the residuals in the prediction. In other terms, it represent the average of the differences between the predicted and actual values.

3.2. Activity

The activity (the variable that we want to predict) is a categorical variable, encoded into numerical values. Three different algorithms will be test :

- A simple Multinomial Logistic Regression (with default parameters)
- A Decision Tree Classifier (with optimized parameters)
- A Random Forest Classification (with default hyperparameters and optimized ones)

The variables used to try to predict the activity are the same as for the Linear Regression, but this time in addition with the Snow Line Elevation. The first step was to replace the 0 values of the activity variable by NaN. Indeed, the value 0 means uncertain, and includes many glaciers whose activity could not be clearly determined. It is possible to predict with this class, but as it has many more glaciers than all the others, the accuracy will be very high, despite many errors for the other classes (because of their smaller number). For a concrete application, it would be useful to integrate this class, and this was done during the calculations for this work, but in this document only the results without this class will be presented, for reasons of readability and because we focus more on comparing the different algorithms. Rows containing NaN have been removed from the dataset, so the number of row is now 4988. The dataset has been again split into training, validation and test set, with 70% of the data for the training set, 15% for the validation set and 15% for the test set.

3.2.1. Logistic Regression

A simple Multinomial Logistic Regression is applied on the three set. The performance of the model were evaluate with the accuracy.

3.2.2. Decision Tree Classifier

A single Decision Tree as been applied to the different sets. First a Decision Tree with the default hyperparameters has been applied on the validation set. To improve the result, a GridSearch was implemented to determine the best parameters among the following : Max Depth (maximum depth of the tree), Max Features (how many features are taken in account for each split), Min Sample Split (minimum number of sample to split a node) and Min Sample Leaf (minimum number to be considered as a node). The Table 1 show the different parameters tested.

Table 1. Hyperparameters tried using GridSearch for the Decision Tree Classifier

Hyperparameter	Range	Best
Max depth	20, 60, 80, 100	80
Max features	1,2,3,4,6,8	8
Min sample split	2,4,8,12	12
Min sample leaf	2,3, 4,5	4

In this work only the results with the best hyperparameters are presented.

3.2.3. Random Forest Classification

As for the Decision Tree Classifier, the model was first implemented with the default parameters. Different hyperparameters were then tried, using GridSearch. Those parameters are presented in Table 2. They are the same as for the Decision Tree Classifier, plus the number of estimator (number of trees in the forest) and bootstrap (True or False, and if False, the whole dataset is used for each tree).

Table 2. Hyperparameters tried using GridSearch for the Random-Forest classification

Hyperparameter	Range	Best
N_estimators	100, 200, 300	200
Max depth	20, 60, 80, 100, 120	80
Max features	1,2,3,4	3
Min sample split	1,2,4,8,10,12,20	8
Min sample leaf	3, 4,5	3
Bootstrap	True, False	True

4. Results

4.1. Multiple Linear Regression

The result of the Linear Regression was measured with r2 score and the Mean absolute error (as explained in section 3.1). The values obtain are registered in the Table 3.

Table 3. Coefficient of determination (R2 score) and mean absolute error (MAE) for the Linear Regression

SET	R2_SCORE	MAE
TRAIN	0.986	85.75
VALID	0.978	84.87
TEST	0.988	85.64

We can reach good coefficient of determination, as the value are near 1 for the three sets. The MAE is around 85, which means that over each set, the average error is 85m. Knowing that the altitude of the Snow Line is between 0 and 6000m for the test set, the result is quite good.

The figure 1 below show the feature importance.

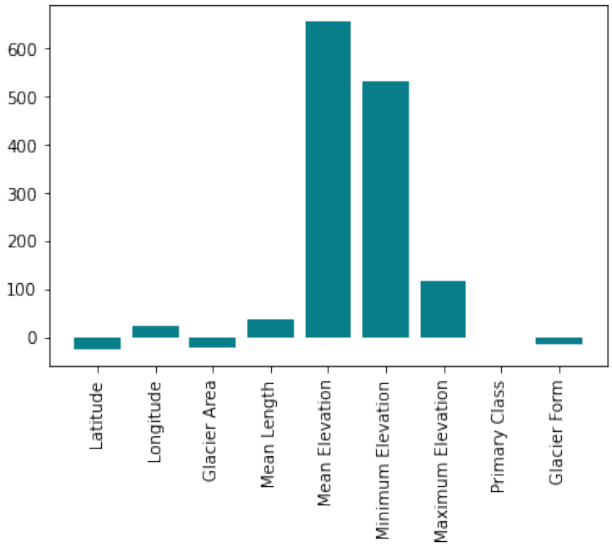

Figure 1. Importance of the different features for the prediction of the Snow Line Elevation

We can see that it's the Mean Elevation and Minimum Elevation that are the most important variable to predict the Snow Line Elevation. Indeed, if we only use the variable Mean Elevation, we can already reach a r2 score of 0.97.

4.2. Activity

The accuracy of the three different models are presented in Table 4.

Table 4. Accuracy for the Logistic Regression (LR), Decision Tree Classifier (DTC), Random Forest with default parameter (RF_1), and Random Forest with best hyperparameters (RF_2)

SET	LR	DCT	RF_1	RF_2
TRAIN	0.642	0.898	1.000	0.932
VALID	0.631	0.772	0.811	0.826
TEST	0.637	0.785	0.815	0.813

A great improvement of the result are noticeable if we go from a logistic regression (LR) to a decision tree (DCT). But then there is only a small increase in accuracy with the Random Forest (RF_1), and no better result with the change of the hyperparameters (RF_2).

The accuracy with the Decision Tree and the Random Forest, which is around 0.8, is quite good though, although this is still too low for this model to be applicable. But the Decision Tree and the Random Forest seems to overfit the data, as the accuracy is much higher for the train set than for the two other sets.

If we want to go in a bit more details, we can create a confusion matrix. It's another way the evaluate the quality of the model. The row of the matrix are the real class and the columns are the predicted class. It's than easily possible to see where the classes were not correctly predict. The figure 3 shows the Confusion Matrix for the test set with the Random Forest with the tuning of the hyperparameters.

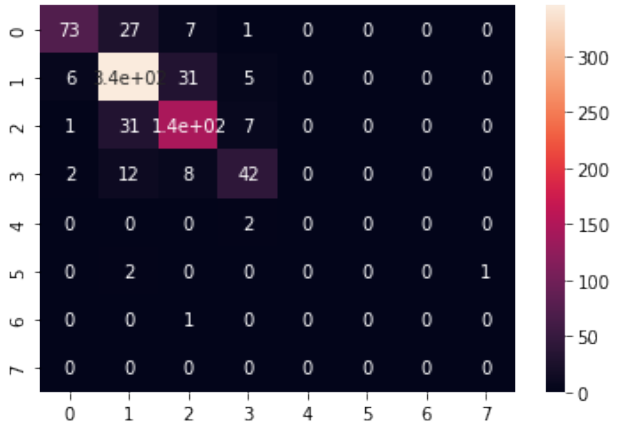

Figure 2. Confusion Matrix for the test test of the Random Forest with the change in hyperparameter (RFC_2)

We can see on it that there is a problem with the last classes, as no glacier is accurately predicted in these classes. The model struggle to classify the last classes present on the dataset. Otherwise we can see that most of the glacier are in the class 1, which correspond to slight retreat. Most of

the errors are classifications in the classes next to them (for example class 1 instead of class 2).

The feature importance for The Random Forest and for the Decision Tree (both with the changed hyperparameters) are shown in figure 3.

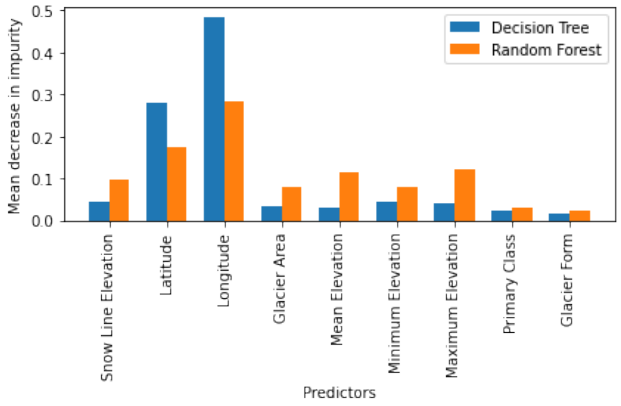

Figure 3. Feature importance for Decision Tree Classifier and Random Forest, both with the change in hyperparameters

In both case, it's the longitude which has the more importance, followed by latitude. The importance of the Snow Line Elevation change pretty much from on model to another. The Decision Tree is mainly based on the longitude.

5. Discussion

The prediction of the Snow Line Elevation can reach really good accuracy. These very good results can be explained by two main reasons. Either there is an overfitting of the data, or it is a very basic prediction, which effectively depends on few parameters. Given that the results (accuracy) are similar for the training, validation and test set, and that we are looking for an altitude, based on other altitudes, we can rather tend towards the second hypothesis.

For the prediction of the activity, it's quite surprising that the Random Forest overfit the data, because Random Forest is supposed to avoid the problem of overfitting by having multiple trees. The result with the Logisitc Regression are not satisfying, but at least there is not overfit of the data.

The model is not able to predict quite good the last classes. Two mains reasons can maybe explain this problem. The major reason is because there are not enough glaciers in those class to train good the model (between 4 and 8 for the last classes), and the prediction is then really difficult. Another reason can be that perhaps the min sample split should be changed, because the result obtain with the Grid Search is a too high value, which is not good for class with less glaciers. But as there are not a lot of glaciers in those classes, we can also said that maybe it's less important that the model don't recognize those activity, as they are rarer. But in reality it's not really the case, as those rarer classes are classes that often need to be monitored, as they can cause natural hazards, such as the class surge.

Concerning not the algorithm but the result in itself, that show that is seems possible to predict the activity of the glacier based on different parameters. We saw that the parameters with that are the most important to predict the activity is longitude, followed by latitude. This seems rather logical and reflecting reality, since advancing glaciers are concentrated in a few places on earth, and when in a given region glacier dynamics are generally the same (with some exceptions). This would mean that with a few parameters easily identifiable by remote sensing (latitude,longitude and elevation for example), it might be possible to update the activity data of glaciers relatively quickly

6. Conclusion

The Linear Regression for the Snow Line Elevation gives really good result, and depends mainly on few predictors (Mean Elevation and Minimum Elevation, that can alone predict really well). It's probably due to the fact that it's a really easy linear prediction.

The classification for the glacier activity give less good results, but still an accuracy of around 0.8. A great improvement in results can be seen when moving from a simple Logistic Regression to a Decision Tree and the the result cannot really be improved with a Random Forest. The reason why the Random Forest algorithm seems to overfit the data still needs to be investigated.

Another think that could be done to improve the model should be to take in account some climate or meteorological parameters.

7. Code and Data Availability

The script used for this project is available under this link : https://github.com/MargauxHofmann/2022_ML_EES/blob/main/ML_Project/ML_GlacierProject_notebook.ipynb

The data used are available under this link : https://www.kaggle.com/datasets/nsidcorg/glacier-inventory?resource=download

And the documentation supporting the data under this link: https://nsidc.org/data/g01130/versions/1

References

Braithwaite, R. J., Raper, S. C. B. (2009). Estimating equilibrium-line altitude (ELA) from glacier inventory data. Annals of Glaciology, 50(53), 127-132. https://doi.org/10.3189/172756410790595930

Jouvet, G., Huss, M., Funk, M., Blatter, H. (2011). Modelling the retreat of Grosser Aletschgletscher, Switzerland, in a changing climate. Journal of Glaciology, 57(206), 1033-1045. https://doi.org/10.3189/002214311798843359

Rabatel, A., Bermejo, A., Loarte, E., Soruco, A., Gomez, J., Leonardini, G., Vincent, C., Sicart, J. E. (2012). Can the snowline be used as an indicator of the equilibrium line and mass balance for glaciers in the outer tropics? Journal of Glaciology, 58(212), 1027-1036. https://doi.org/10.3189/2012JoG12J027

Racoviteanu, A. E., Rittger, K., Armstrong, R. (2019). An Automated Approach for Estimating Snowline Altitudes in the Karakoram and Eastern Himalaya From Remote Sensing. Frontiers in Earth Science, 7. https://www.frontiersin.org/articles/10.3389/feart.2019.00220

Sommer, C., Malz, P., Seehaus, T. C., Lippl, S., Zemp, M., Braun, M. H. (2020). Rapid glacier retreat and downwasting throughout the European Alps in the early 21st century. Nature Communications, 11(1), Art. 1. https://doi.org/10.1038/s41467-020-16818-0

WGMS, and National Snow and Ice Data Center (comps.). (1999). World Glacier Inventory, Version 1 [Data Set]. Boulder, Colorado USA. National Snow and Ice Data Center. https://doi.org/10.7265/N5/NSIDC-WGI-2012-02. Date Accessed 12-18-2022.