School of Management, Economics, Law, Social Sciences, International Affairs and Computer Science

Predicting Danger Level Classes of Avalanches in The Region of Graubünden

A Multiclass Classification Approach

Report

Workshop Fundamentals of Data Science

05.12.2022

Collin Arendsen
(21-617-204)
Florence Bloch
(21-613-484)
Lisa Katharina Rumplmayr
(21-621-222)
Yasmin Mohamed Sufi
(21-622-451)

Table of Contents

Table of Figures

1. Importance of Danger Class Levels

In our project, we want to predict the danger class levels of avalanches in the canton of Grisons. Avalanches cause 22 fatalities on average each year in Switzerland, therefore we want to provide the resorts a model with which they can predict the danger class levels, thereby facilitating the protection of its visitors and reducing the number of lives lost each year. (Zweifel et al., 2021, p. 104).

The danger class level is determined by a number of factors, for example skier penetration depth (Pen_depth), sum of daily height of new snow (HN72_24), relative humidity (RH) etc. As the danger class level rises, so does the likelihood of an avalanche triggering. Hence the numbers can be interpreted as 1: low, 2: moderate, 3: considerable, 4: high, 5: very high. (*Danger Levels - SLF*, n.d.)

2. Data Cleaning

The original data set that we used contained around 292'837 rows and 78 columns. Our first step was to see how many of these values were NaN and if we dropped these how many rows we would be retaining. After dropping the NaN it turns out that we retained about two thirds of the data, around 211'772 rows. As this is a decent amount of information, we proceeded to looking at the column types (object, float etc) and dropped the object types, as they did not add to the information we need for predictability. From the data set, we only looked at station codes that are in Grisons and subset them from the original data. Upon inspecting how many values each of the 5 classes had, class 5 had the least values. In both the case of danger level 4 and 5, it is advised to refrain from open slopes as the avalanche risk is extremely high. Hence, we merged classes 4 and 5 because they have an almost equal effect on skiers and the resorts, and it would help reduce the class imbalance thus making better predictions. After determining which features are the most important for predicting the danger class levels, we took the top 11 most important features and we finally had the specific data set we would proceed to work with.

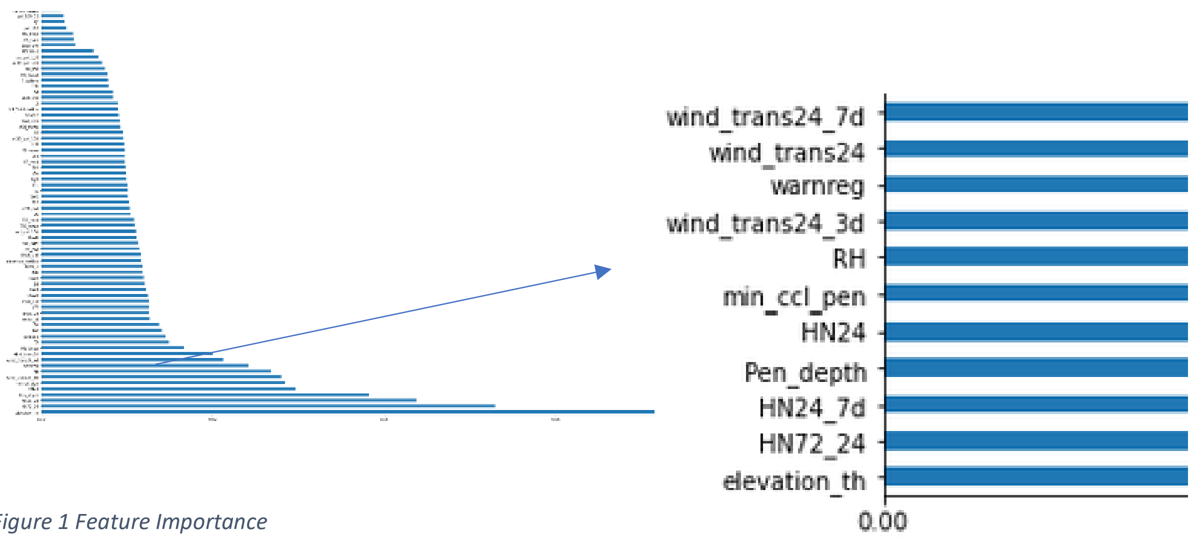

Figure 1 Feature Importance

3. Principle Component Analysis

To gain more insight into the data that we have, we proceeded to do a Principal Component Analysis. With the help of this dimensionality-reduction method, we reduced our dimensionality to 2 in order to visualize the data. Thus, we wanted to see if our data is distinguishable and hence can be grouped into the individual danger classes. The data seemed to be concentrated in the below shown manner.

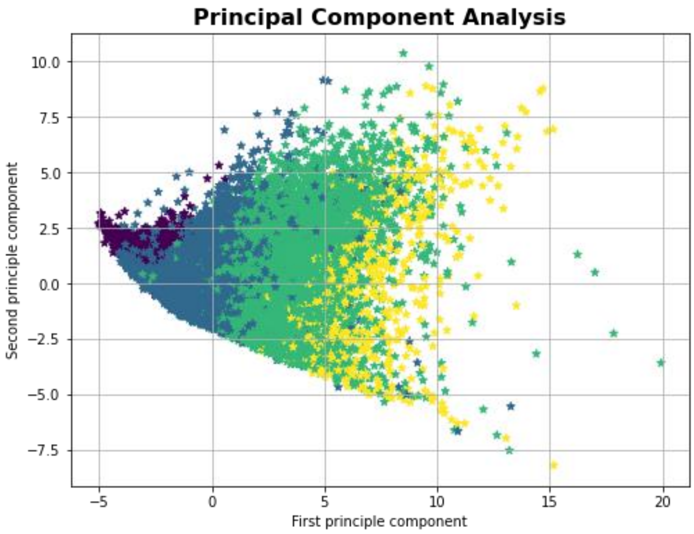

Figure 2 Principal Component Analysis

4. Logistic Regression

We grouped the danger classes into a less dangerous and a more dangerous group. Group 0 contains the danger classes 1, 2 and group 1 contains classes 3, 4 and 5 to perform a statistical analysis using a logistic regression. Class 3 was added into the more dangerous group because almost 50% of death occurs at this danger level (Gefahrenstufen - SLF, n.d.). The goal is to predict a binary outcome, such as 0 or 1 based on prior observations of the data set. The accuracy is around 78%. We proceeded to improving the model and could only successfully improve the recall, which measures how good our model is at correctly predicting positive classes from around 65% to 73%.

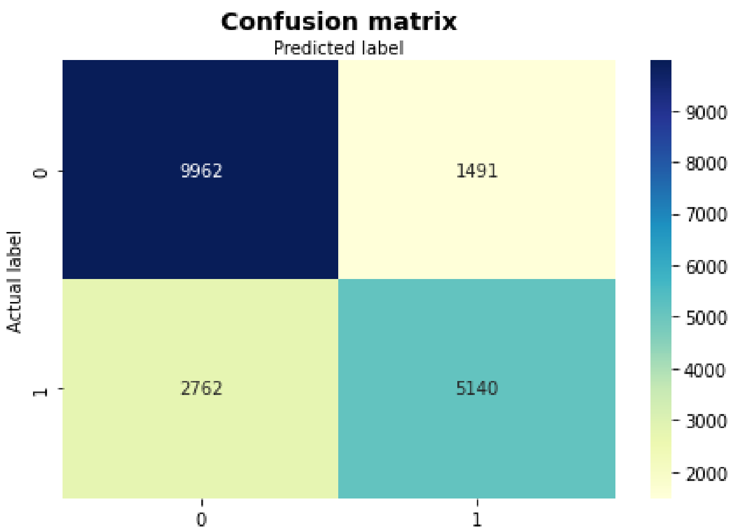

Figure 3 Visualizing Correct Classifications and Misclassifications

5. Models

5.1 Neural Network

One of the most suitable models for our problem, namely a multiclass classification, is Neural Networks. We used this model to predict the danger class level. We divided the data into train (60%), validation (20%) and test (20%) data. The accuracy of our predictions was around 75% on the train data, and 74% on both the validation and test data.

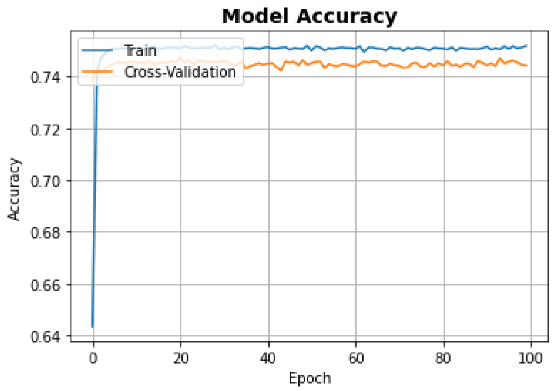

Figure 4 Model Accuracy

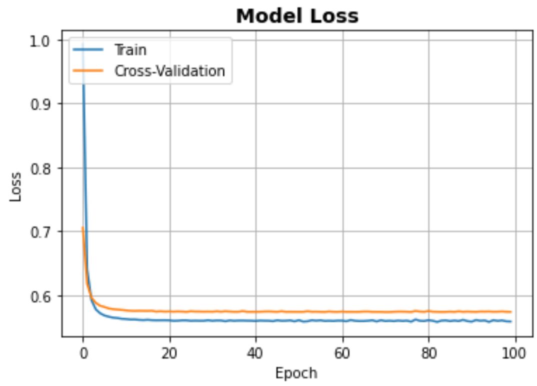

Figure 5 Model Loss

5.2 Decision Trees and Random Forests

After conducting the neural networks, we investigated ways to improve our accuracy using decision trees and random forests. We started by implementing a simple decision tree. In doing thix, we got an error of almost 100% for class 5 as our model was unable to make any predictions due to the low number of values for this danger level. We realized that in both the case of danger level 4 and 5 it is advised to refrain from snow sports away from open slopes as the avalanche risk is extremely high. We therefore decided to merge class 4 and 5 as their implications have an almost equal effect on skiers and the resorts. This improved our precision for the highest danger level to 59%, however, our recall value was still low at 16%. Using the merged classes in a decision tree with cross validation gave us an overall accuracy of 74%. To further improve our recall value, we used random forests. This increased the recall value, but it was still low (30%) for danger level 4. The overall accuracy for random forests was 77%.

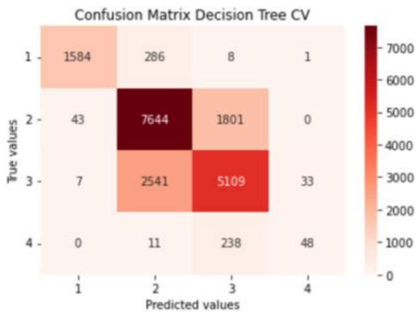

Figure 6 Confusion Matrix Decision Tree CV

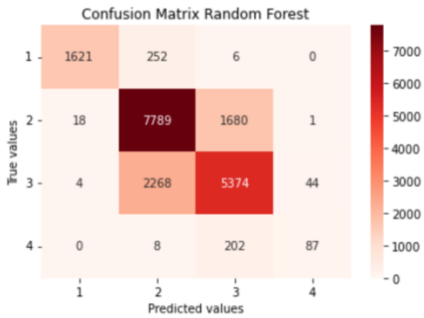

Figure 7 Confusion Matrix Random Forest

To improve our recall value, we wanted to even out our imbalanced data set and tried over as well as under sampling. Given that there are still comparatively few values in the fourth danger category and that the model learns better with more data, it seems logical that the over sampling worked better. With the oversampling we tried random oversampling as well as SMOTE oversampling, whereby SMOTE performed better. Oversampling improved our recall score compared to the unbalanced data set and simultaneously lowered the precision We decided to stick with oversampling compared to the unbalanced data set. Reason being that we consider recall scores more important as our business case is to enable ski resorts to make predictions to increase the safety of its visitors. Stating a too low danger level will have more fatal consequences for them than if they were to state a danger level that was too high. Fatalities brought on by inadequately foreseen avalanche hazards would drive away guests from the ski resort. Compared to the imbalanced data set, our overall accuracy for random forests reduced but only by one percentage point to 76%, however, our recall scores significantly improved (figure 9).

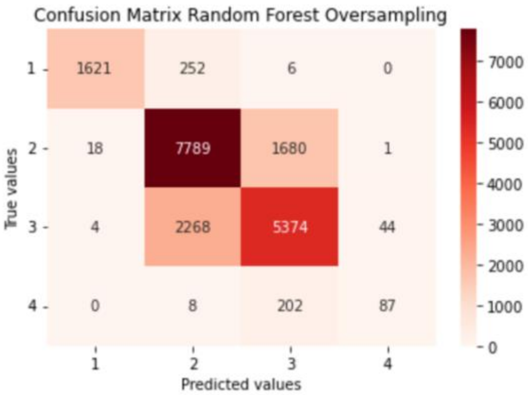

Figure 8 Confusion Matrix Random Forest

	precision	recall	f1-score	support
1.0	0.93	0.90	0.91	1879
2.0	0.77	0.79	0.78	9488
3.0	0.73	0.70	0.71	7690
4.0	0.35	0.62	0.45	297
accuracy			0.76	19354
macro avg	0.70	0.75	0.71	19354
weighted avg	0.76	0.76	0.76	19354

Figure 9 Classification Report Random Forest Oversampling

In a last step we aimed to further improve our accuracy by tuning the hyperparameters of our random forest data. Additionally, we used the booster function with our oversampled data, however, this resulted in the same accuracy as when using random forests, but a lower recall value. We therefore decided that the best model to predict the danger level of avalanches in Grisons is the random forest using the oversampled data. In the table below we summarized all the results from the previous discussed models.

	Av. Precision	**Av. Recall**	**Accuracy**
Decision Tree CV	0.75	0.62	0.74
Random Forest	0.79	0.67	0.77
Decision Tree CV Oversampling	0.61	0.69	0.69
Random Forest Oversampling	0.70	0.75	0.76
Boosting Oversampling	0.71	0.72	0.75

Figure 10 Results Summary of all models.

5.3 Time Series Analysis

To check for seasonal trends, we performed a time series analysis. Here, we only used two columns of information, first the datetime ("Datum") and second the danger level ("dangerLevel"). To get a singular value per day, we computed and used the average danger level for Grisons.

This figure illustrates that there are time breaks between the data points as there were no rows depicting values between May and October. All the predictions from the following models are hence only applicable during the ski season of November to April. Moreover, this figure shows that most values seem to fall somewhere between danger level 2 and 3. It is interesting to note that after 2013 danger level 1 seemed to vanish so that the lowest danger level of is 1.

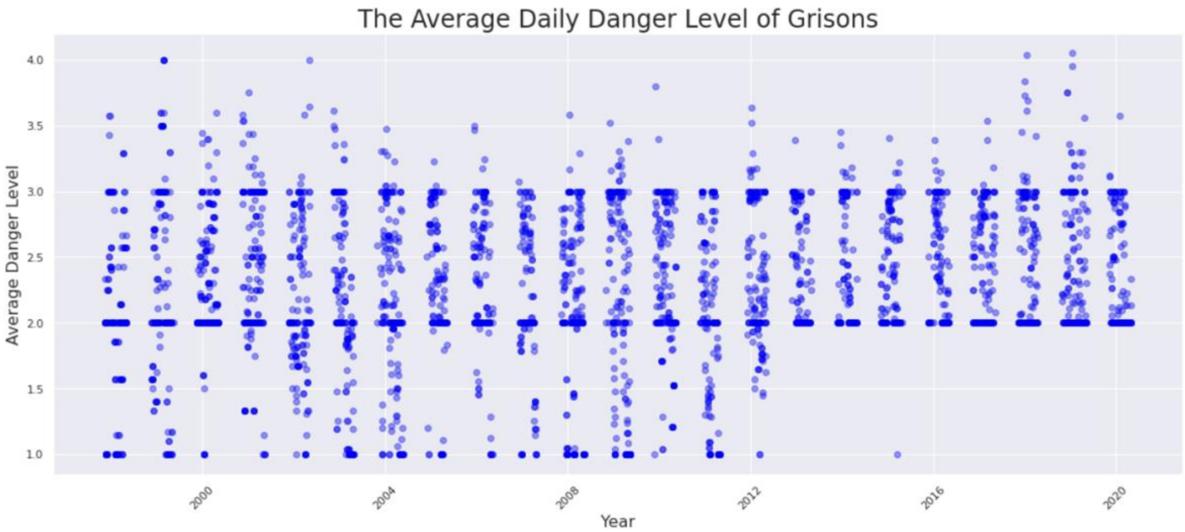

Figure 11 The Average Daily Danger Level of Grisons

To get a better understanding of our values, we computed a rolling mean for 7 days as well as a rolling mean for 30 days to compare the results with the average daily values. The test and training data split was done with 70% training data and 30% test data where the 30% include the most recent rows of data. As time series analysis looks for patterns based on an increase in time, splitting the data randomly into test and training data without regards to an increasing datetime would not result in a good solution. We tried models such as Auto-Arima and Pyaf, but they resulted in the prediction being a straight line as they took the average for all values within the test data as the forecast. However, the model NeuralProphet managed to create a different prediction.

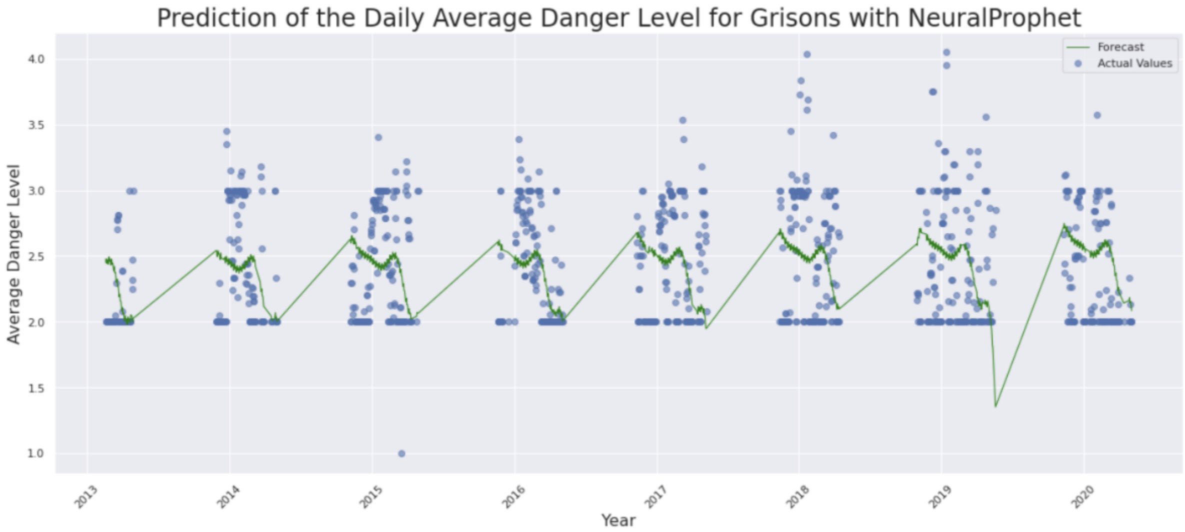

Figure 12 Prediction of the Daily Average Danger Level

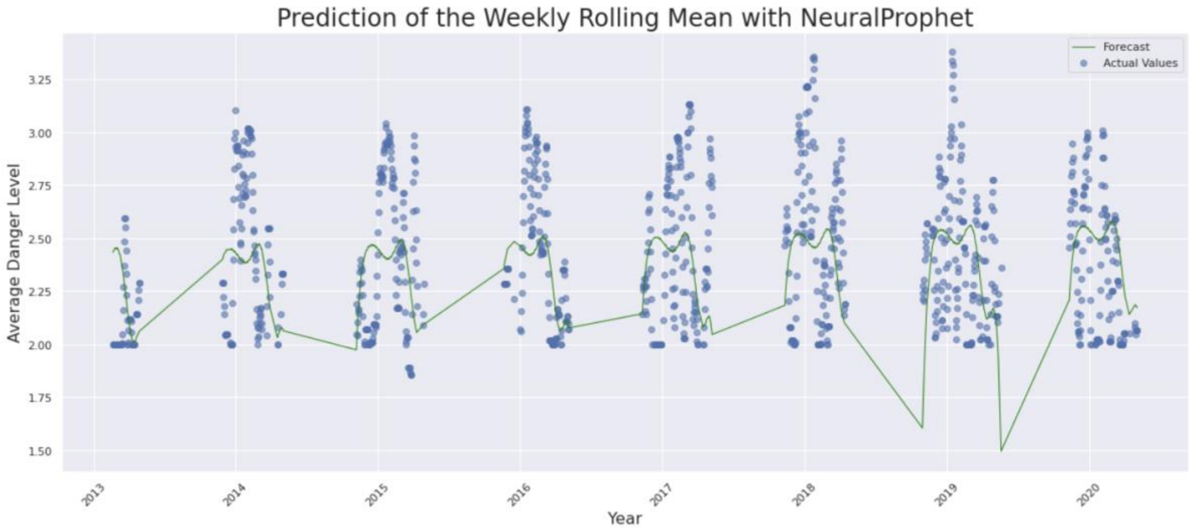

Figure 13 Prediction of the Weekly Rolling Mean

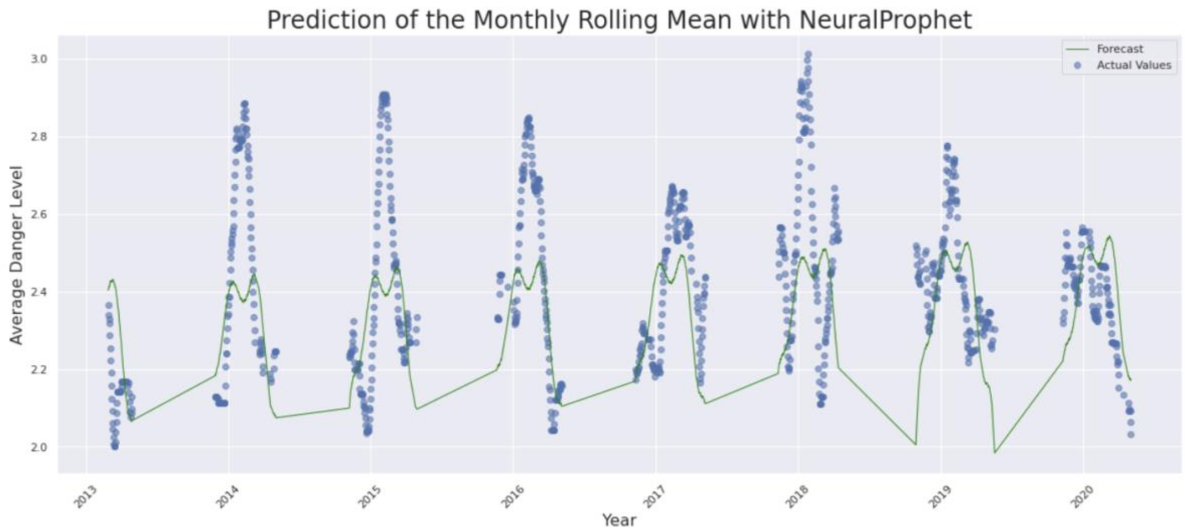

Figure 14 Prediction of the Monthly Rolling Mean

What becomes apparent is that the average danger level seems to be the highest during the peak ski season of December, January and February. This makes sense as during these months there is usually the most snow as figure 15 shows. (*Total Snowfall for Switzerland in November - Current Results*, n.d.).

	Snow Height in cm	
Months	**Chur**	**Davos**
November	10	64
December	20.6	65.3
January	34	95.2
February	24.7	80.5
March	10.3	65.5
April	1.5	51.2

Figure 15 New Snowfall per Month

Thus, Grisons' average daily danger level decreases the more it reaches the outskirts of the winter season, with November, March and April having lower average daily danger levels. The weekly and monthly rolling mean illustrate these findings even more.

6. Conclusion

The dataset we found on EnviDat.ch (Weather, Snowpack and Danger Ratings Data for Automated Avalanche Danger Level Predictions, n.d.) had over 290'000 rows. After deleting NaNs and subsetting the data to only look at stations in Grisons, we had the final data set that we proceeded to work with. Firstly, we ran a principal component analysis on it, which retained the information but reduced the dimensionality to two, thereby making it possible to plot our data and get first impressions on it. After dividing the danger class levels into two, one for high danger and one for low danger, we ran a logistic regression on it.

In order to predict the individual danger levels we used the following models: a neural network, decision trees, and random forests. The neural network gave us an accuracy of 74% on the test data. We wanted to further improve our accuracy by using decision trees and random forests. To increase recall, accuracy, and precision even more, we optimized the n_estimators, max_depth, min_samples_leaf, and oversampled the higher danger level. By doing this, we improved the recall value while achieving an accuracy of 76% with the oversampled random forests.

Lastly, we performed time series analysis by using Auto-Arima, PyAF and NeuralProphet as forecasting models. Auto-Arima and PyAF resulted in the prediction being the mean value while NeualProphet showed that there was a tendency for higher danger levels during the peak ski season of December, January and February.

To achieve an even higher accuracy, finetuning a wide range of hyperparameters for the boosting model could be done in addition with stacking the models together. However, the hyperparameters could also be finetuned to much so that there is overfitting on the training data set. Hence, this is a very sensitive process where a lot could also go wrong. Moreover, stacking all models together would exponentially increase the computation time and thus it would be a trade-off for a higher accuracy.

7. Variable Legend

Variable Name	Variable Description
dangerLevel	Avalanche danger level [1-5]
Datum	Datetime of variable observation
elevation_th	Elevation of station [m]
HN72_24	3 d sum of daily height of new snow [cm]
HN24_7d	7 d sum of daily height of new snow [cm]
Pen_depth	Skier penetration depth [cm]
HN24	24 h height of new snow [cm]
min_ccl_pen	Min critical cut length at a deeper layer of the penetration depth [m]
RH	Relative humidity [–]
wind_trans24_3d	3 d wind drift [cm]
warnreg	Warning regions [1-130]
wind_trans24	24 h wind drift [cm]
wind_trans24_7d	7 d wind drift [cm]

Figure 16 Variable Description

8. Legend to Data Sets on Github:

Original data set source: https://www.envidat.ch/dataset/weather-snowpack-danger_ratings-data

df_p1 -- df_p8 : 2/3 of original data after removing NaNs

df_t1datum and df_t2datum: Grisons subset including the column "datum"

df_t1new and df_t2new: Grisons subset excluding the column "datum"

9. Bibliography

Danger levels - SLF. (n.d.). https://www.slf.ch/en/avalanche-bulletin-and-snow-situation/about-the-avalanche-bulletin/danger-levels.html

Gefahrenstufen - SLF. (n.d.). https://www.slf.ch/de/lawinenbulletin-und-schneesituation/wissen-zum-lawinenbulletin/gefahrenstufen.html

Total Snowfall for Switzerland in November - Current Results. (n.d.). https://www.currentresults.com/Weather/Switzerland/snowfall-november.php

Skiing: Tips & Offers | *Graubünden Tourism*. (n.d.). https://www.graubuenden.ch/en/activities/winter/skiing (For Cover Picture)

Zweifel, Z., B., Pielmeier, P., C., Techel, T., F., Marty, M., C., & Stucki, S., T. (2021). Schnee und Lawinen in den Schweizer Alpen Hydrologisches Jahr 2020/21. *WSL Berichte*, *Heft 116, 2021*(ISSN 2296-3456), 104. https://www.dora.lib4ri.ch/wsl/islandora/object/wsl%3A29347/datastream/PDF/Zweifel-2021-Schnee_und_Lawinen_in_den-(published_version).pdf#page=104

Weather, snowpack and danger ratings data for automated avalanche danger level predictions. (n.d.). Envidat. Retrieved November 11, 2022, from https://www.envidat.ch/dataset/weather-snowpack-danger_ratings-data