Master thesis

Artificial intelligence in ski touring

Prediction of *caution*- and *foot*-sections for

ski touring routes with a machine learning approach

Submitted by

Claudio Furrer

Klosterstrasse 23, 6003 Luzern

claudio.furrer@stud.hslu.ch

Spring semester 2024

Submitted May 10

MSc in Applied Information and Data Science

Lucerne University of Applied Sciences and Arts

Master thesis

Artificial intelligence in ski touring

Prediction of *caution*- and *foot*-sections for ski touring routes with a machine learning approach

Submitted by

Claudio Furrer
Klosterstrasse 23, 6003 Luzern
claudio.furrer@stud.hslu.ch

Supervised by

Martin Rumo
Zentralstrasse 9, 6002 Luzern
martin.rumo@hslu.ch

On behalf of

Skitourenguru GmbH
Markusstrasse 12, 8006 Zürich
represented by Günter Schmudlach
schmudlach@gmx.ch

Spring semester 2024
Submitted May 10

MSc in Applied Information and Data Science
Lucerne University of Applied Sciences and Arts

Abstract

This master thesis is written on behalf of Skitourenguru and deals with the prediction of *caution*- and *foot*-sections of ski touring routes with a machine learning approach. The ski routes dataset maintained by the Swiss Alpine Club (SAC) and swisstopo was sampled with additional static terrain data. A route section is marked as *caution* in the data if the tour is particularly dangerous at a certain part. It is marked as *foot* if the tour has to be travelled without skis. On the one hand, the work aims to predict the *foot*-sections on new, unseen routes. National cartographic authorities may use the model as an aid for marking the *foot*-sections. On the other hand, the man-made *caution*-sections in the current SAC data can be challenged for their consistency. The data, the code and the results are published in a GitHub repository and can be used by interested individuals under the defined licence.

Machine learning models are trained and evaluated using the SAC data enriched with additional geodata. The focus lies on logistic regression models and general additive models (GAM), as these are transparent and relatively easy to communicate to stakeholders. Black box models such as random forests and gradient boosting are trained, whereby these are to be understood as benchmarks and cannot be considered as one of the final models. Furthermore, the focus is not on intensive hyperparameter tuning, but rather on extensive feature engineering, which is carried out with the help of exploratory data analysis and domain knowledge. The domain knowledge is based on a literature research on risks in ski touring. In addition to that, an interview with the individual responsible for the SAC tour portal closes gaps in the literature and provides valuable information on the SAC dataset.

It is described in the interview that the SAC data has been adapted over decades, whereby the classification is not based on strict rules, but is carried out subjectively by different mountain guides. This indicates possible noise in the data. It was also mentioned that the *caution*-sections are often marked in branches, which leads to potential class noise, as these sections tend to be marked too generously. In the visual data analysis, this class noise is particularly evident for the target variable *caution*. Accordingly, data points labelled as *caution* are visible, although they are located in harmless terrain. There occur also data points that are located in dangerous terrain but are not labelled as *caution*. Similar anomalies are also visible for the *foot*-sections, although the class noise appears much less pronounced for this target variable. Another difficulty is the fact that the class of interest is strongly underrepresented in the data (imbalanced data). The exploratory data analysis shows that the terrain features *crevasse* (crevasse zones on glaciers), *fd_risk* (risk of falling) and *ti* (avalanche terrain indicator) are the most promising for modelling *caution*. For the modelling of *foot*, the most promising terrain features are *crevasse*, *ele* (elevation), *fd_risk* and *fold* (folds and edges in the terrain). Because of the class noise in the target variables, a 'ti-filter' is applied to the modelling for *caution* as part of feature engineering, which reduced the class noise to a certain degree (see chapter 5.2.3). For the modelling of *foot*, a 'tunnel-filter' is applied for the same purpose (see chapter 5.2.3). These two filters lead to better results in the *caution*- and *foot*-modelling. The main.py script is written for the modelling part in such a way that it

could be used for modelling both *caution* and *foot*. Different feature engineering methods (scaling, oversampling, undersampling) and filters (ti-filter, tunnel-filter, street-filter) could be adapted, allowing experimentation with different parameters. The variable *ti* was used both as a filter and as a feature when modelling *caution*, which is why the filter values were chosen very carefully (see chapter 5.2.3). Most promising for the *caution*-modelling is the run where the ti-filter was applied on the imbalanced data and points on streets were excluded. For the *foot*-modelling, it is the run where the points in tunnels were excluded from the imbalanced dataset. In addition to the common metrics in machine learning (accuracy, precision, etc.), the metric *confusion score* (see chapter 4.3) has been defined by the client.

The first research question deals with the prediction of *foot*-sections. For the evaluation of the model, it is agreed with the client that the model must be balanced, i.e. the evaluation takes place at a *p*-threshold where the number of false positives correspond to the number of false negatives. The winning *foot*-model is a GAM with the features *ele*, *fd_risk* and *fold*, with a smoother applied to the last variable. It achieves a confusion score of 142. The overall model performance is promisingly high with an accuracy of 0.98, but the model struggles to reliably predict the class of interest with a precision of 0.59. If the *p*-threshold is adjusted, the precision can be increased at the expense of recall (see chapter 6.1). The model is then no longer balanced, which is why the trade-off must be weighed up with the help of expert knowledge. If more false positives are accepted (increasing recall at the expense of precision), the model could mark a conservative preselection of *foot*-sections, which then needs to be validated by experts. The consistency of the markings could be greatly increased with such an approach.

The second research question is about the extent to which the current consistency of the *caution*-sections in the SAC data could be challenged. The class noise is significantly stronger due to the branch-wise markings of the *caution*-sections, which leads to poorer model performance. The winning GAM with the variables *crevasse*, *fd_risk* and *ti* (with a smoother on *ti*) has an accuracy of 0.93 at the mentioned *p*-threshold, but also struggles to reliable predict the class of interest with a precision of only 0.39. Here, as well, precision can only be increased by entering into a trade-off. The resulting confusion score is 308. Nevertheless, the final model largely confirms the findings from the literature and the interview, and exciting insights can be obtained. For example, by considering a one-dimensional logistic regression model to determine the boundary values for the variables used in the winner model. If the variable *crevasse* is greater or equal 5, a route point is *caution*. For the variable *fd_risk*, the route point is *caution* if it has a value of 1’079 or higher. For *ti* from a value of 0.66. With these boundary values and expert knowledge, a consistent rule-based and transparent marking of the *caution*-sections could be introduced.

Even if the models are not yet perfectly precise, more accurate models could be trained using further noise reduction techniques (e.g. DBSCAN) in combination with expert knowledge. The estimation of sections (instead of points) could further smooth out outliers, where first attempts have already delivered promising results. Nevertheless, noise reduction should be carried out carefully and with the help of experts in order to not introduce a bias into the model by altering the ground truth.

Table of contents

1 Introduction

1.1 Background

In Switzerland, more than one million people are being injured and 2'400 people die every year in non-occupational accidents. Therefore, accidents are a challenge for public health. In addition to the suffering and the pain for those affected, these accidents cost 12 billion Swiss francs annually. For those reasons, accidents should be prevented as far as possible. Nowadays, 6% of deaths in leisure time are linked to sports activities. Taking injuries into consideration, sport makes up a decent share of 39% as the cause (BFU, online). Therefore, risky sports such as ski touring must be given special attention. On average, 17 deadly ski touring accidents occur every year in Switzerland (BFU, online). Even this year there were major incidents such as the accident on the Tête-Blanche, where five ski tourers died and one is still being missed (NZZ, online). Avalanches claim three out of five lives in the mountains in winter, making them the most common cause of death in ski touring (Winkler et al., 2023).

As with any sports, an athlete is exposed to various risks during ski touring. In ski touring, one is exposed to risks that have small event probabilities but can potentially have very serious consequences and very large uncertainties. The platform www.skitourenguru.ch (in the following Skitourenguru) starts exactly there. Skitourenguru assigns avalanche risks to thousands of ski tours in the Alps on a daily basis: green for low risk, orange for increased risk or red for high risk. To do this, a high-performance computer uses a digital terrain model to calculate a map of potential avalanche terrain. Each time new avalanche situation reports are issued, the risk indicators for the routes shown are then calculated fully automatically. Skitourenguru is therefore very much involved in the selection and the planning of the tours. But the self-responsibility and assessment in the mountains is left to each tour participant themselves, meaning Skitourenguru has no influence on this matter. It is mainly the planning part where Skitourenguru can contribute to a better risk management (Schmudlach, online).

As a part of the external elective module *SAS Analytics* of the master's curriculum, the author of this master thesis came across the tool Skitourenguru during a presentation by Ulrich Reinke who works for the SAS Institute. As a ski touring beginner and mountain sports enthusiast, the author was excited about Skitourenguru. He therefore explored the platform and contacted Günter Schmudlach (the founder of Skitourenguru) shortly after regarding the possibility of writing a master thesis for Skitourenguru. Not much time had passed until Günter Schmudlach and the author met physically for the first time in June of 2023 and were able to set the framework for this master thesis. For this reason, the author is personally very interested in the topic and excited about the upcoming master thesis.

1.2 Problem statement

Skitourenguru is a platform that can be used in planning of the ski tours, where an athlete may use it as a tool for estimating the difficulty of a ski tour. The Swiss Alpine Club (SAC) together with swisstopo provide a dataset with a large number of ski touring routes in Switzerland. In this dataset, the marked routes are divided into *normal*-, *caution*- and *foot*-sections. The subdivision of the sections is man-made and there are no uniform criteria defined by the SAC. With the help of these sections, athletes are able to select the right spot where they should take off their skis and climb the summit by foot (*foot*-section). Or in which parts of the route they should be particularly careful (*caution*-section). Skitourenguru on the other hand can use the information on these sections and include them in the evaluation of the difficulty of the ski tour or display the segments and ski depots. The problem occurring is that these *foot*- and especially the *caution*-sections in Switzerland are not defined with standardised criteria, but are defined more or less arbitrarily by the mountain guides of the corresponding SAC sections. In some countries the *foot*-sections are very imprecise or not marked at all.

1.3 Research questions

This master thesis is dedicated to the prediction of these *caution*- and *foot*-sections (target variables) for new, unseen ski routes. The purpose of the models trained on the Swiss geo data is to be applied everywhere in the Alps, where the *caution*- and *foot*-sections are not yet defined and therefore need to be predicted. Finally, the *foot*-model can be used by national cartographic authorities, where the *foot*-sections can later be officially integrated into the ski touring routes of these countries. Moreover, the sub question arises as to whether the currently non-rule-based drawings by the SAC of the *caution*-sections in Switzerland can be drawn in a more rule-based with the help of the trained *caution*-model. This leads to the following research questions:

- **Main research question**: Which predictive models can be built and trained using geospatial data as features (input variables) to predict *foot*-sections provided by the SAC?
- **Sub research question:** To what extent can predictive models challenge the consistency of the human-drawn markings by the SAC of the *caution*-sections in Switzerland?

1.4 Thematic limitations

The development and the implementation of a concrete framework (e.g. pre-selection of *foot*-sections by algorithm, validation of these sections by experts, consideration of trade-offs) is not dealt within this paper, i.e. the focus lies entirely on the development and training of the models. Another limitation is the focus on generalized linear models (GLM) and generalized additive models (GAM). This is for the simple reason that these tend to be easier to interpret and communicate to stakeholders. Non-linear *black box* models such as random forests or gradient boosting are included in the work but only serve as a benchmark for the fitted GLM and GAM. Because the tuning of hyperparameters in random forests and gradient boosting on large datasets is computationally expensive, only a manageable selection with a few different hyperparameters was trained during the modelling. Thus, the focus of the modelling part lies much more on extensive feature engineering and exploratory data analysis in order to incorporate domain knowledge rather than extensive hyperparameter tuning.

1.5 Publicity of the results

It is important for the client that the knowledge obtained, and the created model can be used and further developed by the community interested in ski touring and public stakeholders. For this reason, the master thesis and the fitted model will be published in the GitHub repository `RoutesProperties` (details in the appendix) under a *creative commons* [1] license. It allows creators to share their work while retaining certain rights and control over how the work can be used by others without the need for individual, case-by-case permissions. The license terms can be deducted directly from the GitHub repository.

1.6 Structure of the thesis

The following master thesis is divided into the following parts: In chapter 2, the basic knowledge of ski touring and avalanche risks is reviewed on the basis of literature and an interview. In chapter 3, the raw training data and its sampling process is explained and analysed in more detail. In chapter 4, the methodology for training, evaluating, and selecting the machine learning models is described. In chapter 5, feature engineering is discussed in detail and the data is analysed. In addition, in this chapter the models are trained and evaluated, where finally a winner model for each of the target variables *caution* and *foot* will be defined. In the last chapter, the results and the winning models are discussed with regard to the research questions as well as the limitations of the work.

[1] https://creativecommons.org

2 Theoretical framework

In this chapter, the current state of practical and scientific knowledge of risk in ski touring is reviewed. The review addresses managing uncertainties and discusses specific risks in ski touring, with a particular focus on avalanches. Potential risk predictors are being explored for modelling the *caution*- and *foot*-sections. The features are mainly to be found in the literature on probabilistic avalanche science. Literature on the analytical method, which deals with assessing the situation on site, is only of limited help in finding suitable, static terrain predictors. Nevertheless, it provides information on the locations where weaknesses in the snowpack are more likely to occur, and therefore potential features can be derived. In addition to the literature, an interview is conducted with Andreas Eisenhut, the individual responsible for the touring portal of the SAC, who provides valuable insights into the swisstopo dataset, particularly with regard to the data quality and the historical classification of *caution*- and *foot*-sections by the SAC.

2.1 Literature review

2.1.1 Dealing with uncertainties

On a ski tour, an athlete is always confronted with *yes- or no-decisions*. While dealing with uncertainty, randomness and complexity plays a key role in safety. Recognition, decision-making and behaviour in risk situations depend highly on cognitive, emotional and social factors. It has been proven that risk-conscious people cause fewer accidents than those who supposedly believe to have things under control. Additional knowledge and experience is therefore often overcompensated by taking much greater risks (Munter, 2023). In behavioural psychology, a distinction is made between two human thought systems. System 1 works automatically and quickly, largely effortlessly and without voluntary control (intuitive thinking). System 2 focusses attention on rigorous mental activities and complex calculations (deliberative thinking). When dealing with uncertainties, being aware of these two systems is important (Kahneman, 2012). In this context, the irreproducibility of decisions that are made under the same conditions must be acknowledged. Due to the inherent random variability in human decision making process, inconsistent results are a consequence, even if the same individuals are being exposed to the identical information. Compared to rule-based systems (e.g. algorithms), there is a lot of noise in human judgements. Therefore, in many cases it could be advantageous to rely on clear rules and algorithms to ensure consistency and objectivity, especially when it comes to fact-based and repeatable decisions (Kahneman et al., 2021). This emphasises on the importance of dealing with uncertainties in serious planning of a ski tour with the aid of the relevant tour information and the use of deliberative thinking. A rule-based classification approach for the *caution*- and *foot*-sections can make the planning part more objective and highlight the critical sections of a route, so that ski tourers are compelled to be particularly careful at these points of their tour and be aware to not only rely on System 1.

The greatest and deadliest risk in ski touring is being caught in an avalanche (Harvey et al., 2023). To be precise, avalanches claim three out of five lives in the mountains in winter (Winkler et al., 2023). In the context of avalanche risk management, Werner Munter is omnipresent in the modern avalanche literature. In the 90s, he revolutionised the perspective on risks involved in ski touring with his own developed method, which he explained in the book *3x3 Lawinen*. The 3x3 method is a systematic, quantitative and rule-based approach for the assessment of avalanche risks and for dealing with uncertainties, which inherently forces a ski tourer to use Kahneman's System 2 (conscious and thoughtful thinking). With his new probabilistic approach to avalanche risk assessment, Munter challenged the widespread practice of relying primarily on subjective judgement (analytical method). The intuitive assessment on site can be compared to a certain extent to Kahnemann's description of System 1 (intuitive thinking) with corresponding noise in the decision making.

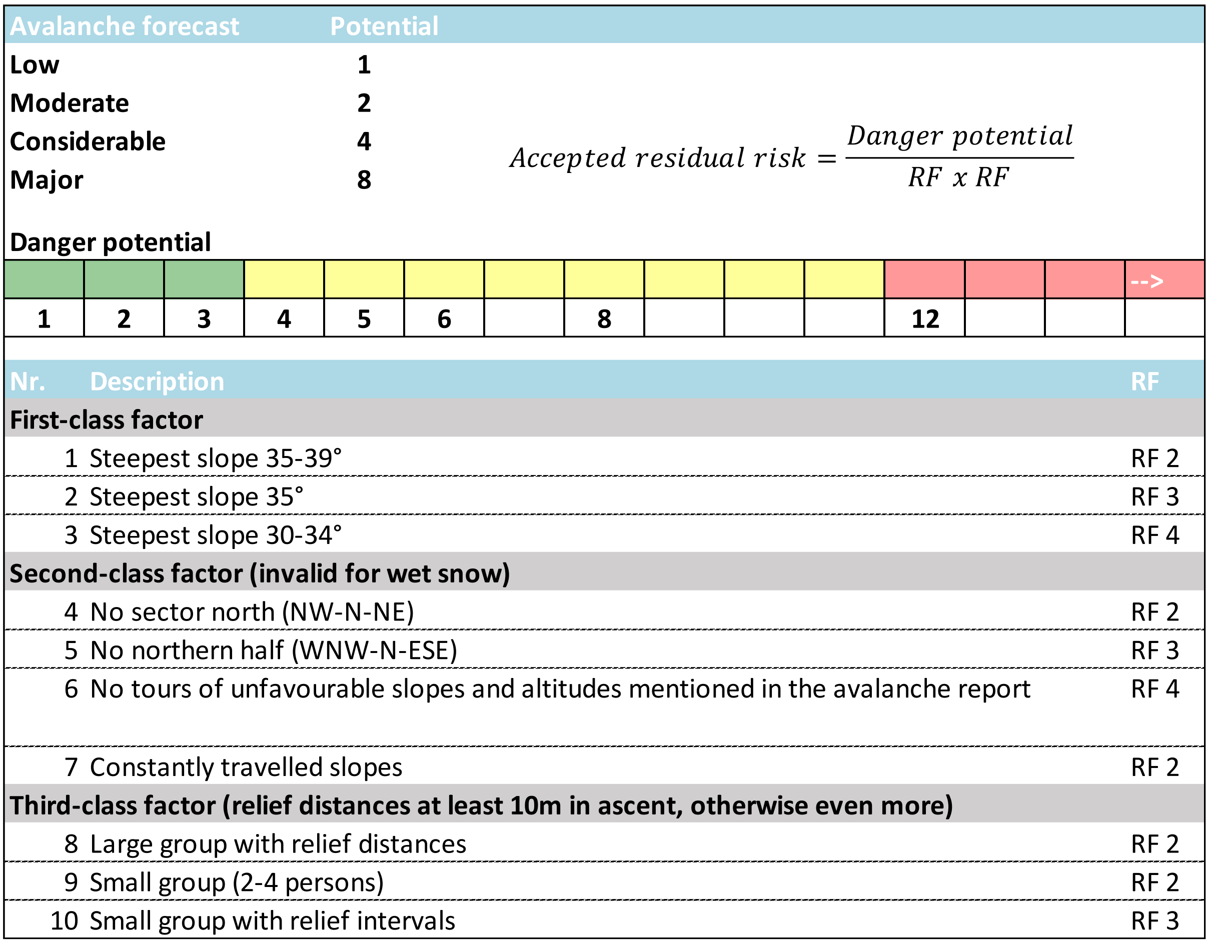
Munter's rule-based method forced a paradigm shift with a significant reduction in fatalities. In his book, he states that man-made snow slabs are responsible for the vast majority of avalanche accidents. He declares the slope gradient as an important danger, since snow slabs can be triggered from a slope gradient of 30° and steeper. He also draws attention to the very large uncertainties of local snowpack analyses, as the composition and load-bearing capacity of snowpacks on the plain differ greatly from those on the slopes. In his opinion, the snow structure of a slope resembles a patchwork quilt, which is randomly composed of strong and weak layers over time. Interpolation of snow layer stability is not reliable in the vast majority of cases due to the lack of homogeneity. Munter was widely criticised by the scientific community and only received recognition when his method was retrospectively proven to be effective (Munter, 2023). In concrete figures, human-triggered snow slabs are responsible for over 90% of avalanche deaths (Winkler et al., 2023).

2.1.2 Risk management on ski tours

The advantages of rule-based risk management seem obvious from the previous chapter, especially with regard to human noise in the decision-making process. Subsequently, methods for risk management on ski tours are examined in more detail, in which first features for the prediction of the passages with *caution* can already be derived from these methods.

An important part of risk management on ski tours is the avalanche situation report. The danger levels 1 to 4 are relevant for the ski tourers (level 5 is very rare). The considered factors of the danger level cover properties of the analytical avalanche science such as snowpack stability, distribution of snowpack stability (frequency of weak layers) and the maximum size of the expected avalanches. The avalanche danger increases exponentially between the individual danger levels. The forecast is subject to uncertainties as there can be deviations within a region (Harvey et al., 2023).

To get a feeling of possible dangers on ski tours, it is worth taking a look at Munter's reduction method (see Figure 1). The method is based on the fact that traditional (intuitive) avalanche assessment on site was a subject to major uncertainties. Instead, Munter favoured a probabilistic approach. His aim was to develop a method in which a tourer does not have to analyse the snowpack on site, but rather takes into account topographical terrain parameters such as slope, aspect and altitude. According to Munter, a look at the snowpack can be worthwhile, but this should never justify a *yes* if the other indications result in a *no* (Munter, 2023). An avalanche accident analysis over the period from 2013 to 2020 confirmed the effectiveness of probabilistic methods, where 85% of avalanche accidents could have been avoided (Fleischmann et al., 2021).



Avalanche forecast	Potential
Low	**1**
Moderate	**2**
Considerable	**4**
Major	**8**

$$Accepted\ residual\ risk = \frac{Danger\ potential}{RF\ x\ RF}$$

Danger potential

🟩	🟩	🟩	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟥	🟥	🟥	-->
1	**2**	**3**	**4**	**5**	**6**		**8**				**12**			

Nr.	Description	RF
First-class factor		
1	Steepest slope 35-39°	RF 2
2	Steepest slope 35°	RF 3
3	Steepest slope 30-34°	RF 4
Second-class factor (invalid for wet snow)		
4	No sector north (NW-N-NE)	RF 2
5	No northern half (WNW-N-ESE)	RF 3
6	No tours of unfavourable slopes and altitudes mentioned in the avalanche report	RF 4
7	Constantly travelled slopes	RF 2
Third-class factor (relief distances at least 10m in ascent, otherwise even more)		
8	Large group with relief distances	RF 2
9	Small group (2-4 persons)	RF 2
10	Small group with relief intervals	RF 3

Figure 1: Elementary reduction method (ERM) (Munter, 2023, p. 126)

Munter uses the method to target his key variables of danger potential (avalanche forecast), gradient (steepest point), slope exposure (shadiest point), number of tracks and group size. An example for the reduction method is a danger potential of 8 (Avalanche report *major*), whereby two risk-reducing factors (1 x RF2, 1 x RF3) were taken. This results in a danger potential of $\frac{8}{2\ x\ 3} = 1.33$. The calculation tells us how accident-prone the selected combination is compared to the accidents in the past.

2.1.3 Risk predictors

Based on the above analysis of the reduction method from Munter and further literature, possible features for the dangers in ski touring are now defined and discussed in detail. In particular, the book *Lawinen. Verstehen, beurteilen und risikobasiert entscheiden* by Harvey, Rhyner and Schweizer is an important source since the authors are employed by the renowned *Institute for Snow and Avalanche Research (SLF)* in Davos.

Slope

Avalanche accident figures clearly demonstrate that most accidents can be attributed to snow slabs (around 95% of accidents). Moreover, they indicate that the snow slabs are almost always caused by humans. Other types of avalanches such as wet snow avalanches, sliding snow avalanches or loose snow avalanches are also a potential danger to ski tourers, but are usually triggered spontaneously and are generally not caused by skiers (only around 5% of accidents). Therefore, attention must be paid to the slab avalanches. These can only be triggered if the slope is steeper than 30°, usually between 35° and 45°. It is important to note that snow slabs can also be triggered remotely (Harvey et al., 2023). Avalanche accidents occur most frequently on terrain that is around 36-40° steep (Schmudlach, 2022). This observation is also supported by Munter's rule of thumb, where the most important principle is not to tour above 39° slope when the danger level is *moderate*. At *considerable*, a tourer should not go above 34° steepness and at *major* not above 30°. If this rule had been applied, two thirds of fatal avalanche accidents would have been prevented (Munter, 2023). It can be derived from the literature that the slope is an important feature for determining *caution*. This feature may also be important for *foot*-sections, since it is hardly tourable above a certain steepness.

Aspect

Munter refers in his book to his analysis of snowpacks where he states that 50% of all weak snow layers were found in the cumulative sectors north-west (NW), north (N) and north-east (NE). In his reduction method, the aspect of these sections is also taken into account via a lower reduction factor (see Figure 1). In the southern half (S), on the other hand, only 25% of the weak spots were found. His own analysis correlates very well with the number of accidents (Munter, 2023). It is also known from analytical avalanche research that north-facing slopes are more susceptible to weak snow layers, as the snowpack hardens less well due to the lack of sunlight. Weak layers are therefore particularly durable in old snow. In addition to that, drift snow (snow transported by wind) and surface frost are particularly problematic on north-facing slopes (Harvey et al., 2023). The literature shows that exposure in the north tends to be more dangerous than in the south, thus this characteristic could be important in terms of *caution*.

Elevation

Munter emphasizes in his book on the fact that the avalanche situation report often divides the touring area into favourable and unfavourable slopes. Slopes above 2'000 meters altitude with exposures from northwest to north to east are often unfavourable (Munter, 2023). This division originates from the fact that the altitude favours the snow cover through different temperatures, particularly during precipitation, when it rains below and snows above. It can be observed that at altitudes below 1'800 meters, the snow temperatures in the Alps are rarely low for long periods of time, which results in weak layers generally forming less frequently below 1'800 meters since they solidify quicker (positive effect on snowpack stability). As the opposite to that, the snowpack at altitudes above 1'800 meters solidifies worse, which makes the slopes more prone to weak layers and therefore more dangerous (Harvey et al., 2023). The altitude can be roughly divided into the following three zones, with different snowpack tendencies and therefore different implications for the risk (depending on the climate, the tree line lies between 1'800 and 2'200 meters):

Altitude categories		
1	**Below the tree line**	In wooded terrain, the snow cover is often variable and the wind has little influence. In open fields and sparse forests, the snow cover is evenly distributed. Air temperatures at this altitude often fluctuate around 0 degrees Celsius, which has a positive effect on the build-up of snow cover. Conditions are critical, especially at very low temperatures. Risk of sliding snow at warmer temperatures.
2	**Large, homogeneous slopes between the tree line and around 100 metres in altitude below the crests or passes**	At this altitude, the terrain is often not very rough. There is also little wind erosion and the snowpack characteristics are often similar across the board. If the structure of the old snowpack is unfavourable, large avalanches are triggered, especially at this altitude.
3	**Summit, ridge or pass locations and up to 100 metres in altitude below**	Pronounced terrain irregularities and predominantly strong winds repeatedly lead to wind-eroded snow surfaces, but also to drift snow. The snow cover can erode in a confined space or over a large area. Even small avalanches can lead to dangerous falls at this altitude. Drift snow or fresh snow is often the main problem. Large avalanches are particularly likely after heavy snowfall.

Table 1: Altitude categories (Harvey et al., 2023, p. 193-194)

The analysis of the terrain characteristics on the base of the tree line (see Table 1) also underlines the generally higher risk above around 1'800 to 2'200 meters, which results from lower temperatures, more frequent wind loads but also a higher risk of falling.

Risk of falling

In high alpine terrain, it is not uncommon to find a cliff at the end of a slope. The literature also describes it as the *terrain trap*. In this case, triggering a snow slab not only leads to a potential burial, but also to a potential fall with fatal injuries (Harvey et al., 2024). After avalanches, fall accidents (e.g. over a cliff) are the second most common cause of death in the Alps during the winter. One in four causes of death can be assigned to fall accidents (Winkler et al., 2023). Thus, the theory attaches particular importance to the risk of falling. For the classification of the *caution*- and *foot*-sections, this feature may therefore be very important.

Terrain fold

The shape of the terrain (e.g. edges) can also be an important predictor, as 40% of avalanche accidents occur on slopes close to ridges, which are prone to drifting snow. Channels are particularly subjected. Valleys require more caution than ridges since the edges of the valleys are particularly steep (Winkler et al., 2023). Additionally, approaching a summit can lead along a ridge that has to be climbed by foot, which makes the terrain fold a promising feature.

Slope size

The size of the slope is also described as a terrain trap in the literature. The bigger the slope, the greater the avalanche can become and the more likely it is for a serious burial to occur. The size of a slope above the ski route is the most important factor for determining the risk of burial (Harvey et al., 2023).

Forest

Munter is being critical of the protective function of a forest, especially in less dense forests. For a forest to have a protective function, it must be very dense. If skiing is possible in the forest or large parts of the sky can be seen from the forest, it tends to provide little or no protection against triggering avalanches (Munter, 2023). However, in addition to potentially serving as protection against avalanches, a dense forest also poses a danger, as mentioned later in the interview.

Traffic

The literature from analytical avalanche science indicates that frequently skied slopes tend to be safer, since the snow cover has already been tested several times for stability. This is particularly true for conditions without new snow. Even with increased danger, a frequently travelled route can be relatively safe if the tracks have not been snowed over again (Harvey et al., 2023). Munter is also convinced that heavily skied slopes are more stable than slopes at the same altitude and exposure that are rarely skied (Munter, 2023). However, since traffic is not a static feature (after a snowfall there are no tracks before the first run), this non-static variable is not suitable for *caution*-prediction of a tour segment, even if it could be measured to a certain extent using modern technologies (e.g. applications like Strava).

Further features

In the analytical avalanche science, there are of course even more features that can be used to predict *caution*. Examples include *wumm*-sounds, group size, snowpack stability (determined by inspecting the snowpack) and so on. However, as analytical avalanche science is strongly focussed on on-site assessment, these features are not relevant for this work since they are not static and therefore cannot be extracted from a map. Therefore, the terrain is the only avalanche-forming factor that does not change over time (Winkler et al., 2023).

2.2 Interview findings

The results of the interview with Andreas Eisenhut, who is responsible for the SAC tour portal, are analysed below. The primary focus is on his view of which criteria (features) have an influence on the *caution*- and *foot*-sections. Additionally, the data quality from the swisstopo data was discussed (see chapter 3.5).

2.2.1 Risk predictors

Generally speaking, Andreas Eisenhut confirms the key features derived from the literature review in chapter 2.1. In his opinion, the most important features for modelling *caution* include slope, risk of falling and crevasse zones of glaciers (Interview, row 347-350). For the following criteria, the knowledge could be extended in addition to the literature.

Glacier

As a large part of the literature had dealt with avalanche science, glaciers were not given much attention in the literature. The results of the interview were able to close this gap. Andreas Eisenhut gave the input that glaciers represent a potential danger, but for him it would be the wrong approach to simply model glaciers in binary terms (*true* or *false*). This is because in the swisstopo dataset, the *caution*-section is mostly shown in the crevasse zones of the glacier. The focus must therefore be placed on the crevasse zones. He created a layer that will be used for sampling the crevasse zones. These zones are relatively static and do not change significantly over time. Even when glaciers melt, the crevasse zones do not move with the glacier. They are mainly static because they do not form randomly on the glacier, but depend on the underlying topography. Narrows, slopes, ridges and other terrain forms lead to stresses and crevasse formations in the overlying glacier (Interview, row 226-269).

Lake

Similar to the crevasse zones, lakes have not been widely discussed in the literature. The interview also provides insightful information here, whereby lakes seem to be a negligible criterion for the modelling of *caution*. Andreas Eisenhut distinguishes between reservoirs and normal lakes, although he does not consider this subdivision necessary for modelling purposes (Interview, row 270-284). The literature confirms that accidents in lakes tend to be rare, but nevertheless represent a danger. Lakes should only be crossed when there is a thick layer of ice and, if possible, routes should be chosen around those (Winkler et al., 2023).

Forest

The interview also provided an interesting perspective on forest density. In the avalanche literature, forestation, especially dense forestation, tends to be a positive, danger-reducing feature. However, Andreas Eisenhut provided a different point of view from a holistic perspective. From his perspective on danger potential (not only avalanche risk), dense forests should rather be considered as *caution* because various risks are associated with those (disorientation, exertion, terrain traps). However, he suspects that in the current swisstopo dataset, dense forests are rarely labelled as *caution*, depending on the author who made the classification (Interview, row 315-334). The influence of the forest density on *caution* in the dataset is therefore very questionable since the relationship may not be evident in the swisstopo data.

Road

Andreas Eisenhut felt that it was important to mention that in most cases a route along a road or a well-developed path is never *caution*. He therefore suggested that this differentiation should somehow be incorporated into the model. His suggestion was to model the entire dataset once, and once only the dataset with the pathless touring terrain (without roads). However, he was not sure whether this was the right approach since he was not familiar with machine learning (Interview, row 90-159).

Rockfall

In the interview, it was also mentioned that feedback from the community was increasingly incorporated into the *caution*-sections (but not yet in the swisstopo dataset). When rockfalls are reported, the area is roughly marked (manually) and the corresponding route section is labelled as such. However, this should not be considered for the thesis, as there is no complete layer for this (Interview, row 285-288).

2.3 Conclusion of the theoretical framework

Both the theory and the interview assign great importance to the slope with regard to *caution*. In the interview, the risk of falling and crevasse zones were named as further, maybe the most important predictors for *foot* and *caution*. Overall, there are many overlaps from the literature and the interview results regarding the avalanche risk. Criteria such as aspect, altitude, forest density and lakes were further mentioned. Additionally, whether the route follows along a road or not also appears to be a reliable predictor for *caution* and *foot*. The following chapter describes the data and the sampling process for further modelling the *caution*- and *foot*-sections.

3 Data

During the kick-off meeting in June 2023, Günter Schmudlach from Skitourenguru provided a solid collection of map data layers on which the route points together with the swisstopo ski touring data will be sampled and later modelled. The present chapter focusses at first on the map data. Then, the swisstopo dataset containing the SAC ski routes is described (target variables). Afterwards, the (re)sampling process is also described in detail. The sampled map data is further described, which contains the terrain features for the prediction (input variables). To increase comprehensibility, map sections of a selection of these layers are also shown. Finally, the data quality is discussed. From this point, variables will be denoted in curly brackets, i.e. {*variable*}.

3.1 Map data

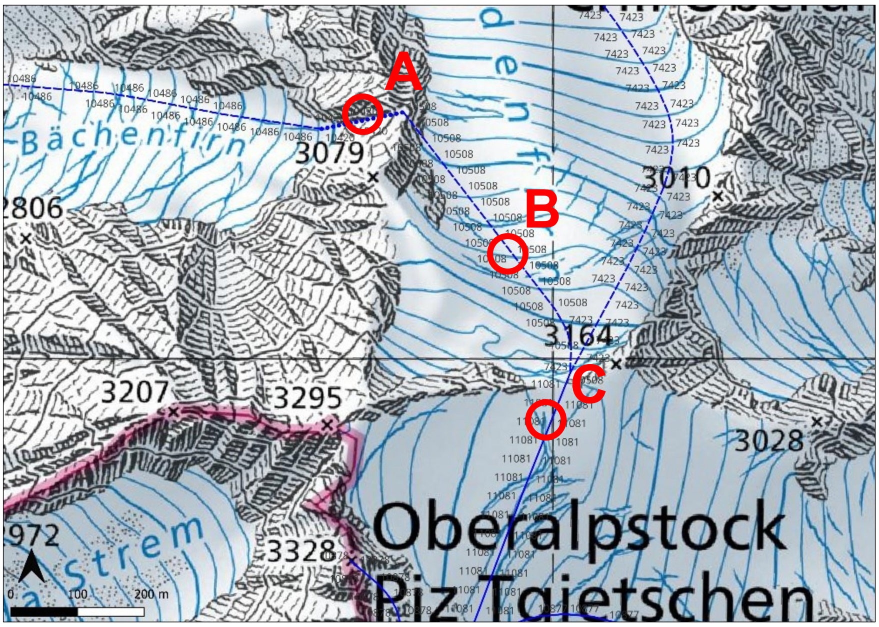
In QGIS[2], a map is deposited as a layer. Since swisstopo publicly provides maps with a very high level of detail (especially in alpine terrain), their map is used as base layer. For the sampling of additional variables, open street map (OSM) is used as well. The map data layers address most of the features identified in the previous chapter. Data regarding whether a route crosses a glacier (including crevasse zones) or a lake is also considered, but has to be sampled. Fortunately, the interviewee Andreas Eisenhut was able to provide a layer for the crevasse zones. The glacier and the lake data had been sampled directly from swisstopo and then combined with the other layers from Günter Schmudlach. For variable selection, it is important that these can be sampled both in Switzerland and abroad. Fortunately, this is the case for Günter Schmudlach's and Andreas Eisenhut's layers. Only then can the main objective of the work, the generalisation to unseen terrain data in France, be achieved.

3.2 Initial ski routes data

SAC operates a tour portal for snow mountain activities. In a shapefile based on LV95[3] (EPSG:2056), which later is reprojected to LV03 (EPSG:21781), SAC provides a variety of ski touring routes in cooperation with swisstopo. The routes used to correspond to the scale of 1:50'000. However, the tour data is now gradually being updated to a scale of 1:10'000. They show the approximate route of the tours. To further analyse this route data, the routes from the shapefile (lines) must be subjected to resampling, where the routes are divided into points (every 10 meters). The following two maps (Figure 2 and Figure 3) illustrate the concept of resampling, where the route trajectories are divided into points. Consequently, these points keep the attributes, i.e. to which section the point belongs. The resampling is performed with QGIS (functions *multipart to singlepart, v. to point*) and a software created by Günter Schmudlach called MapCreator (function *LineResamplingTool*).

[2] Quantum Geographic Information System (QGIS), software for geospatial data analysis

[3] Landesvermessung 1995, geodetic reference system used in Switzerland for mapping, surveying, and cartography



Example

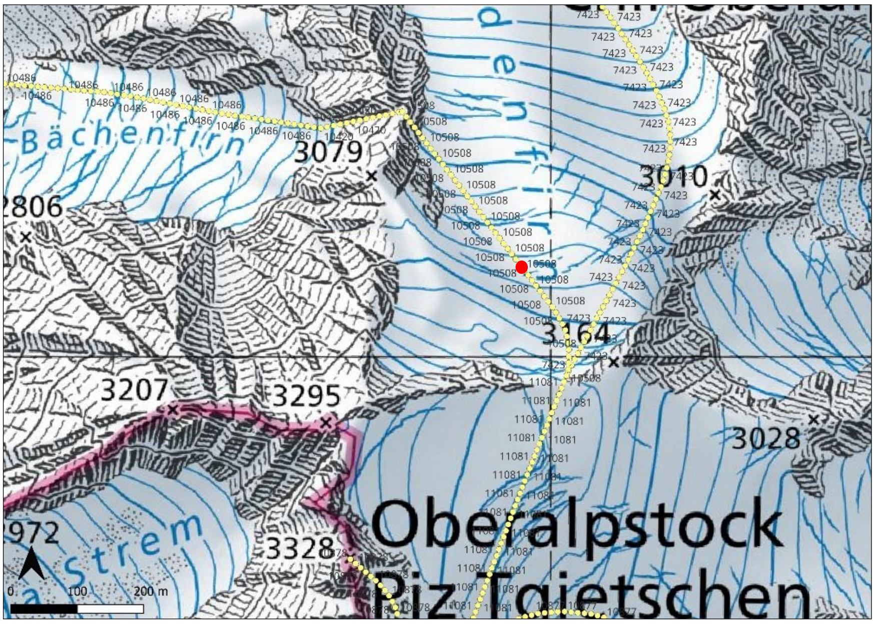
A) ***foot*****-section**
represented by points

B) ***caution*****-section**
represented by dashed line

C) ***normal*** **route**
represented by continuous line

Figure 2: Map with ski routes (line) and different sections

In Figure 2, the ascent after the Bächenfirn leads over a *foot*-section (A – dotted line). The subsequent traverse across the glacier leads over a *caution*-section (B – dashed line) and the descent to the south is a *normal*-section (C – continuous line).



Example

Resampled point(s): The red dot on the left map is just one example. Based on the upper map, the point's attribute *foot* has the value 0 and the attribute *caution* has the value 1. The points can be extended with other georeferenced information in a dataframe and then modelled.

Figure 3: Map with resampled ski routes (points)

As it can be seen in Figure 3, after resampling, the sections can no longer be read out by eye. However, the information is still stored in the attributes of the points. For instance, the point drawn in red would have the value 0 for the binary variable $\{foot\}$ and the value 1 for the binary variable $\{caution\}$.

Variable	Description	Values
ski	The route is a ski tour.	1/0 (True/False)
snowshoe	The route is a snow shoe tour.	1/0 (True/False)
caution	Specific part is a *caution*-section.	1/0 (True/False)
foot	Specific part is a *foot*-section.	1/0 (True/False)

Table 2: Variables from SAC ski touring portal data

In Table 2, the variables from the swisstopo ski routes dataset are listed with their description and characteristics. Only the variables {*caution*} and {*foot*} are relevant for the modelling part, which serve as target variables. In addition, the variables {*ski*} and {*snowshoe*} are only used for filtering, as snowshoe-only tours are not included in the modelling.

3.3 Additional data for resampled ski routes

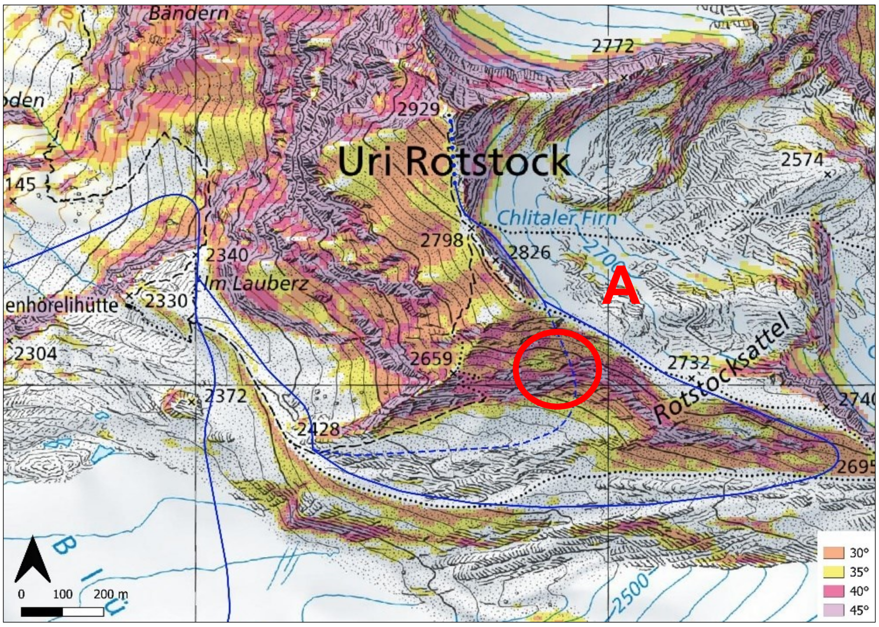
Once the ski touring routes are resampled into points for every 10 meters, the resulting vector can be sampled with further terrain data features using QGIS (function *sample raster values*). Because only one feature layer can be sampled at a time (i.e. {*slope*}, {*fold*}, etc.), a separate layer had to be created for each feature. In QGIS, the feature layers were later merged into a final vector (function *join attributes by location*). The final vector therefore contains the ski routes points sampled with the following terrain features in Table 3:

Variable	Description	Values
aspect	Orientation of the slope.	Decimal (0 to 360) *(later transformed into categorical variable, see 5.1.2)*
country	Indication whether a datapoint is located in Switzerland, Liechtenstein or abroad.	Integer (1 to 2) NA: Abroad 1: Switzerland 2: Liechtenstein
crevasse	If a route runs over a glacier, this variable reflects the extent of the crevasse zone.	Decimal (1 to 7) NA: No crevasse zone, since datapoint is not on glacier 1: Little typical crevasse zone 2: ... 3: ... 4: ... 5: ... 6: ... 7: Very typical crevasse zone
ele	Elevation according to the DEM with 10 m resolution.	Decimal (0 to 5'000 m)
fd	Forest density (in %) with a resolution of 10 m.	Decimal (0 to 1)
fd_maxv	Maximal velocity on a downfall trajectory.	Decimal (0 to 80 m/s)
fold	Slope normal discontinuity raster. The raster shows folds (edges) in the terrain. Calculated from a DEM with 10 m resolution. Negative values indicate concavity (u), positive values indicate convexity (n).	Decimal (-180 to 180°)

forest	Specific point from ski tour lies in forest area.	Binary 1/0 (True/False)
glacier	Specific point from ski tour lies on a glacier.	Binary 1/0 (True/False)
id	Unique identifier for every route point.	Integer
lake	Specific point from ski tour lies on a lake.	Binary 1/0 (True/False)
planc7	The planar curvature calculated from a DEM with resolution 10 m. Negative values indicate convexity (n), positive values indicate concavity (u). The property indicates if a spot is located on a ridge, in a valley or on a homogeneous slope.	Decimal (-300 to 300) *(very seldom outliers above 300 or below -300 were replaced with 300 (-300 respectively), see 5.1.6)*
slope	The slope angle derived from a digital elevation model (DEM) with 10 m resolution.	Decimal (0 to 90°)
street	Category that indicates the distance to the next street. A street is defined as a roadway that can be managed by an agricultural vehicle.	Categorical 1: Next street in distance 0...5 m 2: Next street in distance 5...15 m 3: Next street in distance 15...25 m 0: No street nearby (later transformed into binary variable, see 5.1.2)
ti	Terrain indicator indicates how suitable a terrain point is to trigger an avalanche with minimum required slope angle for release (MRSAR) of 100 m.	Decimal (0 to 1)

Table 3: Additional variables for resampled ski route data

For a better understanding, a selection of the layers is visualised in the following. In Figure 4, the feature {*slope*} is shown. This layer may be particularly important as the feature was described as a promising predictor both in the literature and in the interview. Slopes of 30° and above are emphasised in the illustration, as a snow slab can only be triggered from this degree onwards. There is a similar layer called {*ti*}, which reflects how suitable the terrain is for triggering avalanches with a minimum required slope angle. The {*ti*} layer is made by Günter Schmudlach and also takes the size of the slope into account.



slope layer

A) *caution*-section
Dashed line indicates caution, maybe because of the inherent avalanche risk due to the slope steepness or its maximal fall height?

Figure 4: Slope Layer from Skitourenguru {slope}

Another promising feature from the literature and the interview is the risk of falling. There is no public layer that quantifies the risk of falling. However, Günter Schmudlach has shared his corresponding layer called {*fd_maxv*}. It quantifies the maximal fall down velocity. Figure 5 shows that the Mönch has to be climbed by foot. This is probably due to the risk of falling left and right, the steepness of the route and the rough terrain. The variable {*fd_maxv*} is transformed to {*fd_risk*} in the course of the thesis.

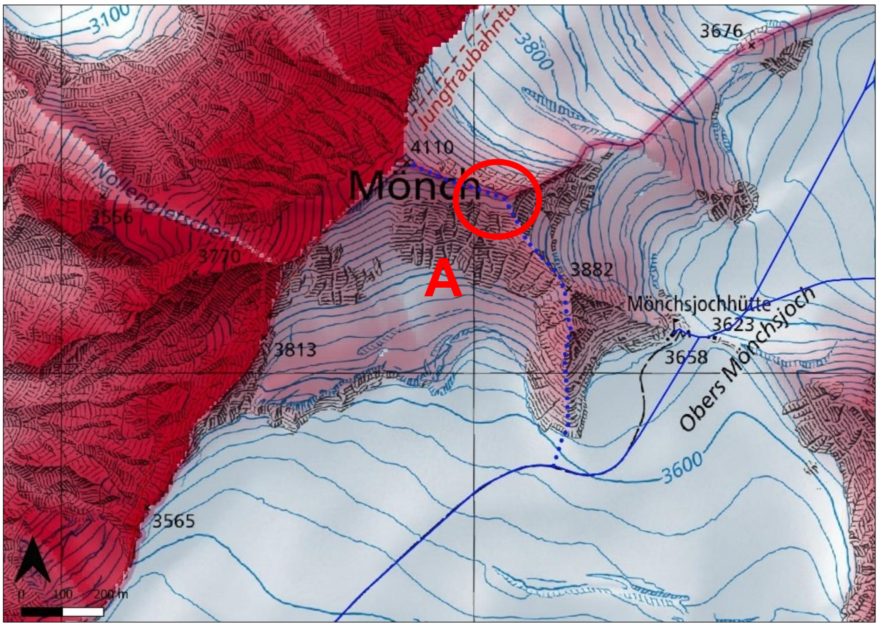


Figure 5: Terrain Max Velocity Layer from Skitourenguru {fd_maxv}

fd_maxv layer

A) *foot*-section
Dotted line (*foot*-section) represents the ascent to the Mönch by foot, maybe because of the maximal fall down velocity?

The crevasse zones were further mentioned in the interview as a potentially important feature for the *caution*-modelling. Andreas Eisenhut's layer contains the crevasse zones, which are scaled in the feature {*crevasse*} from 1 to 7 based on the roughness of the terrain under the glacier.

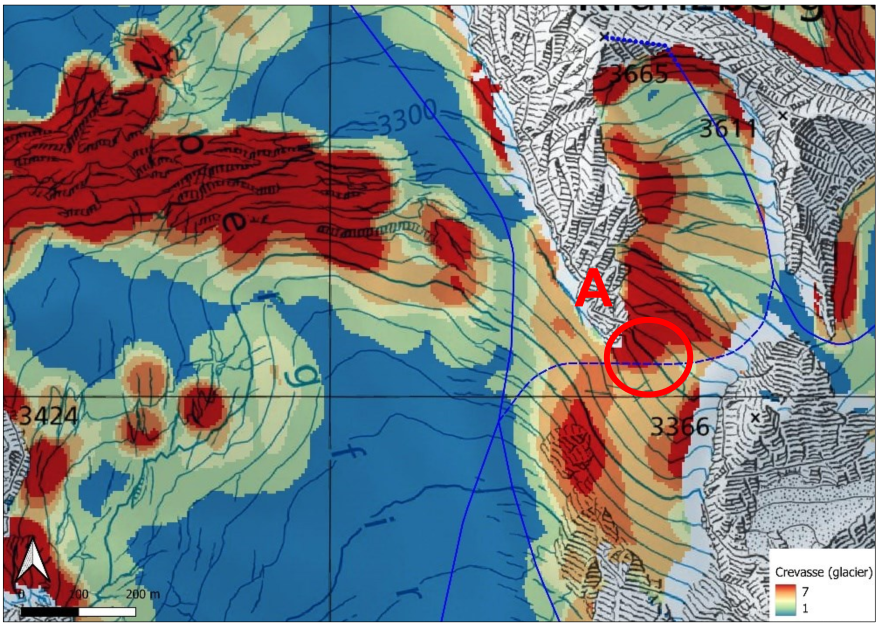


Figure 6: Crevasse Zones Layer from Andreas Eisenhut {crevasse}

crevasse layer

A) *caution*-section
Dashed line (*caution*-section) represents the traverse over a crevasse zone with values up to 7 (very distinct crevasse zone).

The variables aspect {*aspect*}, elevation {*ele*}, forest density {*fd*}, forest {*forest*}, fold {*fold*}, planar curvature {*planc7*}, street {*street*} and avalanche terrain indicator {*ti*} were not visualised in the above figures, but are of course also used in the analysis and the modelling part in chapter 5. The variables country {*country*} and lake {*lake*} are only used for filtering.

3.4 Sampled data

Finally, the vector enriched with additional features mentioned in the previous sub-chapter can now be written to a final geo package database (.gpkg). A comma-separated file (.csv) can be exported from that database, which can be read in with Python and is ready for feature engineering. In the final dataset, a certain autocorrelation exists because the values of a variable in a geographic space are somehow correlated with the values of the same variable in nearby locations (points). If this issue is ignored, spatial autocorrelation may lead to biased model evaluation and inaccurate predictions. Furthermore, the rare event problem (also known as the imbalance problem) should be treated with care. Since the proportion of events {*caution* = 1} or {*foot* = 1}is low in relation to the entire dataset, it may help to ensure that there are enough events in the training dataset. Oversampling the rare case or undersampling the abundant case may lead to a better predictive performance of the trained models (Bruce et al., 2020). The problems of autocorrelation (point distance) and class imbalance are addressed in chapter 5.1. With the transformed and sampled data, different linear models (for deployment) and non-linear models (as benchmark) are developed and trained. The methodological part of the modelling procedure is discussed in more detail in chapter 4.

3.5 Data quality

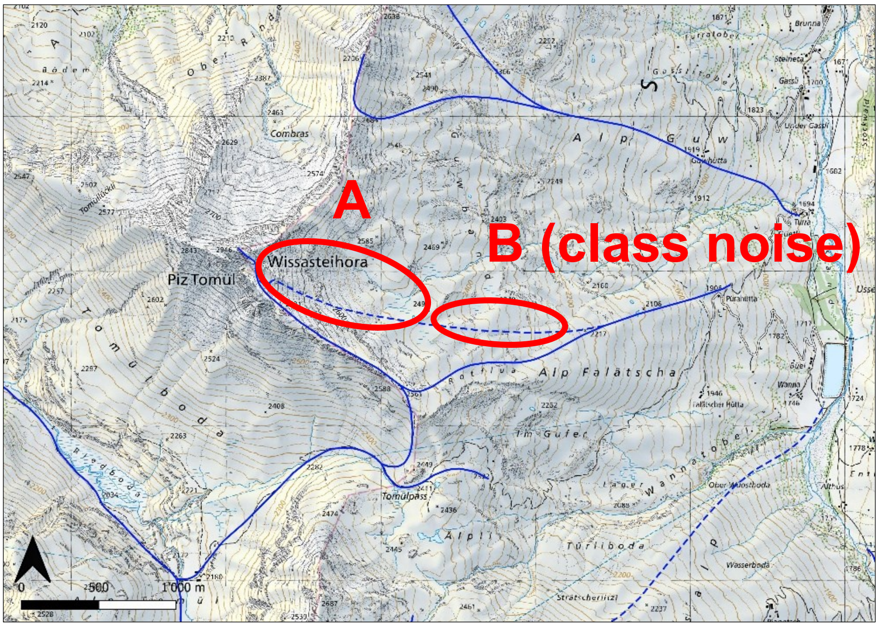
According to the conducted interview with Andreas Eisenhut, there are shortcomings in the data quality of the swisstopo dataset. His statements can be summarised as follows:

Nearby foreign countries

In the past, the route data had been sent by the individual mountain guides to swisstopo, where it was then plotted centrally on the map (at a scale of 1:50’000). At that time, an older software was used, which meant that the markings could not be made as accurately. Although the routes have since been partially revised and plotted more precisely, the data from the neighbouring countries in particular is no longer up to date. When preparing the data, it therefore makes sense to exclude the neighbouring countries due to the lack of precision (Interview, row 168-181). As will be described in chapter 5.1, only points with attribute {*country* = 1} (i.e. Switzerland) are considered for modelling. The areas abroad close to the border were dropped.

Sections of signatures (branch-wise markings)

Andreas Eisenhut mentioned further, that the swisstopo dataset regarding the target variable $\{caution\}$ is often drawn in a branching manner. This means that the marks for $\{caution\}$ tend to be too generous. For example, the descent from Piz Tomül is correctly labelled as $\{caution\}$, but this was drawn for the entire branch of the route. This means that in the part below (after the descent), where the terrain is no longer dangerous, this part of the route to the next junction is still labelled as $\{caution = 1\}$ (although in reality there is no caution needed) (Interview, row 40-80). This inaccuracy in the target variable $\{caution\}$ leads to noise in the data (so called *class noise*), distorts the estimation of the model parameters and also affects the validation (comparison of the predictions with the ground truth). A pragmatic cleaning approach in consultation with Günter Schmudlach is explained in chapter 5.2.3.



Sections of signatures

The dashed line in the centre of the map represents the *caution*-section, here presumably generously drawn from branch to branch.

A) in reality *caution*
The upper part of the tour is correctly classified as *caution.*

B)* in reality <u>not</u> *caution
The lower part of the descent is still labelled as *caution* to the next junction, although in reality this part is probably **not** *caution* according to Andreas Eisenhut ('boring terrain').

Figure 7: Sections of signatures (branch-wise)

From the current point of view, the data quality for the target variable $\{foot\}$ therefore appears to be much better than for the target variable $\{caution\}$. It must also be considered that each data point in the overall dataset is always evaluated with regard to $\{caution\}$ and $\{foot\}$, thus only the following three combinations are possible:

- point has attributes $\{caution = 0\}$ and $\{foot = 0\}$
- or point has attributes $\{caution = 1\}$ and $\{foot = 0\}$
- or point has attributes $\{caution = 0\}$ and $\{foot = 1\}$

There are points like in the third case in the training data for *caution*-modelling. These points are labelled with $\{caution = 0\}$, even if they are in very rough and dangerous *foot*-terrain. In the *caution* modelling, these points also lead to class noise, as they are labelled with $\{caution = 0\}$ (but $\{foot = 1\}$in the background). A solution to this class noise issue is also shown in chapter 5.2.3.

4 Methodology

This chapter describes the methodological approach of the master thesis. The approach is divided into six steps: Problem identification, data preparation, explanatory data analysis, data modelling, evaluation and deployment as shown in Figure 8. Problem identification (1) was dealt within chapters 1 and 2 of the master thesis. The data (2) and the (re)sampling methodology were discussed in chapter 3. The feature engineering and explanatory data analysis (3) is discussed in chapter 5. This chapter also focuses on modelling, evaluation and model selection (4, 5). In the modelling part, various supervised machine learning techniques are used. Chapter 6 summarises the results and draws a conclusion (6).

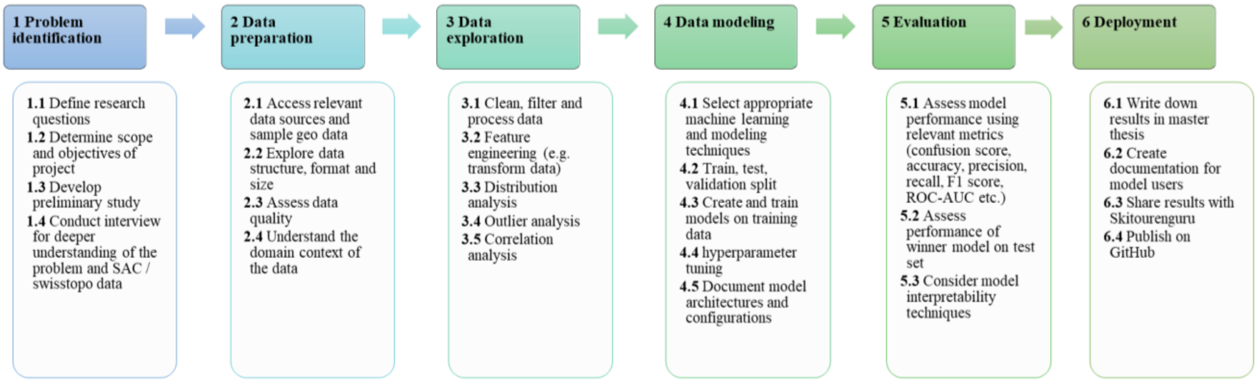

1 Problem identification
1.1 Define research questions
1.2 Determine scope and objectives of project
1.3 Develop preliminary study
1.4 Conduct interview for deeper understanding of the problem and SAC / swisstopo data
2 Data preparation
2.1 Access relevant data sources and sample geo data
2.2 Explore data structure, format and size
2.3 Assess data quality
2.4 Understand the domain context of the data
3 Data exploration
3.1 Clean, filter and process data
3.2 Feature engineering (e.g. transform data)
3.3 Distribution analysis
3.4 Outlier analysis
3.5 Correlation analysis
4 Data modeling
4.1 Select appropriate machine learning and modeling techniques
4.2 Train, test, validation split
4.3 Create and train models on training data
4.4 hyperparameter tuning
4.5 Document model architectures and configurations
5 Evaluation
5.1 Assess model performance using relevant metrics (confusion score, accuracy, precision, recall, F1 score, ROC-AUC etc.)
5.2 Assess performance of winner model on test set
5.3 Consider model interpretability techniques
6 Deployment
6.1 Write down results in master thesis
6.2 Create documentation for model users
6.3 Share results with Skitourenguru
6.4 Publish on GitHub

Figure 8: Schema methodological procedure

4.1 Modelling methodology

4.1.1 Linear models

A separate binary classification model is trained for each of the target variables {*caution*} and {*foot*}. It is important to note that the target variables are binary variables with two levels 0 (*false*) and 1 (*true*). Therefore, the target variables each have two possible outcomes. If such binary classification problems occur, traditional linear models with Gaussian likelihoods are not appropriate anymore. This is because of the fact that the assumption that the target variable follows a normal distribution is inherently not valid any longer for binary variables. Instead, Bernoulli likelihood, where the target variable follows a Bernoulli distribution, that returns a probability of the predicted label to be 1 (or 0), are the better choice (Deisenroth et al., 2020).

To convert the obtained probabilities into binary predictions, a certain threshold has to be selected. A p-threshold of 0.5 would imply that when the model predicts a probability $\{p >= 0.5\}$, it will classify the instance as belonging to the positive class (*true*). If it predicts a probability $\{p < 0.5\}$, the model will classify the instance as belonging to the negative class (*false*). Finding the right threshold requires a proper understanding of the problem. Therefore, the practical implications of false positives (incorrectly classifying 0 as 1) and false negatives (incorrectly classifying 1 as 0) must be considered. For example, regarding a *foot*-section, false negatives (missing a *foot*-section) may have severe consequences, while false positives (incorrectly label a *foot*-section) could cause unnecessary effort by foot on the ski tour. The same applies to the *caution*-section, where false negatives (missing a c*aution*-section) can also have serious consequences. As it is discussed in chapter 5.2.1, the model has to deal with an imbalanced dataset. This means that there are many more {*target variable* = 0} cases than {*target variable* = 1} cases. In consultation with Günter Schmudlach, the goodness of the model can be evaluated at the *p*-threshold where the deviation between false positives (FP) and false negatives (FN) is zero (achieving FP = FN). This implies that the model's errors in predicting positive instances and negative instances are balanced. Finding the right metrics to evaluate the performance of the models at this *p*-threshold is described in more detail later in this chapter.

Generalized Linear Models (GLM)

The following equation represents a multiple linear regression with the error term ε.

$$Y(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

Equation 1: Multiple linear regression model

Without a link function, the right-hand side of the equation above will not produce probabilities. This means that the output will be a linear combination of predictors, but it won't be constrained to the range [0, 1]. Therefore, a link function must be applied, resulting in transforming the multiple linear regression model to a logistic regression model, which is part of the GLM family (Bruce et al., 2020):

$$P\,(Y(x) = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}}$$

Equation 2: Logistic regression model

In Python, there are two common ways to model logistic regression: One approach is to use the library *statsmodels*, which is primarily focused on the statistical analysis, providing detailed summary statistics like p-values and confidence intervals, making it suitable for understanding the statistical properties of the models. Another approach is to use *LogisticRegression* from the library *scikit-learn*, which emphasizes machine learning algorithms and predictive modelling, making it ideal for deploying machine learning models (Medium, online). Since the primary goal is predictive modelling and deploying a machine learning model, *scikit-learn's* focus on machine learning algorithms and predictive modelling makes it the better choice.

4.1.2 Non-linear models

Some machine learning models are considered as *black box models* due to their lack of interpretability compared to GLM. With a General Additive Model (GAM), however, there is a model that lies somewhere in between. A GAM combines the interpretability of linear models with the flexibility to capture non-linear relationships using smooth functions. Non-linear models are able to fit non-linear relationships quite well and are therefore very agile for problem solving (SAS, 2020). GAM, random forests and gradient bosting are shortly described in the following.

General Additive Models (GAM)

GAM are models that are used to model non-linear effects. They extend the framework of linear models to handle non-linear relationships with the help of a smooth function $f_n(x_n)$ between the input variables and the target variable. Although GAM can model complex non-linear relationships, they are relatively simple to explain, which makes them quite powerful (Tanadini, 2023c).

$$Y(x) = \beta_0 + f_1(x_1) + f_2(x_2) \; + \; \ldots + f_n(x_n) \qquad\qquad Y(x) \sim Bernoulli(p(x))$$

Equation 3: General additive model (GAM)

A link function can also be applied to a GAM to model probabilities using a logistic link function.

$$P\ (Y(x) = 1) = \frac{1}{1 + \, e^{-(\beta_0 + f_1(x_1) + f_2(x_2) + \, \ldots \, + f_n(x_n))}}$$

Equation 4: General additive model (GAM) with logistic link function

Unfortunately, GAM is not being implemented in the Python library *scikit-learn*. However, the R package *mgcv* contains a function called *gam()* for fitting a GAM model. Within that formula, smooth terms can be specified. If an input variable is expected to have a non-linear, smooth effect on the target variable, a smoother may help to approximate this effect. The advantage of the *gam()* formula in R is that the edf (estimated degrees of freedom) and therefore the complexity of the smooth term has not to be estimated by the user. If for example the *gam()* function estimates the edf of 2.1 for a given input variable, it implies that this variable has a quadratic effect on the target variable. An alternative to the *mgcv* package in R would be to use the *pyGAM* library in Python. However, this library is relatively young and still under development. For this reason, the GAMs are modelled with *mgcv* in R.

Random forest

A decision tree used for classification problems is called classification tree. It is read from the top down starting at the root node. Each internal node represents a split based on the values of one of the inputs. The inputs can appear in any number of splits throughout the tree. Cases move down the branch containing their input value.

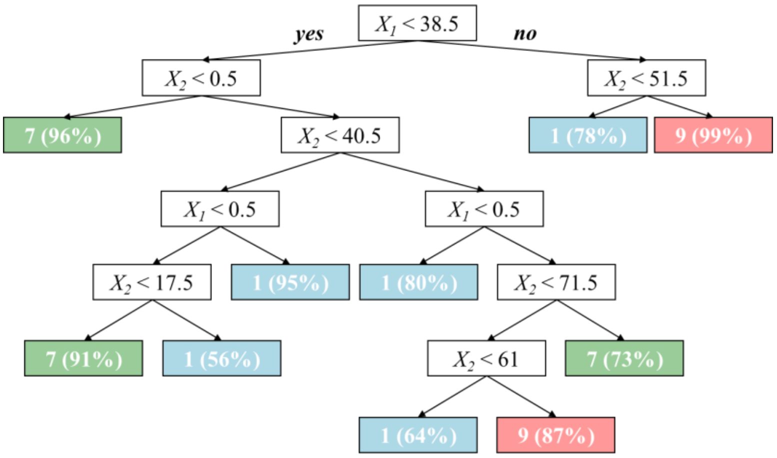


Figure 9: Concept of a decision tree

In a binary tree with interval inputs, each internal node is a simple inequality. A case moves left if the inequality is true and right otherwise. The terminal nodes of the tree are called leaves. The leaves represent the predicted target. The leaves yield the predicted class as well as the probability of class membership. The scheme is shown in Figure 9, which shows an example for handwritten digit classification (only subset of the digits 1, 7 and 9) based on the two input variables x_1 and x_2 (SAS, 2020).

A decision tree, so far, is just a single tree. Instead of growing one single tree, *n* different trees can be modelled. Because decision trees are somewhat unstable models, an ensemble model (random forest) is more robust. A random forest is the combination of multiple decision tree models. Ensemble learning involves combining the predictions of multiple individual models to improve the overall predictive accuracy and generalization. Each tree in the forest is built independently using a random subset of the training data and a random subset of the input features. Each decision tree in the random forest independently predicts a class label for a given input. The random forest aggregates these predictions by majority voting (SAS, 2020). In this master thesis, random forests are modelled with the *RandomForestClassifier* which is part of the *scikit-learn* library. In *scikit-learn*, different parameters can be tried out using *GridSearch* (estimators, max. depth, max. samples, etc.). Only a few parameter combinations were tried out in the modelling due to expensive computing power (long run time).

Gradient boosting

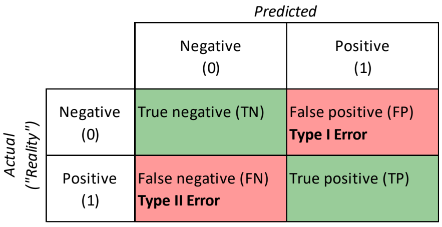
Gradient boosting is a machine learning technique used for both regression and classification tasks. It is an ensemble method that builds a predictive model in the form of an ensemble of weak prediction models, typically decision trees. The key idea behind gradient boosting is to iteratively train new models that predict the residuals or errors of the previous models, thus gradually improving the overall predictive performance. The gradient boosting method can be implemented in Python with the *GradientBoostingClassifier* which is part of the *scikit-learn* library. In gradient boosting, trees are constructed sequentially, with each new tree specifically targeting the errors made by its predecessors. This sequential approach, although challenging to parallelize, often results in superior performance due to the focused correction of previous mistakes. In contrast, random forest adopts an independent construction approach for each tree within the forest. This methodology allows for a simpler parallelization of the training process. Furthermore, trees in random forests are typically shallow, serving to reduce overfitting by capturing simpler patterns in the data (Géron, 2022). For gradient boosting too, only a few combinations were tried out with *GridSearch* due to the high demands on computing power.

4.2 Evaluation methodology

In the model evaluation part, the models have to be evaluated and compared with appropriate metrics (e.g. accuracy, precision, recall, F1 score, ROC etc.). For imbalanced data and the underlying research question, the metrics must be selected carefully. It is not enough for the model to only fit the training data, the predictor needs to perform well on unseen data. The behaviour of the predictor on unseen data can be simulated by cross-validation. The right balance between fitting well to training data and simple explanations of the phenomena have to be found (Deisenroth et al., 2020).

Before the modelling part, the data will be randomly sampled into a training (70%), a validation (20%) and a test set (10%). Random sampling should further reduce the point dependency (similar features of neighbouring points). During cross validation, the training set (70%) will be split into 5 folds. After estimating the parameters on the training set, the performance will be assessed on the validation set (20%) first. The final selected model will be evaluated on the test set (10%). Due to the class imbalance problem of the target variables, the training set is left imbalanced once, oversampled once and undersampled once. However, the validation and test set are always left imbalanced, which reflects the situation in reality. All these tasks can be carried out with help of the functions *train_test_split* (for data split), *SMOTE* (for oversampling), *RandomUndersampler* (for undersampling) and *StratifiedKFold* (for cross-validation) from the *scikit-learn* package.

The performance of the fitted model on the validation and test data can further be assessed with the help of a confusion matrix. As mentioned at the beginning of chapter 4, a difficulty with binary classification is, that one must deal with wrong classifications such as false positives and false negatives. The confusion matrix provides a detailed breakdown of the predictions made by the model compared to the actual class labels in reality (see Figure 10). The matrix can be created directly from the validation data by passing the true labels and the predicted labels to the function *confusion_matrix* from *scikit-learn*. Since the dataset is imbalanced due to only a few cases of {*target variable* = 1}, a large number of true negatives and a relatively small number of true positives, false negatives and false positives can be expected. Carefully choosing the right *p*-threshold and performance metric is particularly important. In the following, different evaluation metrics are discussed in the context of imbalanced datasets.



		Predicted	
		Negative (0)	Positive (1)
Actual ("Reality")	Negative (0)	True negative (TN)	False positive (FP) **Type I Error**
	Positive (1)	False negative (FN) **Type II Error**	True positive (TP)

Figure 10: Confusion matrix

Accuracy

Accuracy is defined as the ratio of the number of correct predictions to the total number of predictions made. It can be misleading when assessing the performance of a machine learning model on imbalanced datasets as mentioned earlier. For example, the accuracy would be high if in a dataset 95% belong to one class and only 5% to the other class. If the classifier assigns all samples to the majority class, then it would get 95% accuracy, which seems high. But the model fails miserably to correctly classify the minority class (i.e. {*caution* = 1} or {*foot* = 1}), which actually would be the class of interest. Despite its high accuracy, the model would therefore be useless (Burger, 2018).

$$Accuracy\ = \frac{\text{True positives} + \text{True negatives}}{(\text{True positives} + \text{True negatives} + \text{False positives} + \text{False negatives})}$$

Equation 5: Accuracy

Precision

Precision measures the accuracy of the model's positive predictions. Improving precision may reduce the number of false positives, which indirectly reduces false negatives. However, it's essential to strike a balance between precision and sensitivity. For the problem with the imbalanced dataset, precision seems to be a more suitable metric than accuracy, but is still not sufficient on its own (Burger, 2018).

$$Precision\ = \frac{\text{True positives}}{(\text{True positives} + \text{False positives})}$$

Equation 6: Precision

Recall (Sensitivity)

Another metric is recall (true positive rate or sensitivity). It measures the ability of the model to correctly identify the positive instances. A higher sensitivity indicates that the model is better at correctly identifying positive cases and reducing false negatives (Burger, 2018).

$$Recall = \frac{\text{True positives}}{(\text{True positives} + \text{False negatives})}$$

Equation 7: Recall

F1 score

The F1 score is the harmonic mean of precision and recall. It provides a balanced measure between precision and recall and is useful when there is an uneven class distribution. A high F1 score indicates that the model has both good precision and good recall (Burger, 2018).

$$F1\,score = 2 \times \frac{\text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})}$$

Equation 8: F1 score

ROC AUC

ROC AUC is insensitive to class distribution and provides a more balanced view of the model's performance. It considers the entire range of classification thresholds rather than a single threshold. This is important because different thresholds can result in different trade-offs between true positive rate (sensitivity) and false positive rate (1 - specificity). In imbalanced datasets, optimizing a single threshold may not be effective, as it could lead to high false positive rates. ROC AUC provides an aggregated measure of model performance across all possible thresholds (Huyen, 2022).

Confusion score

The confusion score is a metric created by Günter Schmudlach that provides a measure of how much confusion there is relative to the true positives. Ideally, the confusion score would be 0. A higher confusion score typically indicates lower precision and/or recall. Therefore, a higher confusion score tends to lower the f1 score as well. For the given classification with imbalanced data, the confusion score is an appropriate metric. The confusion score is defined as:

$$Confusion\,score = \frac{(\text{False positives} + \text{False negatives}) \times 100}{\text{True positives}}$$

Equation 9: Confusion Score

4.3 Model selection methodology

In a conversation with Günter Schmudlach, the issue of model selection was discussed. The difficulty lies in defining the right ratio between false positives and false negatives. From a practical point of view, there is no right solution for the best ratio of the deviation between false positives and false negatives. The answer to the correct ratio lies in the subjective judgement of the problem. Therefore, a pragmatic solution had to be found for this issue. With regard to the model evaluation, it was decided for simplicity reasons that a model should be balanced. The performance of the trained models is therefore always determined with a *p*-threshold value that leads to a difference of 0 between false positives and false negatives. With that, a clear criterion for comparing the models is defined. This means that the metrics mentioned above (e.g. precision, recall, F1 score, confusion score) are calculated at the *p*-threshold value where an equal number of false positives and false negatives are predicted by the model. The performance of the different models is then compared using the defined metrics on the validation set. The model that performs best with as few features as possible is determined as the winner. The main criterion is Günter Schmudlach's confusion score metric. However, the other metrics are likely to correlate with it and are considered as well for a more comprehensive assessment. Once a winning model has been determined, the performance of this model is also evaluated on the test set.

4.4 Implementation in Python and R

For the modelling part, a `main.py` script was written in Python, which is used for modelling both {*caution*} and {*foot*}. As described earlier in this chapter, logistic regression, random forest and gradient boosting are modelled directly in Python with the library *scikit-learn*. Additionally, an R script has been created for the general additive models, which are implemented in the *mgcv* package. However, the `main.py` script in Python was created in such a way that it calls the R script `GAM.R` directly. At the end of the `main.py` script, the scores are added to an existing Excel file containing all the evaluation scores. The script is reusable and uses several functions from the `my_functions.py` script, where for example the filters and methods can be specified. This makes it possible to approach different feature engineering techniques and observe how the evaluation scores of the models change. Thus, before each execution of the `main.py` script, the variable *session_name* must be defined, which is saved in the Excel file at the end. This allows the feature engineering methods used (e.g. ti-filter, street-filter, oversampling, undersampling, scaling) to be traced when the scores of different runs are compared later. For transparency reasons of the individual run, the console output of each run was saved in a `log file` in case something needs to be retraced later. The log files also document certain parameters such as the selected variables with *RFE* or the best parameters according to *GridSearch*. The functions in the `my_functions.py` script are imported into the `main.py` script. The most important functions for modelling are briefly described below, whereby the detailed descriptions and parameters can be viewed directly in the `my_functions.py` script.

function `feature_engineering`

Applies all the feature engineering tasks from chapter 5.1. Filters can optionally be specified in the function call (e.g. `ti_filter=True, tunnel_filter=False, street_filter=False`). Additionally, features to be excluded can also be specified (e.g. `col_drop=List`)

function `train_val_test`

Preprocesses the cleaned input data by splitting it into training, validation, and test set, and exports the sets to CSV files (CSVs are used for reading in with `GAM.R`). Scaling, over- or undersampling can optionally be applied (e.g. `scaling=True, method='oversampled'`). Additionally, the target has to be specified in the function argument (e.g. `target='caution'`).

function `find_threshold`

The function uses a binary search algorithm to find the *p*-threshold value for which FP = FN results in the predictions. Thanks to the binary search, not all thresholds have to be searched and the search time can be greatly reduced as the algorithm iteratively narrows down the search range. Within each iteration, it calls the function `calculate_confusion_matrix` for the given threshold and adjusts the search range accordingly. If the confusion matrix yields FP = FN, the search is terminated and the *p*-threshold found is returned.

function `calculate_confusion_matrix`

Calculates the confusion matrix based on predicted probabilities and true labels at a specified threshold for binary classification. As described in the previous function `find_threshold`, the function is called for as long as it takes to find the optimum *p*-threshold which yields FP = FN.

function `evaluate`

Evaluates the binary classification model on the validation data and computes various classification metrics at the optimized *p*-threshold, where FP = FN. The function writes the performance metrics of the model into the data frame, which is later written into the Excel scoring file.

5 Analysis and modelling

This chapter is dedicated to the analysis and modelling part. Firstly, the data is transformed using feature engineering (NA imputation, transformations, creation of new features). Then the data is visualised and further analysed using explanatory data analysis. This gives an intuition for the features, their distribution and correlation. The next part is modelling, where the models are trained, evaluated and a final winner model is determined for each of the targets $\{caution\}$ and $\{foot\}$.

5.1 Feature engineering

This chapter summarises the results of this `Jupyter notebook`. Further details can be found in the notebook. The notebook is for visualisation purposes only. Nevertheless, the performed cleaning steps are included in the code of the reusable `main.py` script in the function `feature_engineering`.

5.1.1 Missing values

According to Chip Huyen, missing values are a common issue in machine learning. Most machine learning algorithms cannot work with missing features. Not all types of missing values have to be assessed in the same way. These can be divided into the categories MNAR (Missing not at random), MAR (Missing at random) and MCAR (Missing completely at random). Missing values can be either filled with certain values (imputation) or removed (deletion). Deletion is often preferred, because it is easy to perform. If a variable has a lot of missing values, column deletion would be a common approach. The drawback of this approach is that important information might be removed, which may result in lower accuracy of the final model. Another approach would be row deletion, where observations with missing values are removed. This method only works, if the missingness is MCAR and the number of missing values is less than 0.1%. However, removing rows may also remove important information that the model needs to make predictions. Lastly, removing rows of data can introduce bias to the model, especially if the missingness is MAR (Huyen, 2022).

During the (re)sampling process described in chapter 3, values of some variables are missing. This is due to the fact that if there is no value for the corresponding layer (e.g. street, crevasse, etc.) at a certain route point, QGIS sets the value for that observation to missing. The raw dataset has 1’955’978 observations (ex. snowshoe-only routes) and shows the following missing value count (vertical line on x-axis represents 2.5% of observations):

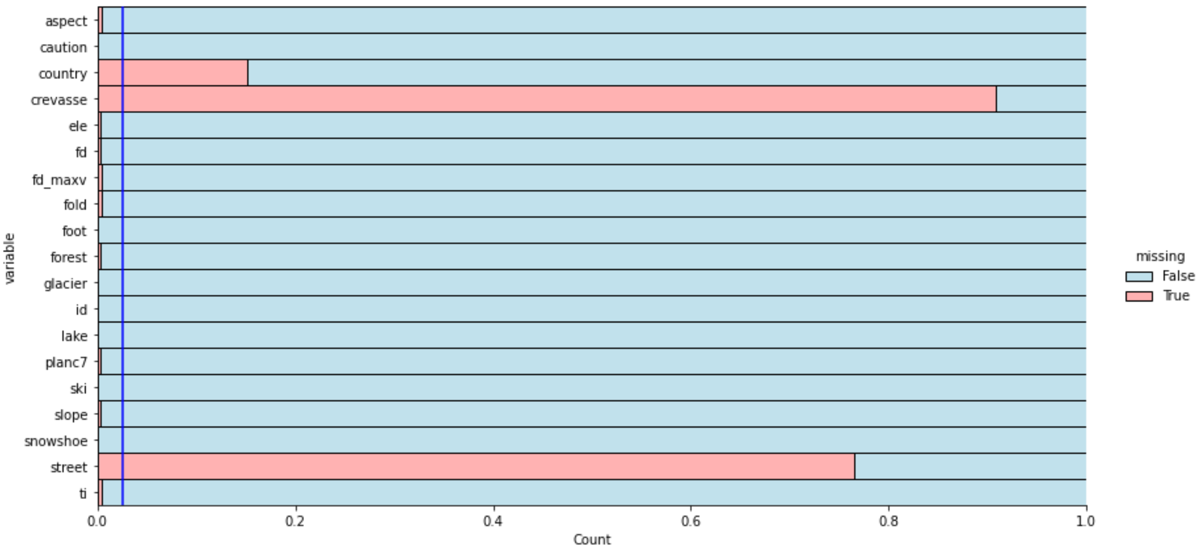


Figure 11: Missing values in the raw dataset after re(sampling)

As to be seen in Figure 11, the three variables {*crevasse*} (90.8%), {*street*} (76.6%) and {*country*} (15.1%) have a particularly high number of missing values. However, this can be explained by the sampling procedure. The variable {*country*} is missing if the data point is located abroad. As foreign countries are excluded later anyway due to poor data quality, this variable has no longer any missing values. The missing values for {*street*} also come from the sampling procedure. If a data point is further than 25 meters away from a street, the attribute of {*street*} is set to missing. The situation is similar for the variable {*crevasse*}. If a data point is not on a glacier, the class of the crevasse zone is set to missing. The variable {*country*} is therefore automatically completed by the filter (exclude foreign countries). The variables {*street*} and {*crevasse*} are easy to impute, which is performed at the end of this chapter.

At second glance, missing values can also be found for the variables {*slope*} (0.37%), *{fd}* (0.37%), {*forest*} (0.37%), {*ele*} (0.37%), {*fold*} (0.42%), *{ti}* (0.47%), {*fd_maxv*} (0.46%), {*planc7*} (0.37%) and {*aspect*} (0.47%). An analysis of the affected data points provided an explanation: The missing values were generated during sampling when the data point is on a lake. Then no value could be assigned by the affected variables, and QGIS set the value to missing. As the lakes will be used as filters for modelling later anyway and these data points will be excluded later, this problem is automatically solved. After excluding the lakes, all variables (except {*crevasse*} and {*street*}) no longer have any missing values. With the following two filters, 299’259 data points were removed, resulting in 1’656’719 observations remaining in the filtered dataset:

- **Filter 1**: exclude observations where {*country* = missing} (i.e. drop data points abroad)
- **Filter 2**: exclude observations where {*lake* = 1} (i.e. drop data points on lakes)

For the variables {*crevasse*} (91.4%) and {*street*} (75.5%), deletion would be the wrong approach because a tremendous amount of information would be lost. As Huyen mentioned, this would be very likely to introduce bias into the model. If no deletion should be applied, the values have to be imputed (i.e. assign or estimate value to missing value). The hard part is to choose a value for the imputation (Huyen, 2022).

The question therefore arises as to what is an appropriate value for the imputation of the missing values of the variables {*street*} and {*crevasse*}. As to be seen in Table 4, the missing values for the variable {*street*} are replaced by the value '4'. During sampling, QGIS generated a missing value if there was no street within a distance of 25 meters. The imputed value '4' therefore indicates that there is no street within 25 meters of the data point. A similar procedure is also used for the variable {*crevasse*}. The missing values can be explained by the fact that a value in the range [1, 7] was only assigned to a data point if it is located on a glacier. A missing value therefore means that the data point is not located on a glacier and is therefore not on a crevasse zone at all. The missing values can therefore be imputed with '0'. After imputation, the dataset no longer has any missing values.

Variable	Before imputation	After imputation
street	Categorical NA: No street nearby 1: Next street in distance 0...5 m 2: Next street in distance 5...15 m 3: Next street in distance 15...25 m	Categorical 1: Next street in distance 0...5 m 2: Next street in distance 5...15 m 3: Next street in distance 0...25 m 4: No street nearby
crevasse	Categorical NA: Data point not on glacier (i.e. no crevasse zone) 1: None to little crevasse zone 2: … 3: … 4: … 5: … 6: … 7: Very typical crevasse zone	Categorical 0: No crevasse zone 1: None to little crevasse zone 2: … 3: … 4: … 5: … 6: … 7: Very typical crevasse zone

Table 4: Imputation operations

5.1.2 Discretization

Discretization is the process of turning a continuous feature into a discrete feature. Therefore, buckets are created for the given values. The following two transformations will result in some loss of information, but on the other hand, the transformations increase the simplicity and interpretability of the model (Huyen, 2022). The variables {*aspect*} and {*street*} are therefore discretized into binary variables before modelling. In the raw dataset, the variable {*aspect*} is on a scale from 0° to 360°. As the literature repeatedly described the orientations northwest to north to northeast as being prone to unstable snow layers, this variable is transformed into a binary variable. Therefore, if the value is between 315° to 360° and 0° to 45°, the newly transformed variable {*aspect_binary*} has the value '1' (red in Figure 12). For values between 45° and 315°, the variable has the value '0' (blue in Figure 12).

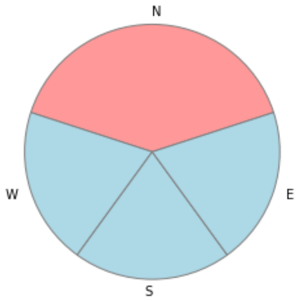


Figure 12: Sectors of newly created variable aspect_binary

The variable {*street*} gives an indication of how far away the data point is from the next street within 25 meters. In the raw dataset, the variable consists of an already discrete interval with the values '1' (next street within 0 to 5 meters), '2' (next street within 5 to 15 meters), '3' (next street within 15 to 25 meters) and '4' (no street within 0 to 25 meters). Nonetheless, the variable {*street*} is further transformed to the variable {*street_binary*}, which is binary and has the value '0' (blue in Figure 13) if no street is within 0 to 25 meters. If there is a street within 25 meters, the variable has the value '1' (red in Figure 13).

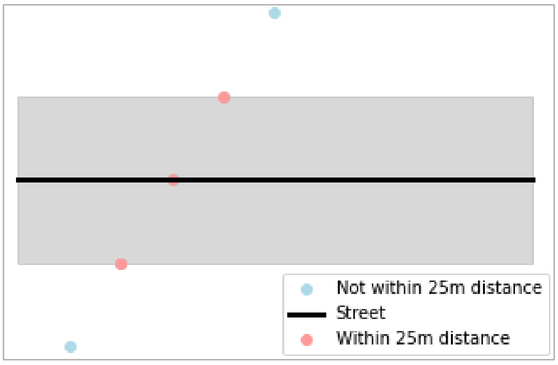


Figure 13: Binary categories of newly created variable street_binary

5.1.3 Scaling

According to Géron, feature scaling is an important transformation that often need to be applied in machine learning. The goal of scaling is to standardize the range of features or variables in a dataset, making them comparable and preventing features with larger scales from dominating those with smaller scales. Some algorithms don't perform well when the numerical features have very different scales. A common scaling technique is normalisation, where the values are shifted and rescaled so that they end up ranging from 0 to 1. Another technique is standardization, where first the mean value is subtracted, and afterwards the result is divided by the standard deviation. Standardized values have a mean of 0 and a standard deviation of 1 (Géron, 2019). For a random forest, scaling is not necessary since this model is not sensitive to the scale of the features. For instance, a decision tree only splits a node based on a single feature. For logistic regression and GAM, scaling may improve their performance (depending on the optimization algorithm). But the interpretability of the output is affected, if scaling is applied to the feature variables, because the coefficients and the smooth functions directly relate to the effect of each predictor on the outcome variable (Analytics Vidhya, online).

Because of the above reasons, scaling is being trialled during modelling, but if the added value in terms of scores is not promising when evaluating the models, the scaling approach is not pursued further. When modelling with Python, however, attention is paid to which algorithms are used to fit the model and whether feature normalisation is recommended or not. In some cases, the *StandardScaler* from the library *scikit-learn* may be used for experimental purposes.

5.1.4 Point distance

When sampling in QGIS, the continuous ski routes were sampled in points. The points now each have a distance of 10 meters to the next point. The result of this is that neighbouring points have relatively similar properties, potentially leading to autocorrelation in the dataset. Autocorrelation is not desirable in most statistical modelling because it violates the assumption of independence and can further influence parameter estimation.

Günter Schmudlach has already discussed this issue with an expert in the past and the result was that correlation can be reduced from a distance of 100 meters. The distance of 100 meters was therefore taken as a reference value for the point spacing and checked again. On the next page, Figure 14 shows how the autocorrelation appears to be present at a point distance of 10 meters (left plot), but has almost disappeared at a point distance of 100 meters (right plot). The slicing was done in Python in such a way that the dataset was first sorted in ascending order according to the variable {*id*} (corresponds to the order of ski routes). Then only every tenth {*id*}-value was kept, which leads to a reduction of the dataset by a factor of ten from 1'656'719 data points to 165'672 data points.

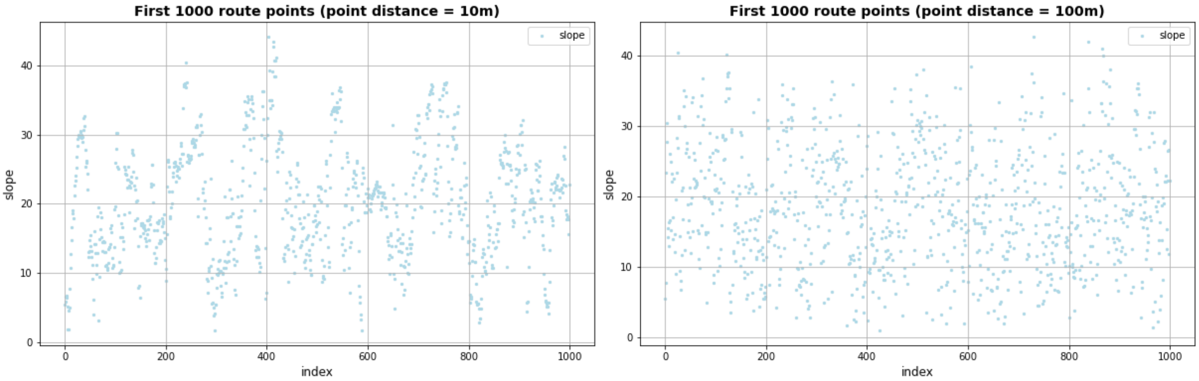


Figure 14: Correlation with different point distance

5.1.5 New feature

According to Skitourenguru, risk is defined as the product (multiplication) of the probability of an event and the consequences of the same event. The event referred to is the down falling of a skier:

- **Probability of a fall**: The steeper the slope, the more likely a fall. It is not known whether the relationship is linear. It is therefore a simplification.
- **Consequences of a fall**: One can mentally drop an object weighing 70 kg from any point in the terrain (e.g. every 10 m) and record the velocities and accelerations during the fall. High speeds and high accelerations presumably lead to more serious consequences (injuries). Together with SLF, Skitourenguru developed a corresponding physical fall model. The maximum speed during the crash can be used as a proxy for the consequences of a crash. This is also a simplification. (Skitourenguru, online)

For this reason and thanks to the advice of Günter Schmudlach, the variable {*fd_risk*} was calculated as the product of {*fd_maxv*} and {*slope*}. The upward outliers of {*fd_risk*} were limited to a maximum value of 2'500.

5.1.6 Further data cleaning

The variable {*planc7*} has very few values that are far below -350 or far above 350. The values outside the interval [-350, 350] are therefore limited to the values -350 respectively 350.

The most important steps regarding feature engineering have thus been completed. In the next chapter, the data is roughly analysed within the framework of explanatory data analysis (EDA). But there will be another filter step regarding data quality, which only became apparent during the following visual analysis of the data.

5.2 Exploratory data analysis (EDA)

This chapter is dedicated to exploratory data analysis (EDA). This part is intended to provide a rough qualitative and quantitative overview of the data before it is modelled. The data is first shown graphically and then shown as summarized statistics. After these two steps, the data is filtered further, as data quality deficiencies became visible due to the branch-wise labelling (see chapter 3.5). Afterwards, outliers are detected and correlation coefficients are analysed (on both the unfiltered and filtered data). A `Jupyter notebook` was also created for this chapter, but only for visualisation purposes and not for reusability. The following visual analysis summarises the key findings from the notebook.

5.2.1 Visual analysis

Target variables

Figure 15 shows the distribution of the binary target variables {*caution*} and {*foot*}. It can be deduced from this that the two variables are very imbalanced. The classes {*caution* = 1} (7.02%) and {*foot* = 1} (1.48%) (red in Figure 15) are strongly underrepresented in the data. The case {*foot* = 1} is rarer than the case {*caution* = 1}. Imbalanced data in machine learning may lead to poor performance, because the distribution of classes is not equal, meaning one class significantly outnumbers the other. In the modelling part, the models are trained on the imbalanced dataset on the one hand, but on the other hand the problem is also addressed by experimenting during the modelling process with techniques such as oversampling and undersampling.

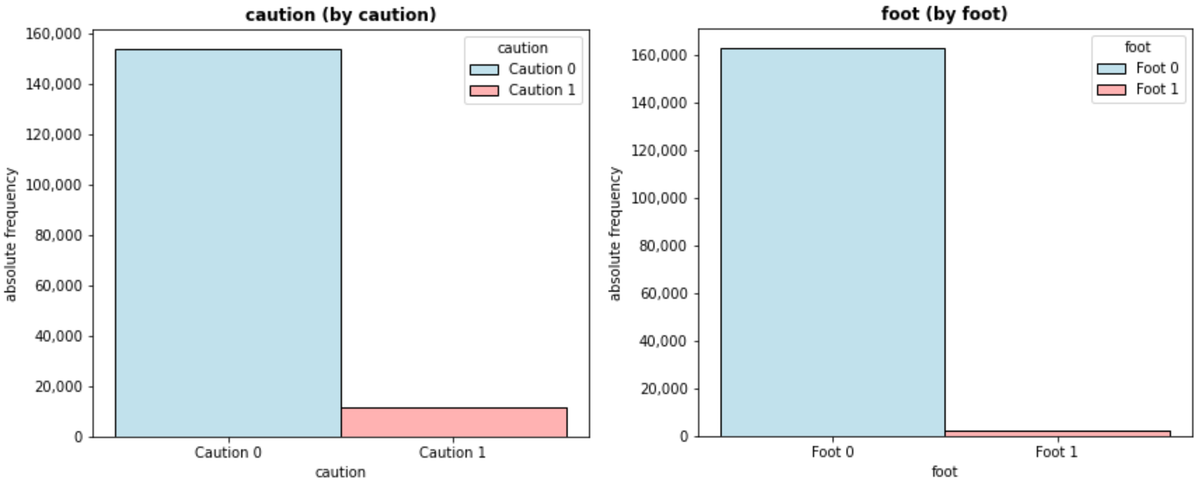


Figure 15: Target variables {caution} and {binary}

Binary feature variables

The remaining binary variables are features that are used for the predictions. It is obvious that the imbalance in the two classes {*caution*} and {*foot*} is also evident in the categorized feature data. If the transformed variable {*aspect_binary*} is considered, it can be seen in Figure 16 that on a relative scale, {*caution* = 1} appears to be slightly more frequent for NW, N and NE, i.e. {*aspect_binary* = 1}. The specific relative proportions can be found in the summary statistics in the next chapter.

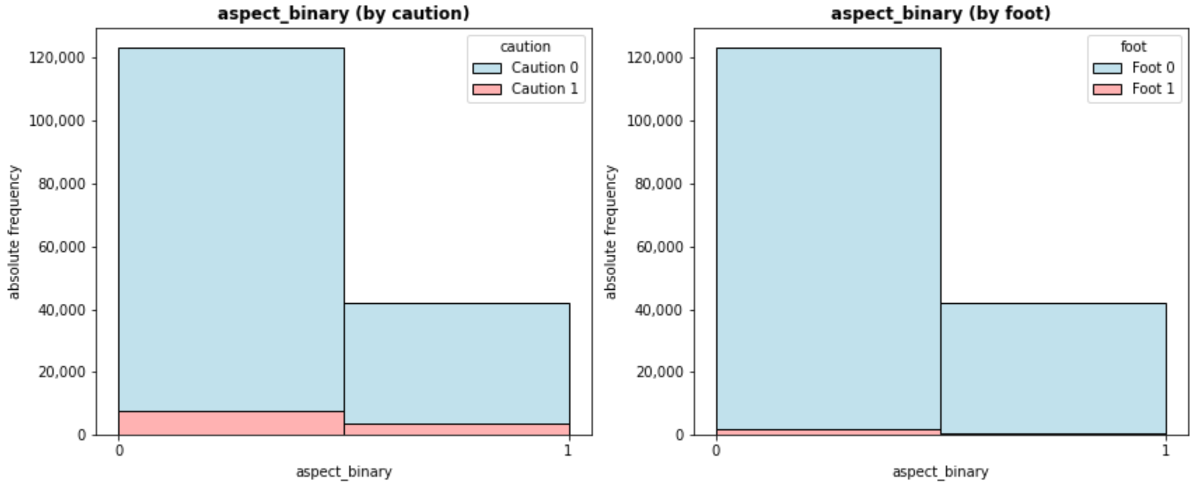


Figure 16: Feature variable {aspect_binary} by categories caution and foot

Two graphical conclusions can be drawn with regard to the binary variable {*forest*} shown in Figure 17: The fact from the interview that forests can be *caution* (due to dangers such as fatigue or loss of orientation) does not seem to be reflected in the data. Additionally, forests seem to reduce the danger in the data, as the terrain can become safer due to a lower avalanche risk according to the literature. But it could also just be a confounder, as forests are only present at a certain elevation, don't occur on glaciers and tend not to be located in very steep terrain.

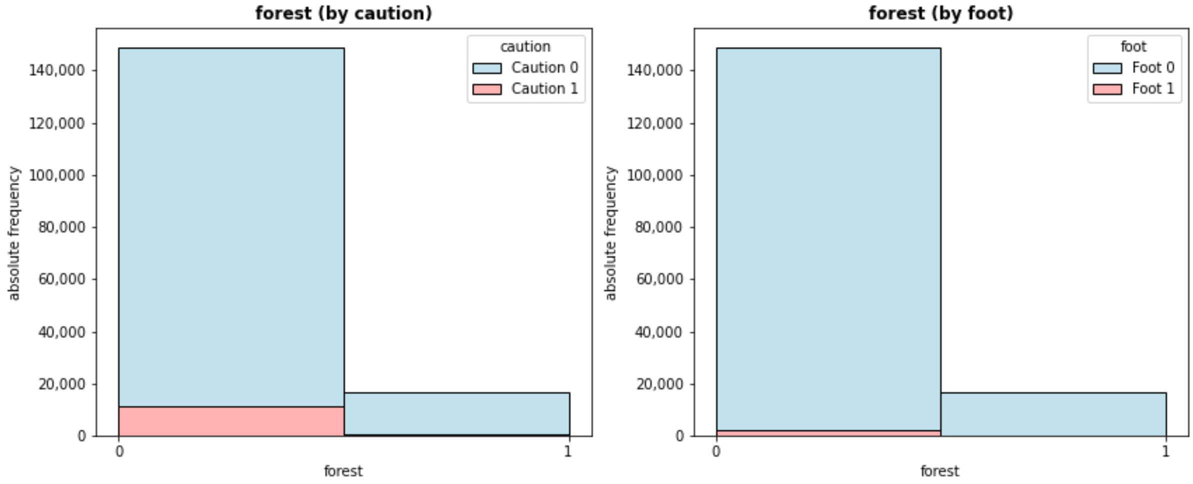


Figure 17: Feature variable {forest} by categories caution and foot

With the variable {*street_binary*}, it is immediately recognisable that values with {*street_binary* = 1} are almost never a *foot*-section. A *caution*-section also seems to occur very rarely on a street. The observation is visualised in Figure 18. This is also in line with the information from the interview where Andreas Eisenhut pointed out that the sections on roads are almost never {*caution* = 1} or {*foot* = 1}. In the modelling part, it will also be experimented with excluding all points where {*street_binary* = 1} from the dataset.

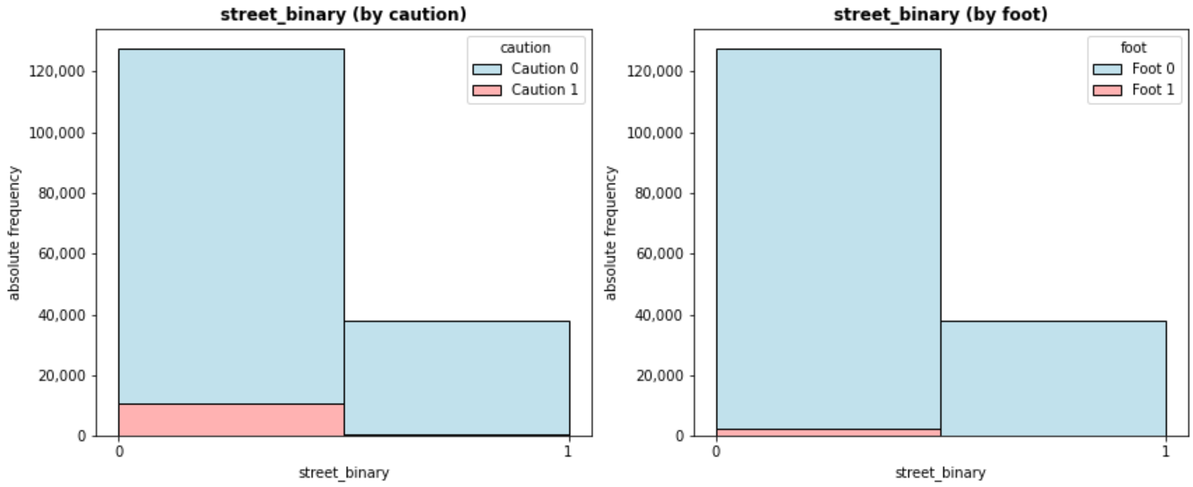


Figure 18: Feature variable {street_binary} by categories caution and foot

Continuous feature variables

The variables {*ele*}, {*fd*}, {*fd_maxv*}, {*fd_risk*}, {*fold*}, {*planc7*}, {*slope*} and {*ti*} are continuous. A selection of the variables is visualised below. For the variable {*ele*}, it can be seen in Figure 19 that the frequency of a *caution*-section seems to increase from an altitude of 1'000 meters. Additionally, the relative frequency for a *foot*-section increases in particular from an altitude of around 3'000 meters. With regard to *caution*-sections, it should be recalled at this point that the *caution*-sections in the data are labelled branch-wise (see chapter 3.5). Thus, the {*caution* = 1} data points at the lower end of the x-axis could be data points that are only *caution*-sections due to the branch-wise marking (but in reality not *caution*). This data quality issue is addressed later in this chapter.

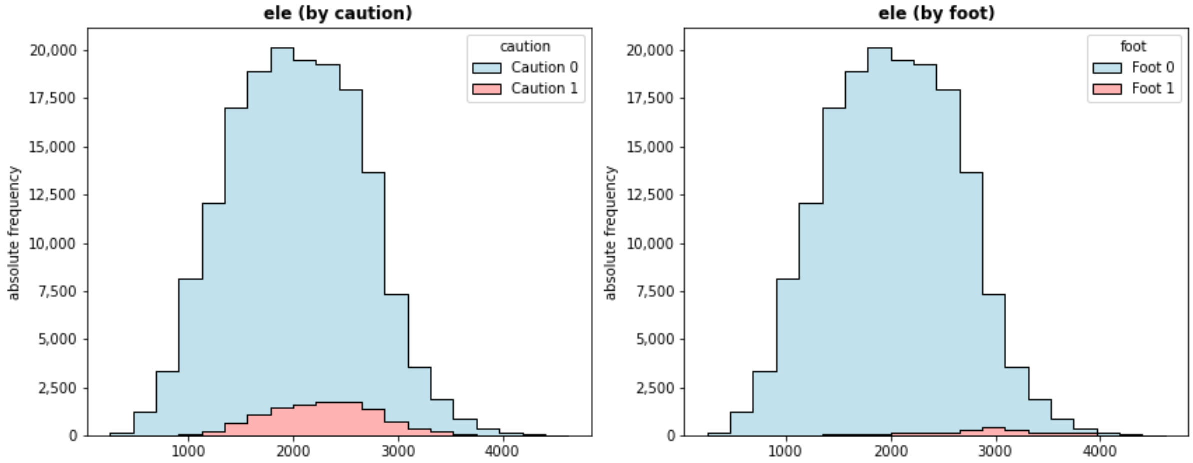


Figure 19: Feature variable {ele} by categories caution and foot

The variable {*slope*} represents the steepness, the variable {*fd_risk*} the risk of falling. According to Figure 20 and Figure 21, the steeper the terrain or higher the risk of falling, the more likely it is for the point to be a *caution*- or *foot*-section. For the variable {*ti*}, it can be seen that the risk increases from 0.6, when categorised by {*caution*}, as to be seen in Figure 22.

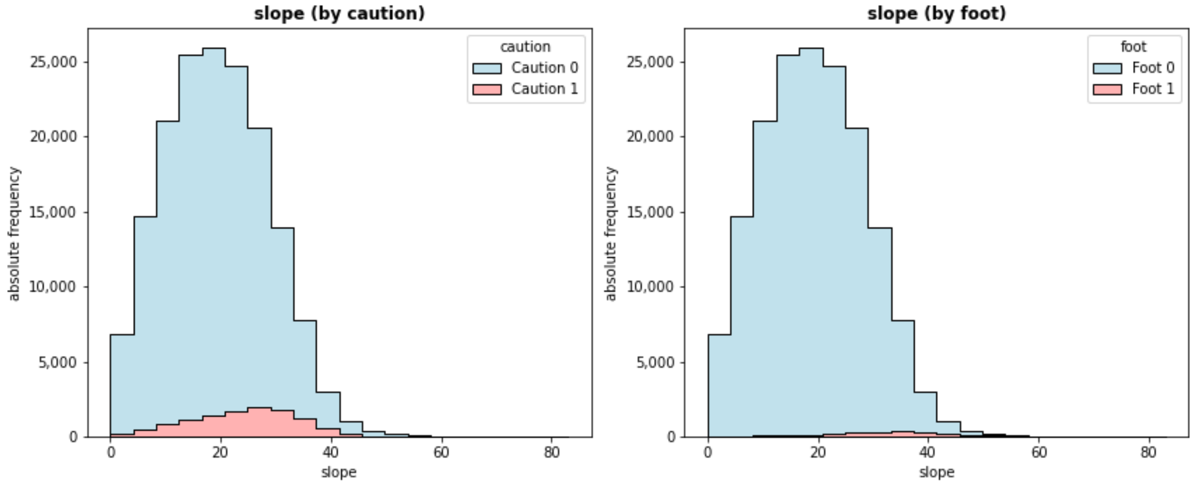


Figure 20: Feature variable {slope} by categories caution and foot

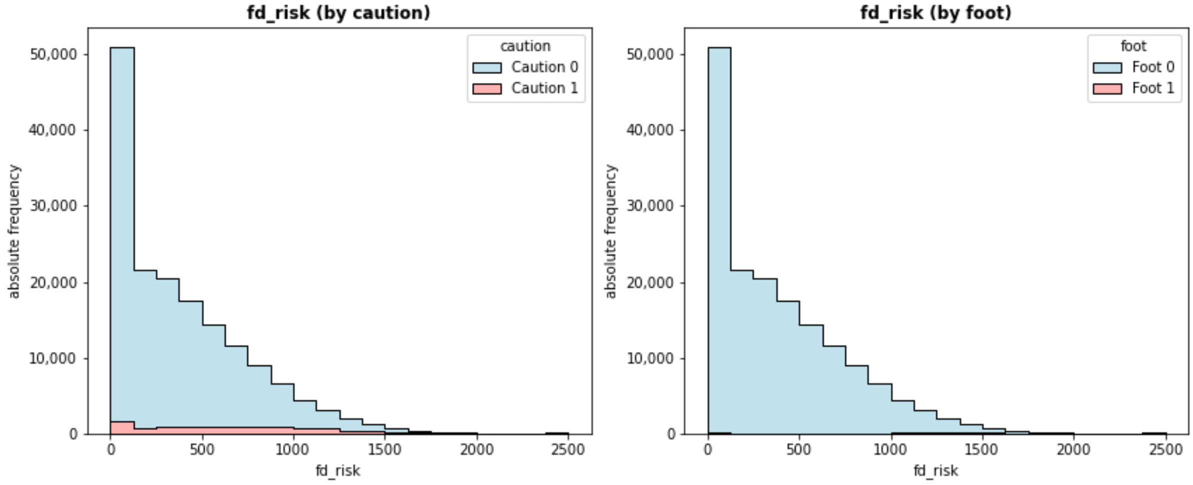


Figure 21: Feature variable {fd_risk} by categories caution and foot

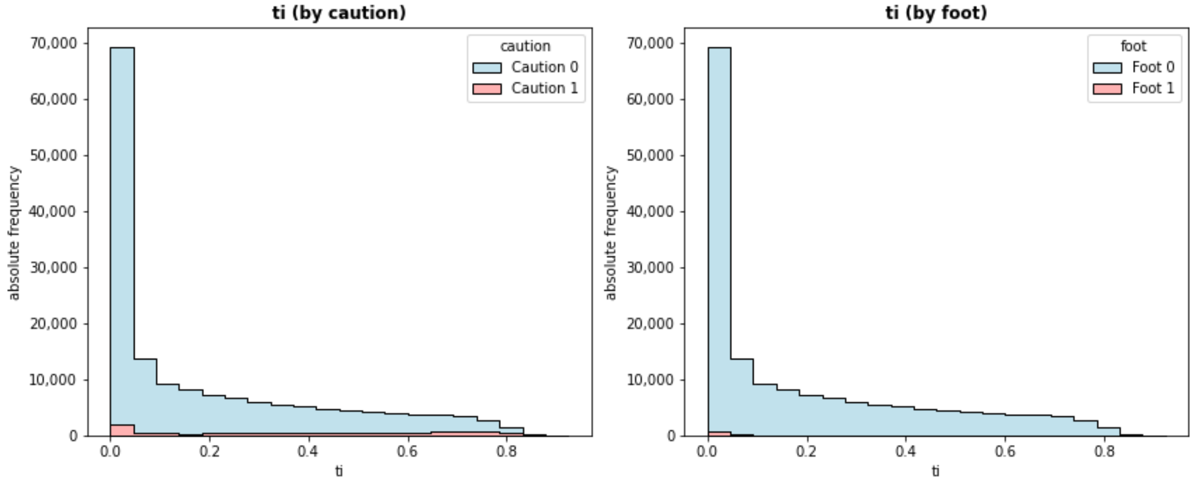


Figure 22: Feature variable {ti} by categories caution and foot

The two plots above for the variables {*fd_risk*} and {*ti*} also show a dense frequency of {*caution* = 1} on the left-hand spectrum of the x-axis. These *caution*-sections in somehow harmless terrain probably also occur from the data quality issue due to the branch-wise marking of the *caution*-sections (i.e. data points that are in reality not *caution*). This class noise problem is addressed later in this chapter.

Categorical feature variables

As the categorical variable {*street*} has been recoded into a binary variable, there is now only one remaining categorical variable in the dataset. Therefore, the variable {*crevasse*} indicates how crevasse-like the terrain is. A value of '0' corresponds to no crevasse zone (i.e. data point lies not on a glacier), while a value of '7' corresponds to a very typical crevasse zone. In the following, the histogram is plotted with a filter if the point is located on a glacier, i.e. excluding all values {*crevasse* = 0}. It is obvious that most of the route points are not located on a glacier, i.e. {*crevasse* = 0}. However, if only the subset with the glaciers is considered, i.e. {*crevasse* ≠ 0}, then it can be seen, particularly in the breakdown by {*caution*}, that the relative proportion of {*caution* = 1} increases with a higher {*crevasse*} value, as shown in Figure 23. This is also recognisable for the variable {*foot*}. Specific relative frequency numbers can be found in the following chapter on summary statistics.

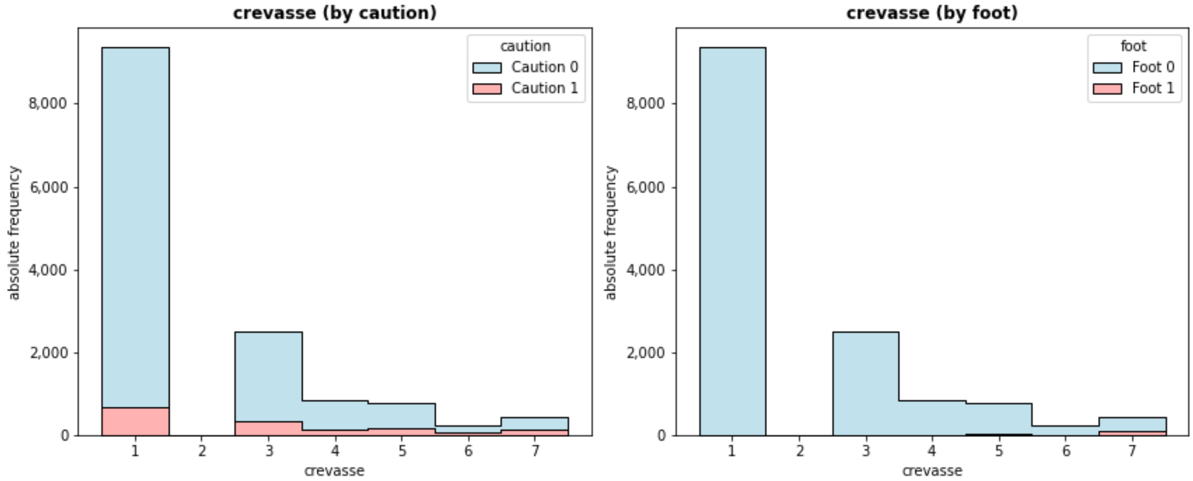


Figure 23: Variable {crevasse} by the categories caution and foot (filter: {crevasse ≠ 0})

5.2.2 Summary statistics

Now that the variables have been visually analysed in the previous chapter, some key figures for the summary statistics are provided. The methodology of the calculations can be viewed in the Jupyter Notebook, whereby the analysis was carried out with the *Pandas* library in Python. The following analysis of the summary statistics is divided into the variable types continuous, binary and categorical.

In Table 5, the key figures are divided into the categories $\{caution = 0\}$ and $\{caution = 1\}$ or $\{foot = 0\}$ and $\{foot = 1\}$. If the target variable $\{caution\}$ is analysed, an increase in the mean value can be observed for most variables if $\{caution = 1\}$ compared to $\{caution = 0\}$. The 0.5 percentile stands for the median. Only for forest density $\{fd\}$ and curvature $\{fold\}$ does the mean value decrease when $\{caution = 1\}$. The spread around the mean value is discussed in more detail in the outlier chapter. The observations are similar for the target variable $\{foot\}$, whereby the mean value increases particularly strongly for the variables $\{ele\}$, $\{fd\_maxv\}$, $\{fd\_risk\}$, $\{fold\}$ and $\{planc7\}$ if the target variable $\{foot = 1\}$.

Summary statistics by {*caution* = 0}								
	ele	fd	fd_maxv	fd_risk	fold	planc7	slope	ti
count	154’036	154’036	154’036	154’036	154’036	154’036	154’036	154’036
mean	2’007	11.64	17.63	377.26	0.29	-1.78	19.03	0.19
std	632	26.45	12.90	374.93	14.61	36.03	9.35	0.22
min	250	0.00	0.00	0.00	-105.51	-350.00	0.02	0.00
0.25	1’530	0.00	5.27	48.21	-7.65	-8.17	12.01	0.01
0.5	1’994	0.00	18.96	292.98	-1.72	0.14	18.57	0.08
0.75	2’470	0.00	27.32	585.87	6.71	8.51	25.48	0.32
max	4’622	100.00	115.86	2’500.00	114.86	350.00	82.99	0.92
Summary statistics by {*caution* = 1}								
	ele	fd	fd_maxv	fd_risk	fold	planc7	slope	ti
count	11’636	11’636	11’636	11’636	11’636	11’636	11’636	11’636
mean	2’292	5.99	25.79	685.39	-2.12	2.97	24.23	0.40
std	539	19.40	12.80	477.98	15.45	30.98	9.65	0.27
min	763	0.00	0.00	0.00	-121.68	-350.00	0.38	0.00
0.25	1’911	0.00	18.63	299.23	-9.55	-5.09	17.34	0.13
0.5	2’299	0.00	28.14	656.48	-3.27	2.26	24.92	0.42
0.75	2’653	0.00	34.77	1’017.56	5.82	11.30	31.23	0.65
max	4’531	100.00	113.02	2’500.00	129.28	350.00	71.68	0.91
Summary statistics by {*foot* = 0}								
	ele	fd	fd_maxv	fd_risk	fold	planc7	slope	ti
count	163’225	163’225	163’225	163’225	163’225	163’225	163’225	163’225
mean	2’016	11.33	17.88	386.12	-0.29	-0.82	19.21	0.20
std	622	26.14	12.71	370.23	13.67	34.50	9.27	0.23
min	250	0.00	0.00	0.00	-121.68	-350.00	0.02	0.00
0.25	1’550	0.00	6.13	58.48	-7.79	-7.68	12.20	0.02
0.5	2’010	0.00	19.38	305.25	-1.95	0.37	18.84	0.09
0.75	2’473	0.00	27.69	604.89	6.45	8.78	25.78	0.35
max	4’531	100.00	113.02	2’500.00	129.28	350.00	82.99	0.92
Summary statistics by {*foot* = 1}								
	ele	fd	fd_maxv	fd_risk	fold	planc7	slope	ti
count	2’447	2’447	2’447	2’447	2’447	2’447	2’447	2’447
mean	2’768	5.71	39.87	1’251.52	27.72	-42.95	32.09	0.28
std	739	19.09	17.20	688.66	36.73	72.31	12.84	0.27
min	491	0.00	0.00	0.00	-98.44	-350.00	0.34	0.00
0.25	2’339	0.00	35.47	799.58	-3.44	-68.68	23.69	0.02
0.5	2’893	0.00	41.85	1’266.91	30.75	-23.78	32.89	0.19
0.75	3’193	0.00	48.35	1’713.14	57.24	1.14	39.91	0.51
max	4’622	100.00	115.86	2’500.00	111.50	350.00	82.54	0.89

Table 5: Summary statistics for continuous variables

In the table above, the minimum values and the first quartile of the variables {*fd_maxv*}, {*fd_risk*} and {*ti*} for the category {*caution* = 1} raise questions, because it is very likely that these points in reality are non-dangerous terrain. Once again, it can be assumed that this has to do with the data quality issue of the branch-wise marking of the sections (class noise). A data point with a value of {*ti* < 0.25} is highly doubtful as to whether this is really a *caution*-section in reality.

The binary variables are analysed according to a similar logic in Table 6, so that a distinction is also made between the category of the target variable. The conditional count within a class is also shown, i.e. if the variable {*aspect_binary* = 1}, a further division of the count into the categories {*caution* = 0} and {*caution* = 1} is made. The same division applies to the target variable {*foot*}. It is not surprising to see that if the variable {*street_binary* = 1}, only 1.8% of the cases involve a *caution*-section. The situation is similar for {*forest* = 1}, where a *caution*-section only applies in 3% of cases. For these two feature variables, the proportions are even lower if *foot*-sections are considered. Very questionable were *foot*-sections that lie on a road. At first glance, it is difficult to explain why one should walk on a road without skis. However, analysing the points where {*foot* = 1 and *street_binary* = 1} on the map provided the answer: these are *foot*-sections of the route that lead through a tunnel. Since these points potentially lead to class noise because they lie on harmless terrain, the modelling has experimented with the exclusion of tunnels (in the following **'tunnel-filter'**). If a point lies on a glacier, 11% of the points are attributable to a *caution*-section. For the *foot*-sections, only 2.2% are attributable to this variable.

Summary statistics by {*caution*}										
	True (1)	Caution (1) when True (1)	Caution (1) % when True (1)	Caution (0) when True (1)	Caution (0) % when True (1)	False (0)	Caution (1) when False (0)	Caution (1) % when False (0)	Caution (0) when False (0)	Caution (0) % when False (0)
aspect_binary	**42'266**	3'782	9.0%	38'484	91%	**123'406**	7'854	6.4%	115'552	94%
caution	**11'636**	11'636	100.0%	0	0%	**154'036**	0	0.0%	154'036	100%
foot	**2'447**	0	0.0%	2'447	100%	**163'225**	11'636	7.1%	151'589	93%
forest	**16'640**	506	3.0%	16'134	97%	**149'032**	11'130	7.5%	137'902	93%
glacier	**13'878**	1'523	11.0%	12'355	89%	**151'794**	10'113	6.7%	141'681	93%
street_binary	**37'992**	695	1.8%	37'297	98%	**127'680**	10'941	8.6%	116'739	91%

Summary statistics by {*foot*}										
	True (1)	Foot (1) when True (1)	Foot (1) % when True (1)	Foot (0) when True (1)	Foot (0) % when True (1)	False (0)	Foot (1) when False (0)	Foot (1) % when False (0)	Foot (0) when False (0)	Foot (0) % when False (0)
aspect_binary	**42'266**	494	1.2%	41'772	99%	**123'406**	1'953	1.6%	121'453	98%
caution	**11'636**	0	0.0%	11'636	100%	**154'036**	2'447	1.6%	151'589	98%
foot	**2'447**	2'447	100.0%	0	0%	**163'225**	0	0.0%	163'225	100%
forest	**16'640**	169	1.0%	16'471	99%	**149'032**	2'278	1.5%	146'754	98%
glacier	**13'878**	301	2.2%	13'577	98%	**151'794**	2'146	1.4%	149'648	99%
street_binary	**37'992**	100	0.3%	37'892	100%	**127'680**	2'347	1.8%	125'333	98%

Table 6: Summary statistics for binary variables

In Table 7, the categorical variable {*crevasse*} with its eight levels is analysed in more detail. Level '0' indicates, that the data point is not on a glacier, which is true for the majority of the data points. If a closer look at the highest level '7' is taken, which indicates very typical crevasse zone terrain, only 1.5% of the data points are assigned to this category. However, it is interesting to note that within this level, around 33% (146 / [146 + 296]) of the data points are {*caution* = 1}. This ratio is not as high in any other class. In other words, the chance of {*caution* = 1} increases when the crevasse zone level is higher. This pattern is not as pronounced if the categories of the target variable {*foot*} are considered.

	Summary statistics by *{caution}*				Summary statistics by *{foot}*			
Crevasse Level	Caution (0)	Relative Frequency	Caution (1)	Relative Frequency	Foot (0)	Relative Frequency	Foot (1)	Relative Frequency
0	141’403	91.8%	10’052	86.4%	149’192	91.4%	2’263	92.5%
1	8’686	5.6%	688	5.9%	9’361	5.7%	13	0.5%
2	11	0.0%	1	0.0%	12	0.0%	0	0.0%
3	2’150	1.4%	342	2.9%	2’484	1.5%	8	0.3%
4	717	0.5%	142	1.2%	852	0.5%	7	0.3%
5	594	0.4%	185	1.6%	750	0.5%	29	1.2%
6	179	0.1%	80	0.7%	240	0.2%	19	0.8%
7	296	0.2%	146	1.3%	334	0.2%	108	4.4%

Table 7: Summary statistics for categorical variables

5.2.3 Distribution analysis

In this chapter, the figures from the summary statistics in the previous chapter become a little more recognisable. Outliers correspond to events that are uncommon. A boxplot can be interpreted as a one-dimensional graph of numerical data. It includes the minimum, the 25th percentile (Q_1), the median, the 75th percentile (Q_3) and the maximum. With a boxplot, outliers can be determined with the interquartile range ($Q_3 - Q_1$). The resulting range can further be multiplied by 1.5 (rule of thumb), which gives a wider boundary than the boxplot itself. Any data points that fall outside this boundary are determined to be outliers. A wider range of the IQR implies that the data is more spread out (Rumsey, 2011).

As described in the previous chapters, the data quality for the {*caution* = 1} data points is questionable due to the generous branch-wise marking. Another problem with the *caution*-sections are the data points that are {*foot* = 1}, as these are also typical *caution*-terrain (i.e. ridges, highly exposed terrain) in reality, but these are labelled {*caution* = 0} as explained in chapter 3.5. In consultation with Günter Schmudlach, the following filter (in the following ‘**ti-filter**’) was applied to the data for modelling *caution*-sections:

- drop *foot*-sections (i.e. {*foot* = 1})
- drop *caution*-sections with very low terrain indicator value (i.e. {*caution* = 1 and *ti* < 0.25}) and drop *non-caution-sections* with very high terrain indicator value (i.e. {*caution* = 0 and *ti* > 0.75})

The desired effect of this filter is that the points marked by the branch-wise marking, which in reality are the opposite class due to poor data quality, disappear. In addition, the dropping of the *foot*-sections (i.e. points with attributes {*foot* = 1 and *caution* = 0}) eliminates points that are labelled as non-*caution* in the data, but in reality correspond to *caution* (*foot*-terrain is always very dangerous). To visualise the effect of the filter, both the boxplots before and after the ti-filter are shown below for the categorisation according to *caution*-sections. A greater spread in the feature variables is to be expected between {*caution* = 0} and {*caution* = 1}. In the modelling chapter, calculations are performed both once with and once without the ti-filter, as this can be set to *true* or *false* in the `main.py` script.

In the following, the continuous variables are categorised once into {*caution*} (before and after the ti-filter) and once into {*foot*} in a boxplot. An impression can be gained of the median of the target variables divided into ‘0’ and ‘1’, and on the other hand, the number of outliers (outside 1.5 x IQR) can be read for the features. Figure 24 shows a difference in the median for the continuous variables {*ele*}, {*fd_maxv*}, {*fd_risk*}, {*slope*} and {*ti*}. The differences in the median of these feature variables between the two levels of the target variable might indicate predictive power and are more likely to be selected in the model. The variable {*ti*}, which indicates how suitable a terrain point is to trigger an avalanche, is particularly noticeable, where the median of the category {*caution* = 1} is 0.42, while the median of the category {*caution* = 0} is only 0.08. On the other hand, the median for the variables {*fd*}, {*fold*} and {*planc7*} appear to be more or less similar for the category levels {*caution* = 0} and {*caution* = 1}.

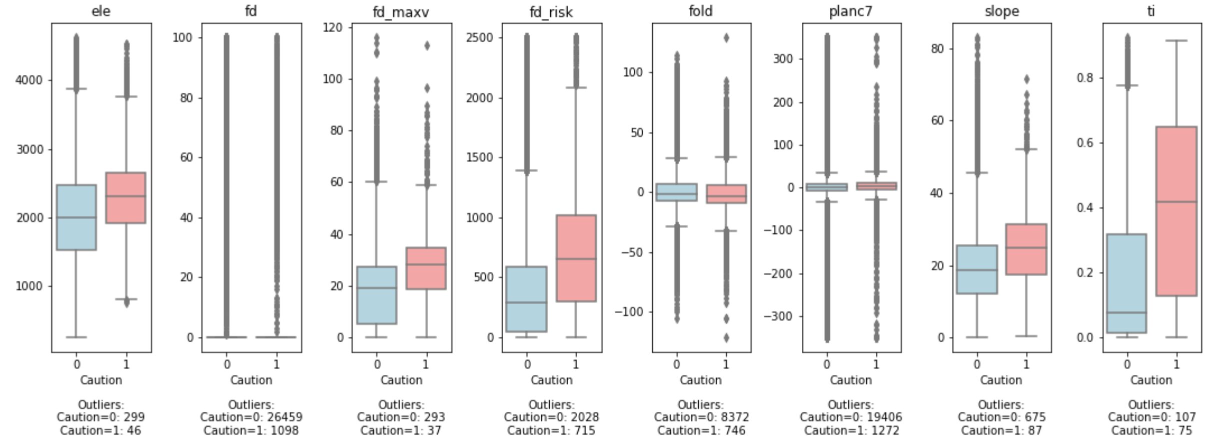


*Figure 24: Boxplots for continuous variables (categorized by {caution}, **unfiltered** data)*

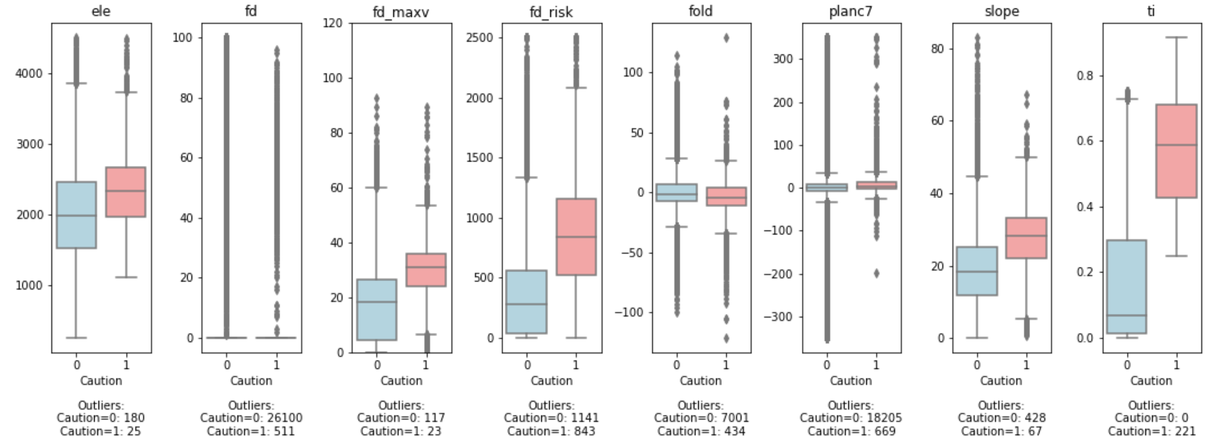


*Figure 25: Boxplots for continuous variables (categorized by {caution}, **ti-filtered** data)*

When the ti-filter is applied in Figure 25, the median between {*caution* = 0} and {*caution* = 1} drifts further apart, especially for the variables {*fd_maxv*}, {*fd_risk*}, {*slope*} and {*ti*}. This is advantageous in modelling, as there will be less class noise and a stronger signal in the data.

Interestingly, a slightly different picture can be seen in the *foot*-sections in Figure 26. For the variables {*ele*}, {*fd_maxv*}, {*fd_risk*}, {*fold*}, {*planc7*}, {*slope*} and {*ti*}, the median clearly deviates if {*foot* = 1}. It can further be seen that the median of the variable {*ele*} in the category {*foot* = 1} deviates even more strongly upwards. Thus, from around 3'000 meters above sea level, it is therefore far more likely that a data point belongs to a *foot*-section. Compared to the previous figure, the median not only deviates upwards, but also downwards, as the variable {*planc7*} shows. For the variable {*fold*}, a pronounced difference in the median is now also recognisable, which indicates *foot*-sections on ridges. The risk of falling, which is quantified by the variable {*fd_risk*}, may also be a very suitable predictor of a *foot*-section, as the median of the category {*foot* = 1} is considerably higher. On the other hand, there are many {*foot* = 0} points that have an outlier value for the variable {*fd_risk*}. These tend to originate from points that are located locally in steep, exposed terrain. Addressing these outliers is difficult, as their exclusion potentially introduces a bias into the dataset. For this reason, the tunnel-filter mentioned in the previous chapter is used, but this filter only addresses the noise in the {*foot* = 1} level (which is the class of interest).

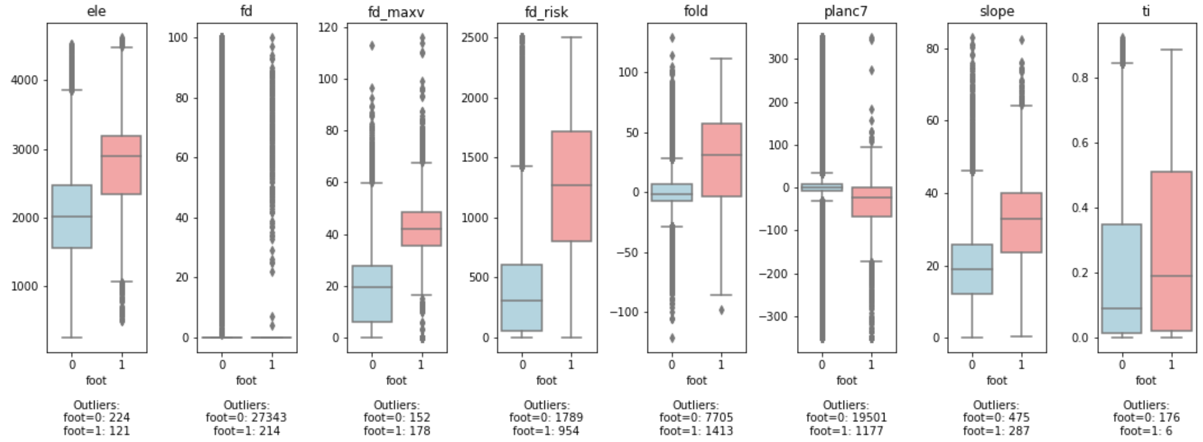


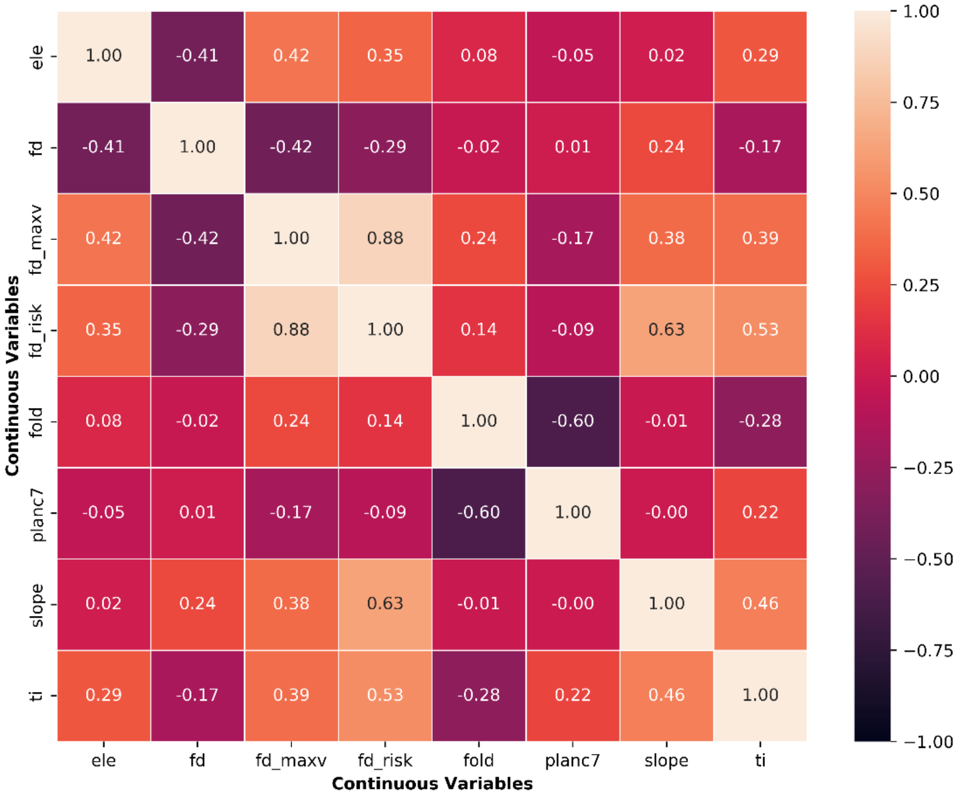
Figure 26: Boxplots for continuous variables (categorized by {foot})

Outliers outside the interquartile range can be seen for all variables. In larger datasets, however, it is common that more outliers can be seen in absolute figures. Expressed as a relative ratio, the variables {*fd*} (≈16%), *{planc7}* (≈12%) and *{fold}* (≈5%) in particular have many outliers. For the time being, the outliers are not transformed any further. The variable {*fd*} represents the forest density on a scale from [0 to 1]. Since many data points are not in a forest (i.e. have the value '0'), the IQR is '0' and all data points with a forest density > '0' are outliers. For this reason, this variable is used with care in the modelling. The binary variable {*forest*} is probably more suitable for modelling. The variable {*planc7*} provides information about the curvature, but is likely to correlate negative with {*fold*} (see next chapter). As {*fold*} tends to have fewer outliers, this variable is preferred over {*planc7*}. In the modelling of {*foot*}, the variable {*planc7*} is nevertheless taken into account for experimentation.

5.2.4 Correlation analysis

In the previous chapter, the individual variables were analysed using a boxplot. In this chapter, the findings are examined in more detail using a correlation analysis. The deviation of the median values already allowed some conclusions to be drawn as to whether the variables have an influence on {*caution*} and {*foot*}. A further approach can be carried out using correlation analysis. In the following, only the correlations of the continuous variables are initially considered. In a second step, the binary and categorical variables were also included, but the correlation had to be calculated using a different calculation method. The standard correlation coefficient (also called Pearson's r) can be calculated in Python using the *.corr()* method from the *Pandas* library for continuous variables. The coefficients are in the range [-1 to 1]. If it is close to 1, it means that there is a strong positive linear correlation between the two variables. When the coefficient is close to -1, it means that there is a strong negative linear correlation. Finally, coefficients close to 0 mean that there is no linear correlation (Géron, 2019).

The correlation matrix in Figure 27 shows a negative correlation between the forest density {*fd*} and the elevation {*ele*} with an r of -0.41. This can be explained by the fact that forest cover tends to decrease with increasing altitude up to the tree line. In addition, {*fd*} also correlates negatively with the maximal fall down velocity {*fd_maxv*} and {*fd_risk*}, which could be related to the fact that the higher the fall risk, the steeper the terrain and thus the lower the forest cover (e.g. steep rock slopes).



	ele	fd	fd_maxv	fd_risk	fold	planc7	slope	ti
ele	1.00	-0.41	0.42	0.35	0.08	-0.05	0.02	0.29
fd	-0.41	1.00	-0.42	-0.29	-0.02	0.01	0.24	-0.17
fd_maxv	0.42	-0.42	1.00	0.88	0.24	-0.17	0.38	0.39
fd_risk	0.35	-0.29	0.88	1.00	0.14	-0.09	0.63	0.53
fold	0.08	-0.02	0.24	0.14	1.00	-0.60	-0.01	-0.28
planc7	-0.05	0.01	-0.17	-0.09	-0.60	1.00	-0.00	0.22
slope	0.02	0.24	0.38	0.63	-0.01	-0.00	1.00	0.46
ti	0.29	-0.17	0.39	0.53	-0.28	0.22	0.46	1.00

*Figure 27: Correlation analysis of continuous variables using Pearson's r, **unfiltered** data*

Figure 27 also shows that with an r of 0.42, the variable {*fd_maxv*} tends to correlate positive with the variable elevation {*ele*}. This can be explained by the fact that the higher a point is, the higher is the probability of a high maximal fall down velocity. The r value of {*slope*} and {*ti*} is 0.46, which indicates a positive correlation. This can be explained by the fact that the steeper a gradient is in the variable {*slope*}, the more likely it is that the point is also typical avalanche terrain in the variable {*ti*}. A similar picture emerges for the variable {*fd_risk*} and {*ti*} with an r of 0.53. Thus, the variables {*fd_maxv*}, {*fd_risk*}, {*slope*} and {*ti*} appear to correlate slightly positive with each other. The variables relating to curvature {*fold*} and {*planc7*} correlate negatively, which is due to their calculation methods.

In a second step, a correlation matrix was created in which the binary and categorical variables were also taken into account. The point biserial correlation from the *SciPy* library is used to measure the relationship between a binary variable, x, and a continuous variable, y. Like other correlation coefficients, it lies in the range [-1 to 1] with 0 implying no correlation. Correlations near -1 or 1 imply a strong linear relationship (SciPy, online). Figure 28 shows the correlation matrix for all variable types. The underlying data is unfiltered, so the branch-wise marking of {*caution*} is not yet taken into account in this matrix.

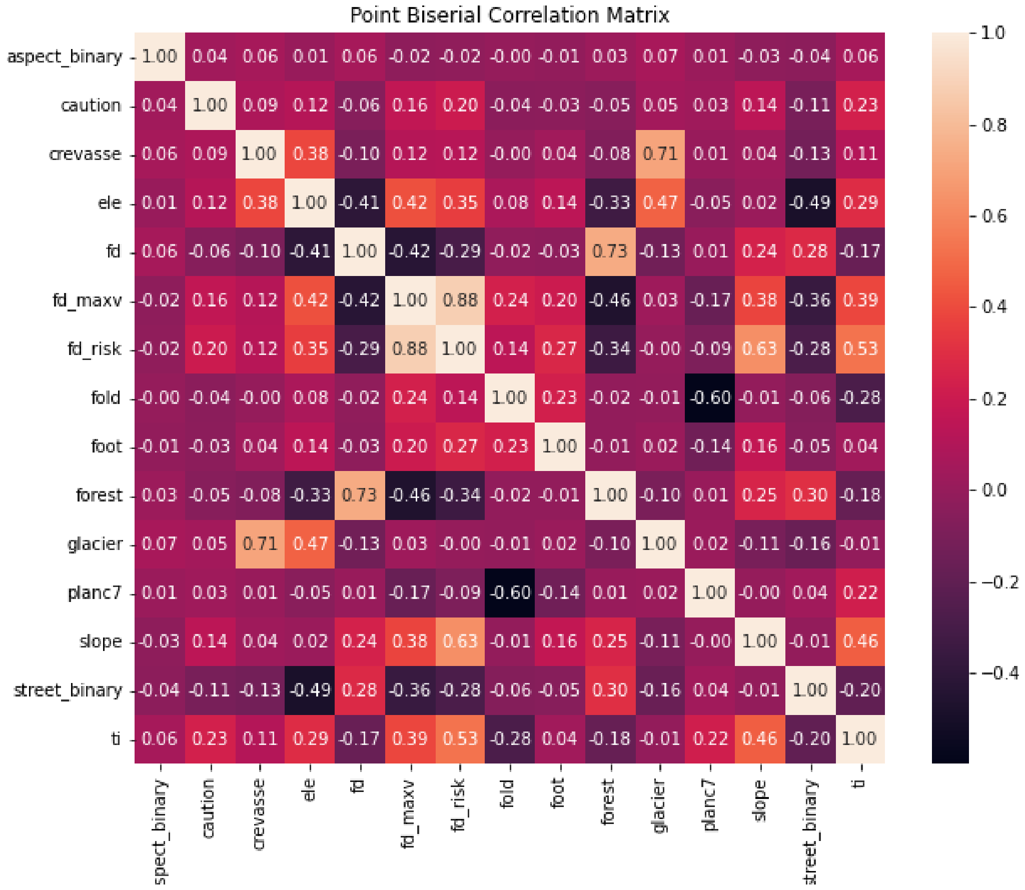
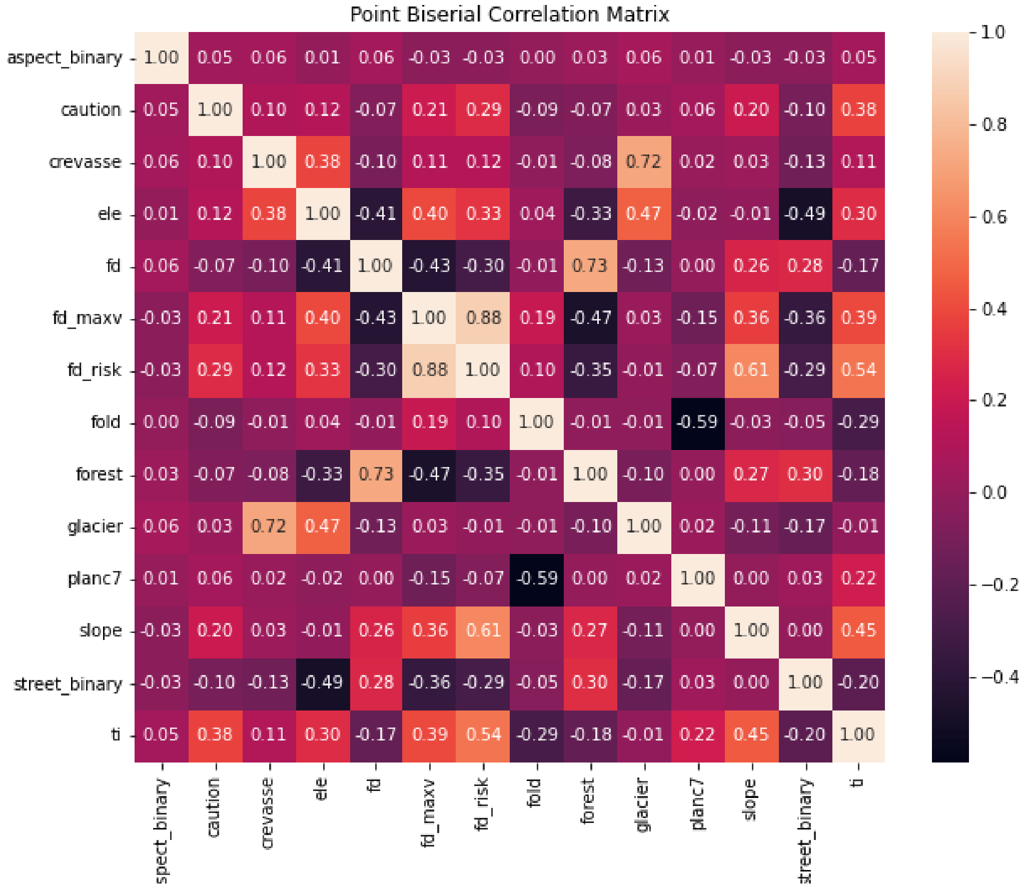


Figure 28: Correlation analysis of all variables using point biserial correlation, ***unfiltered*** *data*

The target variables are now included in the matrix. The target variable **{*caution*}** correlates slightly positive with {*ti* | r = 0.23}, {*fd_maxv* | r = 0.16}, {*fd_risk* | r = 0.20}, {*slope* | r = 0.14} and {*ele* | r = 0.12}. The other target variable ***{foot}*** also correlates slightly positive with {*fold* | r = 0.23}, {*fd_maxv* | r = 0.20}, {*fd_risk* | r = 0.27}, {*slope* | r = 0.16} and {*ele* | r = 0.14}. This implies that as the values of these features increase, the probability of the logit function also tends to increase. Thus, these feature variables might be good predictors of the target variables. On the other hand, the correlation coefficients are still relatively far away from 1, which only indicates a weak positive correlation. If only the features are considered, the matrix also illustrates a strong positive correlation between the features {*glacier*} and {*crevasse*} with an r of 0.71 as well as between the features forest density {*fd*} and {forest} with an r of 0.73. Correlation among the feature variables indicates collinearity, which means that the second variable provides little or no additional information and is therefore suboptimal for modelling. For this reason, it makes sense to exclude the binary variable {*glacier*} and the continuous variable forest density {*fd*}, and instead only use {*crevasse*} and {*forest*}. The preference for {*crevasse*} over {*glacier*} is also confirmed by the expert interview (see chapter 2.2.1).

In Figure 29, the correlation matrix is shown for the filtered dataset for the *caution*-sections. As previously mentioned, the *foot*-sections and the points where {*caution* = 1 and *ti* < 0.25} or {*caution* = 0 and ti > 0.75} are excluded when applying the ti-filter.



	spect_binary	caution	crevasse	ele	fd	fd_maxv	fd_risk	fold	forest	glacier	planc7	slope	street_binary	ti
aspect_binary	1.00	0.05	0.06	0.01	0.06	-0.03	-0.03	0.00	0.03	0.06	0.01	-0.03	-0.03	0.05
caution	0.05	1.00	0.10	0.12	-0.07	0.21	0.29	-0.09	-0.07	0.03	0.06	0.20	-0.10	0.38
crevasse	0.06	0.10	1.00	0.38	-0.10	0.11	0.12	-0.01	-0.08	0.72	0.02	0.03	-0.13	0.11
ele	0.01	0.12	0.38	1.00	-0.41	0.40	0.33	0.04	-0.33	0.47	-0.02	-0.01	-0.49	0.30
fd	0.06	-0.07	-0.10	-0.41	1.00	-0.43	-0.30	-0.01	0.73	-0.13	0.00	0.26	0.28	-0.17
fd_maxv	-0.03	0.21	0.11	0.40	-0.43	1.00	0.88	0.19	-0.47	0.03	-0.15	0.36	-0.36	0.39
fd_risk	-0.03	0.29	0.12	0.33	-0.30	0.88	1.00	0.10	-0.35	-0.01	-0.07	0.61	-0.29	0.54
fold	0.00	-0.09	-0.01	0.04	-0.01	0.19	0.10	1.00	-0.01	-0.01	-0.59	-0.03	-0.05	-0.29
forest	0.03	-0.07	-0.08	-0.33	0.73	-0.47	-0.35	-0.01	1.00	-0.10	0.00	0.27	0.30	-0.18
glacier	0.06	0.03	0.72	0.47	-0.13	0.03	-0.01	-0.01	-0.10	1.00	0.02	-0.11	-0.17	-0.01
planc7	0.01	0.06	0.02	-0.02	0.00	-0.15	-0.07	-0.59	0.00	0.02	1.00	0.00	0.03	0.22
slope	-0.03	0.20	0.03	-0.01	0.26	0.36	0.61	-0.03	0.27	-0.11	0.00	1.00	0.00	0.45
street_binary	-0.03	-0.10	-0.13	-0.49	0.28	-0.36	-0.29	-0.05	0.30	-0.17	0.03	0.00	1.00	-0.20
ti	0.05	0.38	0.11	0.30	-0.17	0.39	0.54	-0.29	-0.18	-0.01	0.22	0.45	-0.20	1.00

*Figure 29: Correlation analysis of all variables using point biserial correlation, **ti-filtered** data*

What can be recognised in the above correlation matrix of the filtered data is that the correlations tend to become stronger in connection with the target variable {*caution*}. Thanks to the filter, the feature variables {*ti* | r = 0.35} (+0.12) , {*fd_maxv* | r = 0.20} (+0.02), {*fd_risk* | r = 0.29} (+0.09) and {*slope* | r = 0.20} (+0.06) now correlate slightly stronger with the target variable **{*caution*}.** This is due to less noise in the data, which should have a positive effect on the modelling of the *caution*-sections.

The correlation coefficients in the previous three matrices only measure linear correlations. Thus, the correlation coefficients may miss out on non-linear relationships. That's why it is important to not only calculate the correlation coefficients but also visualize the variables against each other. Plotting the data visually provides a clearer understanding of the strength and direction of the relationship between variables (Géron, 2019). For these reasons, the continuous variables have been plotted against each other at the end of the `Jupyter Notebook`. There are no non-linear patterns that can be recognised in the scatter plots, and the rather weak linear correlations are further confirmed. The fact that the majority of the calculated correlation coefficients for the linear relationship lie in the range [-0.5, 0.5] therefore seems plausible. Especially when looking at the feature variables, it is not a bad thing if there are no correlations between them. If there are strong correlations between the feature variables, this could indicate collinearity, which is not desirable.

5.2.5 EDA results

The results of this EDA chapter are summarised in a **feature importance table** the appendix. The features that promise an influence on the target variable are labelled with *Yes* (+ = *rather promising*, + + = *very promising*) or *No* (− = *rather no influence to be expected*, − − = *almost certainly no influence to be expected*). In the event that a feature has an undetected, unexpected and non-linear influence on the target variable, all variables are still used as input in the modelling at the beginning. Models with different combinations of variables are trained and approaches with automatic variable selection such as *RFE* from *scikit-learn* are also tested. The most important features were defined for both {*caution*} and {*foot*} on the basis of the domain knowledge and the EDA results:

- most promising features for *caution*-modelling: {*crevasse*}, {*fd_risk*}, {*ti*}
- most promising features for *foot*-modelling: {*crevasse*}, {*fd_risk*}, {*fold*}, {*ele*}

When the *black box* models (random forest, gradient boosting) are trained, both a model with all features and a model with only the most promising features are trained and hypertuned. Because the number of permutations is increasing and thus also the computing power, the GAMs are modelled up to a maximum of four dimensions.

5.3 Modelling *caution*

As described in chapter 4 on methods, the Python script `main.py` and the associated functions in `my_functions.py` were used for modelling. The `GAM.R` script was executed directly via Python for modelling the GAMs. The models' evaluation scores were saved in the excel file caution_scores.xlsx.

5.3.1 Modelling

For the modelling of {*caution*}, the `main.py` script was executed a total of six times. Each session trained more than 1'000 models, the majority of them were linear regression models, about 50 GAMs, 2 random forests and 2 gradient boosting models. There have been relatively many logistic regression models trained, because they are relatively inexpensive to train in terms of computing power, and therefore a lot of different permutations were tried out. The permutations provided a certain amount of information as to which combinations are more suitable for linear modelling. For the GAMs, a limited number of predefined permutations were trained based on the prior domain knowledge, the results of the EDA and the results of the logistic regression with the most promising variables. Additionally, smoothers have been applied to different variables in each case. However, as the interpretability decreases with an increasing number of smoother, a maximum of two smoother were used. The GAM modelling started with a one-dimensional model up to a maximum four-dimensional model (focus mainly on permutations with promising variables). Interaction terms were also experimented with. Random forests and gradient boosting were relatively expensive in terms of computing power (long runtime), which is why only two models each (one with all variables, one with most promising variables) were trained. *GridSearch* was used for random forests and gradient boosting, where different hyperparameters were tried out and the best model (according to F1 score) was selected. Since the focus is not on the last two model types anyway, it is justifiable to train only a few of these models as benchmarks (no excessive hyperparameter tuning). Each of the six executions of the `main.py` script is characterised as follows (in brackets the average confusion score of the top 10 models):

- Run 1 – imbalanced data, without ti-filter (*avg. 489*)
- Run 2 – imbalanced data, with ti-filter (*avg. 319*)
- Run 3 – undersampled training data, with ti-filter (*avg. 321*)
- Run 4 – oversampled training data, with ti-filter (*avg. 326*)
- Run 5 – imbalanced data, with ti-filter, with scaling (*avg. 320*)
- Run 6 – imbalanced data, with ti-filter, with street-filter (*avg. 305*)

In the first run, the dataset was used, where only the necessary feature engineering tasks from chapter 5.1 were carried out. In the second run, the ti-filter (see chapter 5.2.3) was additionally applied to the dataset, which has already led to significantly better results probably thanks to less class noise and a stronger signal. Different feature engineering methods from *scikit-learn* were applied in runs 3 to 5. In these runs, over- and undersampling were both once applied to the training dataset to address the class imbalance. Both methods did not lead to significantly better scores, which is why these methods were not pursued any further. Even the standard scaling method did not result in better scores, where a scaler was fitted to the training data and then applied to the training, validation and test data (fitting the standard scaler only on training data because of data leakage). The last run achieved the best scores, where both the ti-filter and the street-filter were applied to the imbalanced data.

5.3.2 Model comparison

Because the 6th run was the most promising in terms of scores (imbalanced, with ti-filter, with street-filter), the scores for this run are shown graphically below. The application of the two filters is justified since the ti-filter was discussed with Günter Schmudlach (pragmatic addressing of class noise) and the street-filter is to a certain extent comprehensible because, according to the interview with Andreas Eisenhut, points on a street are never $\{caution = 1\}$. If the model were to be used once in production, all data points where $\{street\_binary = 1\}$ applies should subsequently be excluded before modelling and afterwards labelled with $\{caution = 0\}$, as these points were not included in the modelling, but are always $\{caution = 0\}$ according to expert knowledge.

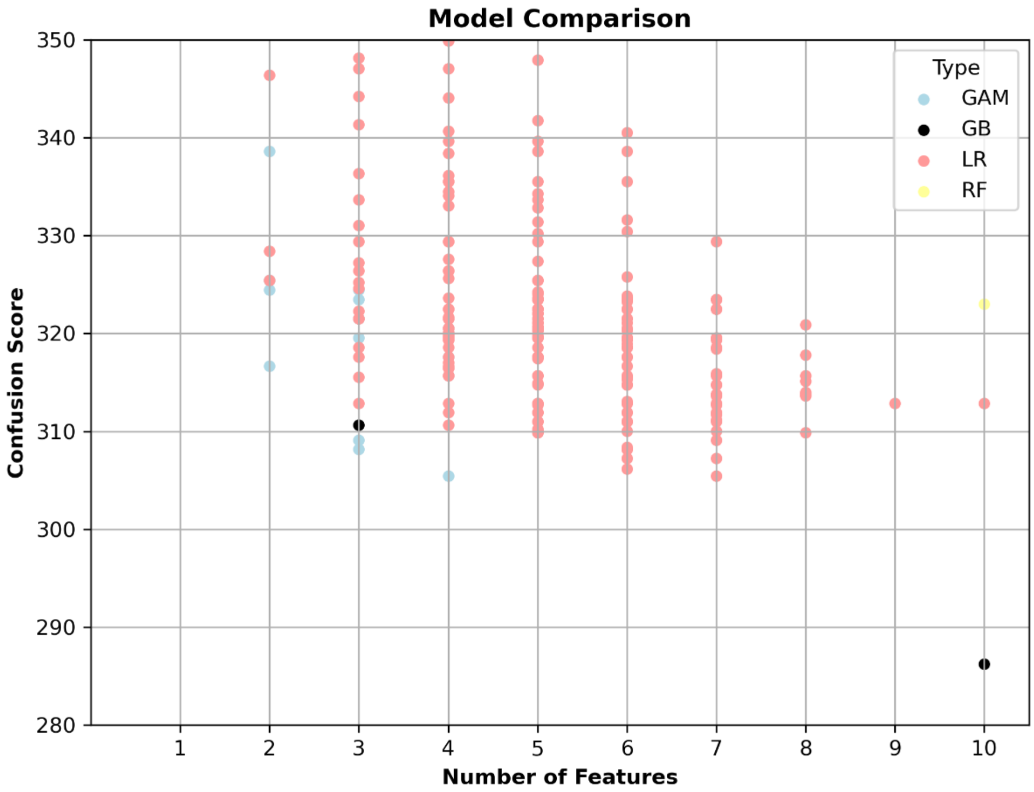


Figure 30: Model comparison for run 6 - imbalanced, with ti-filter and street-filter (score window 280 – 350)

Figure 30 above shows that the best confusion score of 286 was achieved by a gradient boosting model with all ten variables. A negative correlation between confusion score and number of variables is recognisable. The more variables selected, the lower (better) the confusion score tends to be. But it can be observed that the added value of an additional variable is not particularly great, as the scores are still over 300 even with more than three variables. The performance of the random forest is also rather disappointing, as the best of its model only achieves a confusion score of 323 (and was outperformed by some logistic regressions). Some GAMs appear to have performed fairly well, achieving a relatively low confusion score between 305 and 320 with only two to three variables.

Based on the domain knowledge and prior EDA knowledge, the three variables {*crevasse*}, {*fd_risk*} and {*ti*} were determined as the most promising ones for *caution*-modelling (see chapter 5.2.5). The question now arises as to what extent these variables can be confirmed in the modelling. In the figure below, the performance of the trained models is plotted, whereby a model is highlighted in red if it contains the corresponding promising feature. For example, models that contain the feature {*ti*} are plotted in red in the left-hand plot. There is a recognisable pattern that models with the variable {*ti*} tend to yield a better confusion score (this applies in particular to one- and two-dimensional models). This can also be observed for the variable {*fd_risk*}. The pattern is not so clear for the variable {*crevasse*}, which is only found more frequently in the better scores for the three-dimensional models and above.

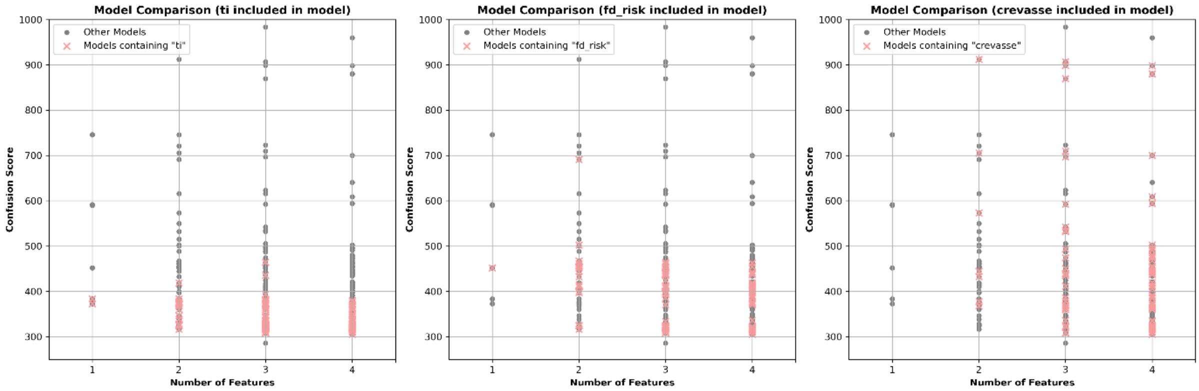


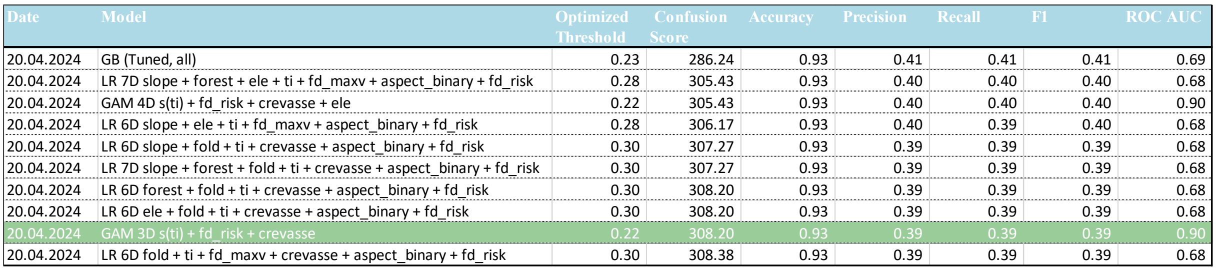
Figure 31: Analysis of the impact of the most promising EDA variables (score window 250 – 1'000)

Most models achieve an accuracy greater than 0.90, but this is largely due to the correct classification of true negatives. As this is an imbalanced classification problem, it makes little sense to consider accuracy. This score says nothing about how well a model can predict the desired class {*caution* = *1*}. The average precision score is 0.34, with the gradient boosting model achieving a score of 0.41. This indicates that the models have difficulties predicting the class of interest reliably enough at the selected *p*-threshold.

5.3.3 Model selection

Now that an overview of the scores of the individual models has been obtained, a winning model can be determined. According to the rule of Occam's razor, of several possible adequate explanations for one and the same fact, the simplest theory is preferable to all others. A theory is simple if it contains as few variables and hypotheses as possible and if these are logically related to each other, from which the facts to be explained follow (Bruce et al., 2020). This has the particular advantage that the results can be communicated to stakeholders, who may not be familiar with machine learning and data science.

The best 10 models (according to confusion score) are listed in Figure 32. The other metrics (precision, recall, F1 score) correlate strongly with the confusion score, which is due to their calculation methods. As already mentioned, the black box models should only be considered as a benchmark. Only a logistic regression model or a GAM can be considered as the winner. Among these models, the deviation in the performance metrics is relatively small (confusion score varies between 305 and 309). For this reason, the winner is the simplest model. The GAM with the 3 features s({*ti*}), {*fd_risk*} and {*crevasse*} is therefore favoured and determined as winner over the GAM with 4 features and the logistic regression models with 6 to 7 features. The coefficients of the model are described in more detail in the discussion in chapter 6.2.



Date	Model	Optimized Threshold	Confusion Score	Accuracy	Precision	Recall	F1	ROC AUC
20.04.2024	GB (Tuned, all)	0.23	286.24	0.93	0.41	0.41	0.41	0.69
20.04.2024	LR 7D slope + forest + ele + ti + fd_maxv + aspect_binary + fd_risk	0.28	305.43	0.93	0.40	0.40	0.40	0.68
20.04.2024	GAM 4D s(ti) + fd_risk + crevasse + ele	0.22	305.43	0.93	0.40	0.40	0.40	0.90
20.04.2024	LR 6D slope + ele + ti + fd_maxv + aspect_binary + fd_risk	0.28	306.17	0.93	0.40	0.39	0.40	0.68
20.04.2024	LR 6D slope + fold + ti + crevasse + aspect_binary + fd_risk	0.30	307.27	0.93	0.39	0.39	0.39	0.68
20.04.2024	LR 7D slope + forest + fold + ti + crevasse + aspect_binary + fd_risk	0.30	307.27	0.93	0.39	0.39	0.39	0.68
20.04.2024	LR 6D forest + fold + ti + crevasse + aspect_binary + fd_risk	0.30	308.20	0.93	0.39	0.39	0.39	0.68
20.04.2024	LR 6D ele + fold + ti + crevasse + aspect_binary + fd_risk	0.30	308.20	0.93	0.39	0.39	0.39	0.68
20.04.2024	GAM 3D s(ti) + fd_risk + crevasse	0.22	308.20	0.93	0.39	0.39	0.39	0.90
20.04.2024	LR 6D fold + ti + fd_maxv + crevasse + aspect_binary + fd_risk	0.30	308.38	0.93	0.39	0.39	0.39	0.68

Figure 32: Ranking {caution}-models (top 10 by confusion score)

If a two-dimensional model had to be chosen as the winner for the selection, then the GAM model with the features s({*ti*}) and {*fd_risk*} would be the winner. This model has a confusion score of 316, which is why, depending on the trade-off between performance and simplicity, it could also be chosen as the winner. The best two-dimensional model for a logistic regression would be a model with the features {*ti*} and {*fd_risk*}, which leads to a confusion score of 325. The best one-dimensional model is a logistic regression model with {*ti*} as feature, resulting in a confusion score of 372. The model performance on the test set is evaluated in chapter 6.2.

5.4 Modelling *foot*

The target variable {*foot*} was modelled in the same way as {*caution*} in the previous chapter, i.e. with help of the scripts `main.py`, `my_functions.py` and `GAM.R` as described in the methods chapter. This means that also more than 1'000 models were trained in each run. The models' evaluation scores were listed in the excel file foot_scores.xlsx.

5.4.1 Modelling

However, the characteristics of the individual runs are different compared to the *caution*-modelling because, for example, according to the interview with Andreas Eisenhut, the branch-wise marking on the *foot*-sections is not an issue at all. The ti-filter was therefore never used in the modelling of {*foot*}. What could be recognised by studying the map, however, is that if a section of the route goes through a tunnel, these passages are almost always marked as a *foot*-passage (which makes sense, as there is never snow in longer tunnels). In their attributes, however, these points are often harmless (suggesting {*foot* = 0} instead of {*foot* = 1}). For this potential reason for class noise, all tunnel sections were excluded in the third run (i.e. points where {*foot* = 1 and *street_binary* = 1}. Because the over- and undersampling techniques were not promising in the previous modelling for {*caution*}, these methods were no longer pursued for the modelling of {*foot*}. In concrete terms, the single runs look like the following (in brackets the average confusion score of the top 10 models):

- Run 1 - imbalanced (*avg. 173*)
- Run 2 - imbalanced, with standard scaling (*avg. 172*)
- Run 3 - imbalanced, with tunnel-filter (*avg. 147*)
- Run 4 - imbalanced, with street-filter (*avg. 148*)

Here, as well, the filters were used very carefully, as depending on the definition of the filters, a very strong bias can be introduced into the trained model. This much can be said in anticipation that the modelling for {*foot*} tended to be more precise than the modelling of {*caution*}, as the following chapter shows. In the second run, standard scaling did not lead to a better performance of the models on the validation dataset. The tunnel filter in the 3rd run, on the other hand, led to significantly better results, as the points located in a tunnel and labelled as *foot*-sections were excluded from the dataset. This probably reduced the class noise somewhat, which led to more precise predictions. When all points lying on a street (as well as the {*street_binary*} variable) were excluded with the street-filter in the 4th run, no noticeably better performance was achieved than with the tunnel-filter.

5.4.2 Model comparison

When looking at the scores in Figure 33, it is interesting to note that the best score was not achieved by a black box model but by a GAM with a confusion score of 141. Gradient boosting achieved the second best confusion score with 144 points, but had to use ten features. The performance of the GAM should be recognised in that it is built with only 3 features. In general, GAMs with two to three variables perform very well in relative terms, and there is also a recognisable tendency for additional variables to add only marginal predictive value.

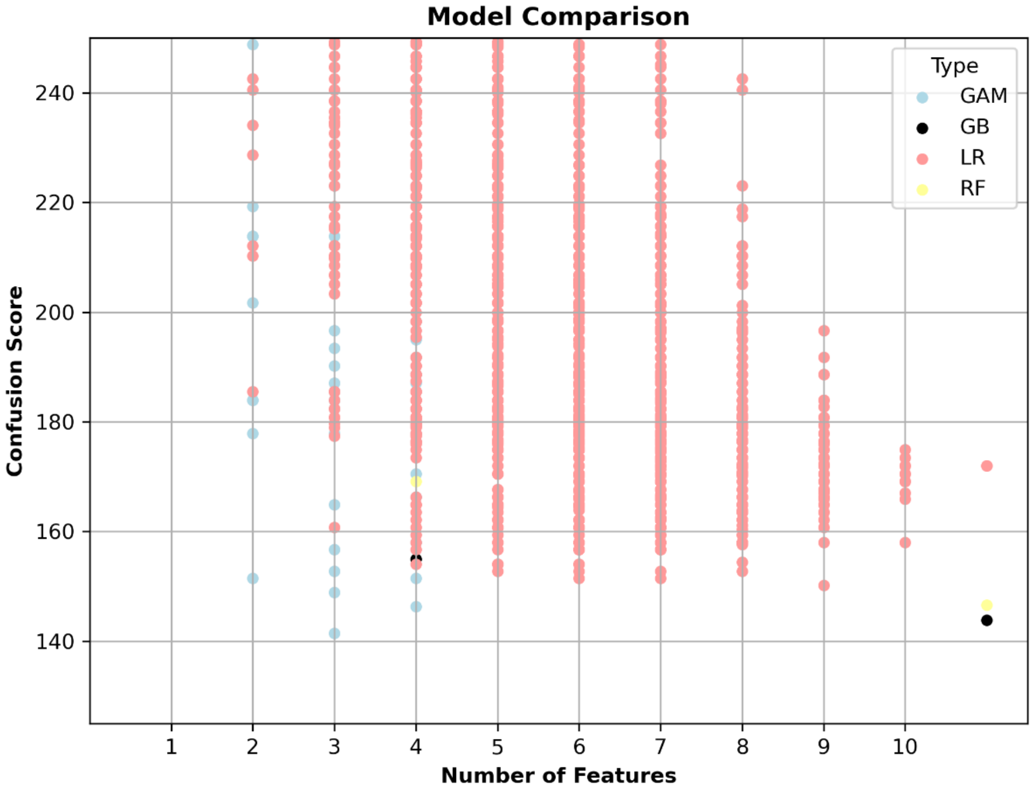


Figure 33: Model comparison for run 3 - imbalanced, with tunnel-filter (score window 125 – 250)

Based on the domain and expert knowledge and the EDA chapter, {*crevasse*}, {*fd_risk*}, {*fold*} and {*ele*} were defined as the most promising variables for the *foot*-modelling. Figure 34 now shows how often these four promising variables are listed in the models. It would be expected that these tend to be listed in low-dimensional models that have a good confusion score.

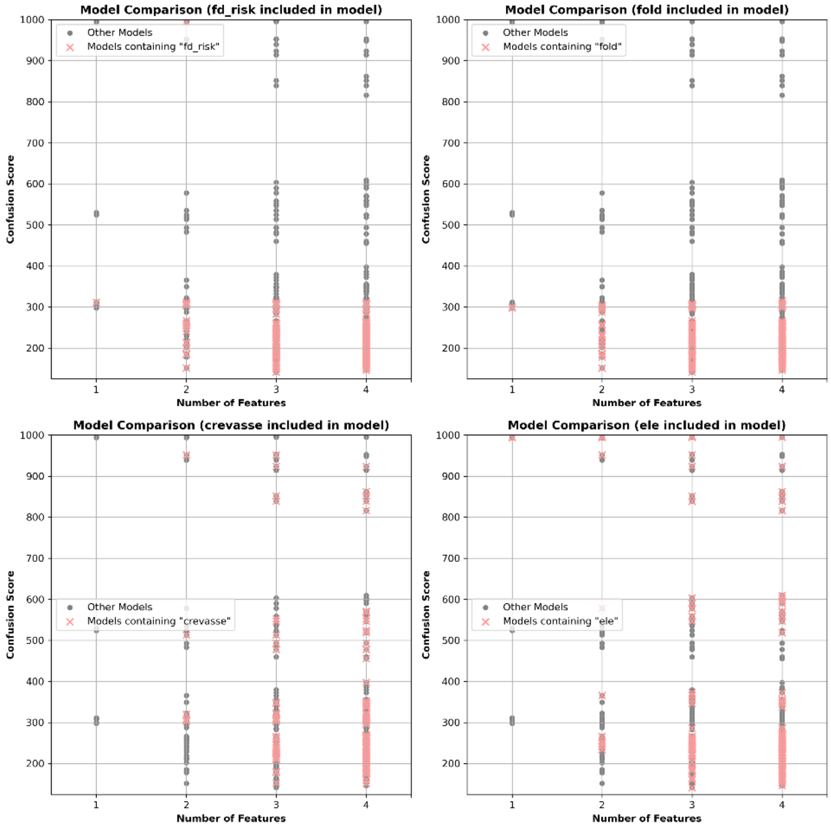
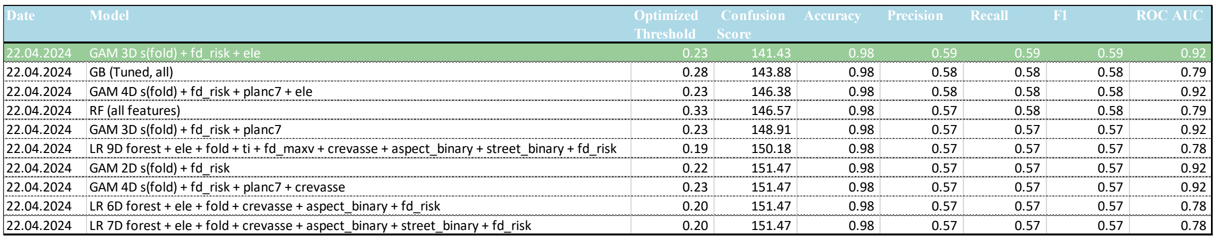


Figure 34: Analysis of the impact of the most promising EDA variables (score window 150 – 1'000)

It is indeed recognisable that the promising variables {*fd_risk*} and {*fold*} are very often present in the models that have a relatively low confusion score. In contrast, the pattern is less pronounced for {*crevasse*} and {*ele*}, which rather indicates that these two variables provide less additional information. The following chapter takes a closer look at the ten best *foot*-models.

5.4.3 Model selection

What is interesting in the ranking below is that the GAMs in the top 10 contain almost the same variables. However, the addition of {*planc7*} does not seem to bring any additional benefit, but leads to confusion. The GAM with the three variables {*fold*} (with smoother), {*fd_risk*} and {*ele*} is therefore chosen as the winner. However, depending on the trade-off between precision and simplicity, one could also argue that the two-dimensional model with {*fold*} (with smoother) and {*fd_risk*} is the winner.



Date	Model	Optimized Threshold	Confusion Score	Accuracy	Precision	Recall	F1	ROC AUC
22.04.2024	GAM 3D s(fold) + fd_risk + ele	0.23	141.43	0.98	0.59	0.59	0.59	0.92
22.04.2024	GB (Tuned, all)	0.28	143.88	0.98	0.58	0.58	0.58	0.79
22.04.2024	GAM 4D s(fold) + fd_risk + planc7 + ele	0.23	146.38	0.98	0.58	0.58	0.58	0.92
22.04.2024	RF (all features)	0.33	146.57	0.98	0.57	0.58	0.58	0.79
22.04.2024	GAM 3D s(fold) + fd_risk + planc7	0.23	148.91	0.98	0.57	0.57	0.57	0.92
22.04.2024	LR 9D forest + ele + fold + ti + fd_maxv + crevasse + aspect_binary + street_binary + fd_risk	0.19	150.18	0.98	0.57	0.57	0.57	0.78
22.04.2024	GAM 2D s(fold) + fd_risk	0.22	151.47	0.98	0.57	0.57	0.57	0.92
22.04.2024	GAM 4D s(fold) + fd_risk + planc7 + crevasse	0.23	151.47	0.98	0.57	0.57	0.57	0.92
22.04.2024	LR 6D forest + ele + fold + crevasse + aspect_binary + fd_risk	0.20	151.47	0.98	0.57	0.57	0.57	0.78
22.04.2024	LR 7D forest + ele + fold + crevasse + aspect_binary + street_binary + fd_risk	0.20	151.47	0.98	0.57	0.57	0.57	0.78

Figure 35: Ranking {foot}-models (top 10 by confusion score)

Among the top ten, the scores are generally relatively similar, because they correlate with each other (similar precision, recall, F1 score, etc.). Overall, the accuracy of the model is very high at 0.98, which means that 98% of the predictions are correct. However, the precision of 0.59 of the winner model indicates that the model struggles to reliable classify the class of interest. Moreover, among the logistic regression models, only models that contain significantly more than three variables are represented in the top ranks. A two-dimensional logistic regression model with {*fold*} and {*fd_risk*} has a confusion score of 185 and a precision of 0.51.When {*ele*} is added, the logistic regression model has a slightly better confusion score of 161 and a precision of 0.55. The model performance on the test set is discussed in the next chapter.

What is interesting for the modelling of both targets {*foot*} and {*caution*} is that the automatic feature selection function *RFE* from the library *scikit-learn* had not worked well for the imbalanced classification problem. The log files show that {*aspect_binary*}, {*forest*} and {*street_binary*} were often selected by *RFE*, perhaps because the majority class could be classified well with these features. However, the variables of the winner models were rarely selected by *RFE*. In other words, the features selected by *RFE* do not yield a good precision score for the model. This phenomena therefore emphasises the importance of a solid domain knowledge of the problem and an extensive explanatory data analysis.

6 Discussion and conclusion

The research questions are discussed in this chapter. On the one hand, the *foot*-winner model is analysed, in particular with regard to the choice of the optimal *p*-threshold and the associated trade-off between precision and recall. In the second research question, an attempt is made in particular to identify boundary values as a recommendation for the SAC in order to improve their data consistency.

6.1 Prediction of *foot*-sections with a machine learning approach

The first research question was about the extent to which *foot*-sections can be classified using a machine learning approach. The model was trained on the data with the tunnel-filter (class noise reduction). Figure 36 shows the performance of the winning GAM with the three features {*ele*}, {*fd_risk*} and {*fold*}, with a smoother on the last feature.

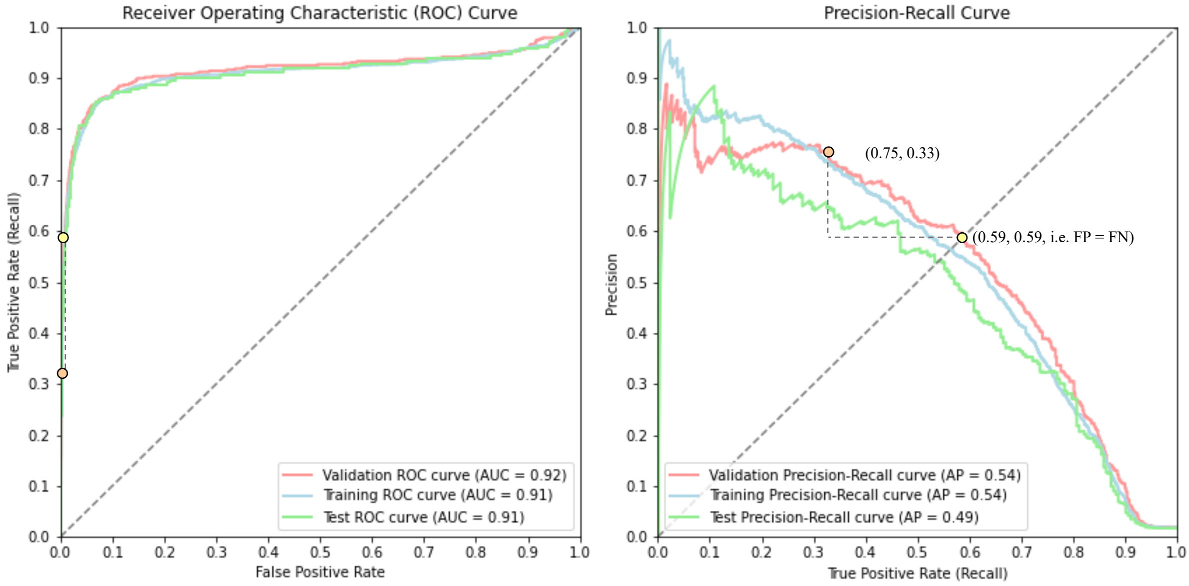


Figure 36: Performance on training, test and validation set for the winning foot-model

On the left, the ROC curve shows the TP-rate (recall) on the y-axis and the FP-rate on the x-axis. The diagonal represents random guessing. The best result would be at the top left, where a model achieves a high TP-rate and a low FP-rate. In general, it can be seen that the ROC curve for all three datasets is close to each other. This indicates a robust classifier, which seems not to overfit the training data. Overall, the *foot*-model is better than random guessing (accuracy of 0.98), but it struggles to predict the class of interest {*foot* = 1} at the *p*-threshold where FP = FN. The 'ideal' model would be at the point closest to the top left corner. But then the assumption of a balanced model, where FP = FN results, would be violated. If an attempt is made to adjust *p*-threshold to optimise the precision (TP / (TP + FP)), as in the right-hand plot from 0.59 to 0.75, then the trade-off becomes apparent. The model would then only predict positive classes if it is very certain. It is therefore obvious that the true positive rate or recall (TP / (TP + FN)) will then fall. The plot on the right shows how this falls from 0.59 to 0.33. The same pattern, but in the opposite direction, is true if the recall (TP / (TP + FN)) is trying to be optimized.

The confusion score of 142 (rounded up) can be interpreted to mean that if there are 100 TP, then 71 FP and 71 FN are predicted by the model. The *p*-threshold determines whether the model predicts more FP or FN. From a conservative point of view, it could make sense to increase recall at the expense of precision (accept more FP). If the model is then used for the marking of the *foot*-sections, the predictions must be validated by experts. Even if the model performance with regard to the class of interest {*foot* = 1} has a certain degree of uncertainty, interesting facts can be gained from the winner model. For example, the importance of the features {*fold*} (which is probably related to the ridges), {*fd_risk*} (highly exposed areas) and {*ele*} (high-alpine terrain) is a strong confirmation of the layers created by Günter Schmudlach (especially the layer for the risk of falling) on the one hand and the ski touring literature on the other. In order to make the model more precise, it would also be interesting to analyse and clean the *foot*-sections for further class noise. For example, section-wise approaches, where a section (e.g. 100 meters) is predicted instead of a point, could smooth out class outliers. For the sampling procedure, points would in this approach be sampled every ten meters, where afterwards the median would be calculated. The precision can also be modified by selecting a different *p*-threshold based on expert knowledge and thus interpreting the trade-off differently. The class noise could be further addressed with clustering approaches such as DBSCAN (density based clustering) to remove potential outliers. However, this would also need to be done with expert knowledge, as this could introduce a potentially very high bias in the dataset by altering the ground truth.

6.2 Consistency of *caution*-sections in the SAC ski touring data

The second research question was about to what extent the existing *caution*-sections in the SAC dataset are consistent or not. The analysis of the modelling showed that there are definitely inconsistencies due to strong class noise in the dataset. Even the winning non-linear GAM model with an accuracy of 0.93 still only has a precision of 0.39 at the *p*-threshold where FP = FN (yellow point in Figure 37).

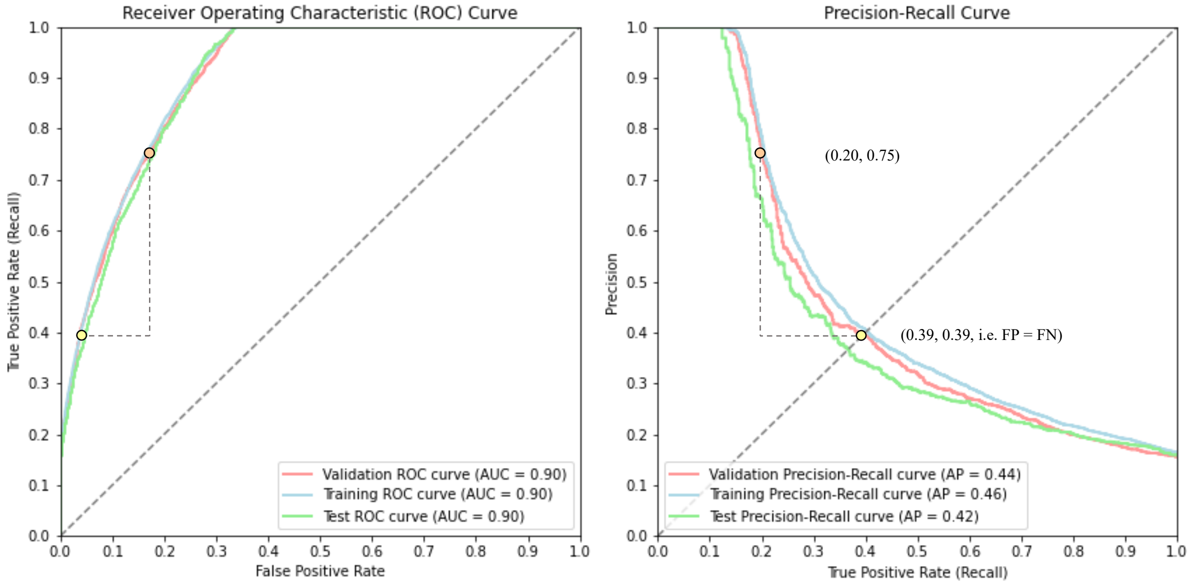


Figure 37: Performance on training, validation and test set for the winning caution-model

In contrast to the *foot*-sections, however, the trade-off between precision and recall appears to be less pronounced when modelling $\{caution\}$. If *p*-threshold is adjusted such that a precision of 0.75 results (orange point in Figure 37), the recall decreases from 0.39 to around 0.20 (right plot). However, what is much more significant is the change in the false positive rate (left plot). This increases from around 0.03 to 0.18. The effect of the change in the FP-rate (when recall is increased) now appears to be relatively strong, whereas it was less strong when modelling the *foot*-sections in the previous chapter. The *caution*-model also struggles, even more so than the *foot*-model, to predict the underrepresented class reliably. Nevertheless, it is possible to roughly determine from the data where the boundary values for the most important features for *caution*-modelling would lie in a one-dimensional model, as shown below. These boundary values may help the SAC to question existing markings. If a rule-based section-marking framework would be created in the future, these boundaries may be used as rough guide values. Thus, for a simple one-dimensional logistic regression model, the estimated parameters (intercept, coefficient) can be used to calculate the boundary value at a certain *p*-threshold (i.e. the one where FP = FN) from which a point is $\{caution = 1\}$. For the variables from the winner model for $\{caution\}$, these boundary values were calculated, i.e. for $\{crevasse\}$, $\{fd\_risk\}$ and $\{ti\}$. The calculations were performed in the Script `winners.R` and documented in the `summaries` directory. The values in the following are rounded to two decimals (exact values for recalculation are in the `summaries` directory). It is important to recall that the variable $\{ti\}$ was used both as a filter and as a feature for the *caution*-model. The filter values were therefore chosen carefully (see chapter 5.2.3).

$$logit(p) = a + b * \{variable\ value\}$$

Equation 10: one-dimensional logistic regression model

$$\{variable\ value\} = \frac{(logit(p) - a)}{b}$$

Equation 11: inverse of one-dimensional logistic regression model solved for variable boundary-value (for given p, a and b)

For the variable **{*crevasse*}**, according to a simple one-dimensional logistic regression model (only fitted on data points that lie on a glacier), a prediction is $\{caution = 1\}$ if the variable has a value of **5** or higher. This boundary value has been calculated with an estimated coefficient b of 0.62 and an intercept a of -4.22, where *p*-threshold is set to 0.25 (FP = FN).

For the variable **{*ti*}**, the resulting boundary value is **0.66**. In other words, if the feature has a value of 0.66 or higher, the simple logistic regression model would label the point as $\{caution = 1\}$. Before the parameter estimation, the street points were excluded, as Andreas Eisenhut suggested. Thus, the estimated parameters are 7.08 (coefficient b) and -5.51 (intercept a) at a *p*-threshold of 0.31 (FP = FN).

If the third variable **{*fd_risk*}** is considered, a boundary value of **1'079** can be calculated at a *p*-threshold of 0.19 (FP = FN). For the calculation, the estimated parameters are 0.003 (coefficient b) and -4.35 (intercept a). If the feature has a value of 1'079 or more, the points are predicted as $\{caution = 1\}$. This parameter estimation was also calculated on the datasets without streets.

These estimates provide a rough indication, but should be treated with care, firstly because they were still estimated on a dataset characterised by class noise (estimate tends to be too conservative). In addition, a one-dimensional logistic regression model is a very strong simplification and not as precise as a multidimensional GAM. And finally, the estimated values also depend on the selected *p*-thresholds, which were chosen at the point FP = FN. With the help of the above boundary values and additional expert knowledge, *caution*-sections could be defined entirely without machine learning, namely with the following if/else approach:

- **if** {*ti* >= 0.66}, then label data point as {*caution* = 1}, **else** check…
- … **if** {*fd_risk* >= 1'079}, then label data point as {*caution* = 1}, **else** check…
- … **if** {*crevasse* >= 5}, then label data point as {*caution* = 1}, **else** label data point as {*caution* = 0}

With the if/else approach above, a very simple rule-based classification would be implemented, from where the boundary values are derived from the SAC dataset. The boundary values would of course have to be confirmed or further adjusted by experts, but the implementation would lead to a much more consistent SAC dataset. Moving away from a previously subjective approach, characterised by many different mountain guide opinions, towards a rule-based, transparent and comprehensible approach. When modelling {*caution*}, the winner model was less precise than the *foot*-model. What is nevertheless very promising is the fact that the SAC data confirmed the previous interview and literature research. According to the interview, it is in particular the avalanche risk (represented in the data by {*ti*}), the fall risk (represented in the data by {*fd_risk*}) and the crevasse risk (represented in the data by {*crevasse*}). The winning *caution*-model prevailed with exactly these three variables, which is astonishing. This suggests that a reliable model could probably be trained with a dataset characterised by less noise. The statement from the interview that forestation has a small or contradictory influence on a *caution*-section is also reflected in the modelling results, as the feature {*forest*} is not used in the final model.

6.3 Reflexion and outlook

In the model selection, both easily communicable models and black box models were trained. Even if the winning models have a high accuracy, the models should be integrated into a framework where experts still validate the predictions of the models due to the lack of precision. Accepting more FP than FN would make the model predictions conservative (principle of caution), where the experts would tend to have to remove some of the model's predicted *foot*-sections because they were predicted too generously. What could lead to even better results, however, would be to cluster the data beforehand in order to get rid of even more class noise (i.e. outlier detection with DBSCAN). The problem with this, however, is that if someone starts to exclude points from the entire dataset, the ground truth will be manipulated. On the basis of these results, it would then be tricky to generalise to the valuation principles applied by the SAC (because the ground truth shifts away from the initial man-made SAC markings).

If, on the other hand, clustering is only performed on the training dataset, the model quality can only ever be tested using the noisy validation data, which is unlikely to provide promising results. One possibility would therefore be to obtain a dataset of a selection of routes from the SAC itself that does not contain any class noise (especially no branch-wise markings) and is therefore very close to the ground truth. A model would then be trained on the basis of the clustered training data and the performance would be tested using the accurate 'clean' route set created by the SAC. Another promising alternative that was also tried out briefly in the master thesis would be to predict route sections instead of route points. In consultation with Günter Schmudlach, however, this approach was not pursued further, as the focus is on point estimations (and not sections). This would tend to average out the class noise. However, the output of the model would then be subject to greater uncertainty because, for example, although 25 sections could be estimated accurately, they still would only reflect average terrain features. The thesis has confirmed much of the ski touring risk literature by analysing and modelling the SAC data, but further sophisticated data cleaning practices or expert knowledge is required to address the class noise. Therefore, the predictions of the current models should always be validated by experts afterwards, but can already make a very valuable contribution to a higher consistency by providing a pre-selection of the markings.

6.4 Acknowledgement

It was a great opportunity for me to deal with the topic of risks in ski touring for a year as part of my master thesis. Dealing with ski tours on the one hand and geodata on the other was fun. But I was particularly pleased to have carried out this master thesis project with inspiring people. Special thanks go to Günter Schmudlach, the man behind Skitourenguru, who made this master thesis possible. He also invested several hours in coaching to give me the necessary knowledge about the data and its sampling. I could have listened to him talking about the Skitourenguru project for hours. Another highlight for me was the ski tour together with him to the Winterhorn in Uri, where I was able to gain even more knowledge about ski touring in real-life terrain. I have learnt a lot from the thesis for my future ski tours. In a broader sense, Günter Schmudlach gave me a lot of inspiration and new perspectives on my educational and professional path, for which I am very grateful, because I have been dealing with this question a lot over the last year. Kudos also to Martin Rumo from Lucerne University of Applied Sciences and Arts, who took the time to supervise this project scientifically. I would also like to thank him for the pleasant cooperation and his personal interest in the topic of the master thesis. Andreas Eisenhut from the SAC also made a valuable contribution with his time for the interview, filling in gaps in the literature and providing valuable information on the swisstopo dataset. He also deserves many thanks. Not to forget all my close friends who supported me during this master thesis project.

Appendix

GitHub repository

https://github.com/skitourenguru/RoutesProperties/tree/main

Code

`main.py`

https://github.com/skitourenguru/RoutesProperties/blob/main/code/main.py

`my_functions.py`

https://github.com/skitourenguru/RoutesProperties/blob/main/code/my_functions.py

`GAM.R`

https://github.com/skitourenguru/RoutesProperties/blob/main/code/GAM.R

`winners.R`

https://github.com/skitourenguru/RoutesProperties/blob/main/code/winners.R

Resampled SAC/swisstopo data

https://github.com/skitourenguru/RoutesProperties/tree/main/data

Requirements

Python Version: 3.10.7

R Version: 4.2.3

`requirements.txt`

https://github.com/skitourenguru/RoutesProperties/blob/main/code/requirements.txt

Creative common license

The following license is integrated in the GitHub repository:

https://creativecommons.org/licenses/by-sa/4.0/deed.en

The Creative Commons Attribution-ShareAlike 4.0 International Public License (CC BY-SA 4.0) allows others to use, share, and modify the work for any purpose, even commercially, as long as they give credit to the authors and release their new creations under the same terms. It ensures that the work remains freely available and adaptable while requiring users to share alike. There are conditions regarding attribution, sharing adaptations, and no additional restrictions. The license also includes disclaimers of warranties and limitations of liability.

Bibliography

Analytics Vidhya (online). *Feature Scaling: Engineering, Normalization, and Standardization (Updated 2024).* Retrieved from https://www.analyticsvidhya.com/blog/2020/04/feature-scaling- machine-learning-normalization-standardization/

Beratungsstelle für Unfallverhütung (BFU) (2023). *Skitouren. Sicher auf verschiedene Gipfel*. Retrieved from https://www.bfu.ch/de/ratgeber/skitouren

Beratungsstelle für Unfallverhütung (BFU) (2023). *Status 2023: Statistik der Nichtberufsunfälle und des Sicherheitsniveaus in der Schweiz*. Retrieved from https://www.bfu.ch/api/publications /bfu_2.505.01_Status%202023%20%E2%80%93%20Statistik%20der%20Nichtberufsunf%C3%A4lle%20und%20des%20Sicherheitsniveaus%20in%20der%20Schweiz.pdf

Winkler, K., Brehm, H.-P. & Haltmeier, J. (2023). *Bergsport Winter. Technik. Taktik. Sicherheit*. Thun/Gwatt: Weber Verlag.

Bruce, P., Bruce, A. (2020). *Practical Statistics for Data Scientists: 50 Essential Concepts*. United States: O'Reilly Media.

Burger, S. V. (2018). *Introduction to Machine Learning with R: Rigorous Mathematical Analysis*. Taiwan: O'Reilly Media.

Deisenroth, M. P., Faisal, A. A., Ong, C. S. (2020*). Mathematics for Machine Learning*. United Kingdom: Cambridge University Press.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. United States: O'Reilly Media.

Harvey, S., Rhyner, H., Schweizer, J. (2023). *Lawinen. Verstehen, beurteilen und risikobasiert entscheiden.* Outdoor-Praxis.

Huyen, C. (2022). *Designing Machine Learning Systems: An Iterative Process for Production-ready Applications.* O'Reilly Media.

McKinney, W. (2022). *Python for Data Analysis. Data Wrangling with Pandas, NumPy, and Jupyter*. United States: O'Reilly Media.

Medium (online). *A comparison of sklearn and statsmodel's logistic regression function*. Retrieved from https://ai.plainenglish.io/a-comparison-of-sklearn-and-statsmodels-logistic-regression-function-4340e9fd29dd

Mersch, J., Fleischmann, M., Mittermayer, H. (2021). *Lawinen: Erkennen - Beurteilen - Vermeiden*. Austria: BERGWELTEN.

Munter, W. (2023). *3x3 Lawinen. Risikomanagement im Wintersport (7. Auflage)*. Italy: TAPPEINER.

Neue Zürcher Zeitung (online). *«Die Opfer haben alles unternommen, um zu überleben»: Protokoll des Skitouren-Dramas im Wallis*. Retrived from https://www.nzz.ch/schweiz/fuenf-von-sechs-vermissten-skitourengaenger-im-wallis-tot-aufgefunden-ld.1821600

Ravanna, A., Saxena, S., Thompson, J., Pinheiro, C. (2020). *Machine Learning Using SAS® Viya. Course Notes*. United States: SAS Institute.

Reis, J., Housley, M. (2022). *Fundamentals of Data Engineering: Plan and Build Robust Data Systems*. Japan: O'Reilly Media.

Rumsey, D. (2011). *Statistics for dummies.* Indianapolis: Wiley Publishing.

Schmudlach, G. (2022). *Avalanche Risk Property Dataset (ARPD). User Manual. V3.1.2.* Retrieved from https://info.skitourenguru.ch/download/data/ARPD_Manual_3.1.2.pdf

Schmudlach, G. (2023). *Skitourenguru im Interview*. Retrieved from https://www.alpin.de/home /interviews/55697/artikel_guenter-schmudlach-eigenverantwortung -kann-man-nicht-durch-einen-algorithmus-ersetzen.html

SciPy (online). *Scipy. Point biserial*. Retrived from https://docs.scipy.org/doc/scipy/reference/ generated/scipy.stats.pointbiserialr.html

Skitourenguru (2023). *Absturzrisiko. Was bedeutet Absturzrisiko?* Retrieved from https://info.skitourenguru.ch/index.php/fdrisk

Tanadini, M. (2023a). *Cross validation Lab 7.* [course script], Lucerne University of Applied Sciences and Arts.

Tanadini, M. (2023b). *Extending the Linear Model Lab 6: Generalised Linear Models Lab*. [course script], Lucerne University of Applied Sciences and Arts.

Tanadini, M. (2023c). *Non-linearity Lab.* [course script], Lucerne University of Applied Sciences and Arts.

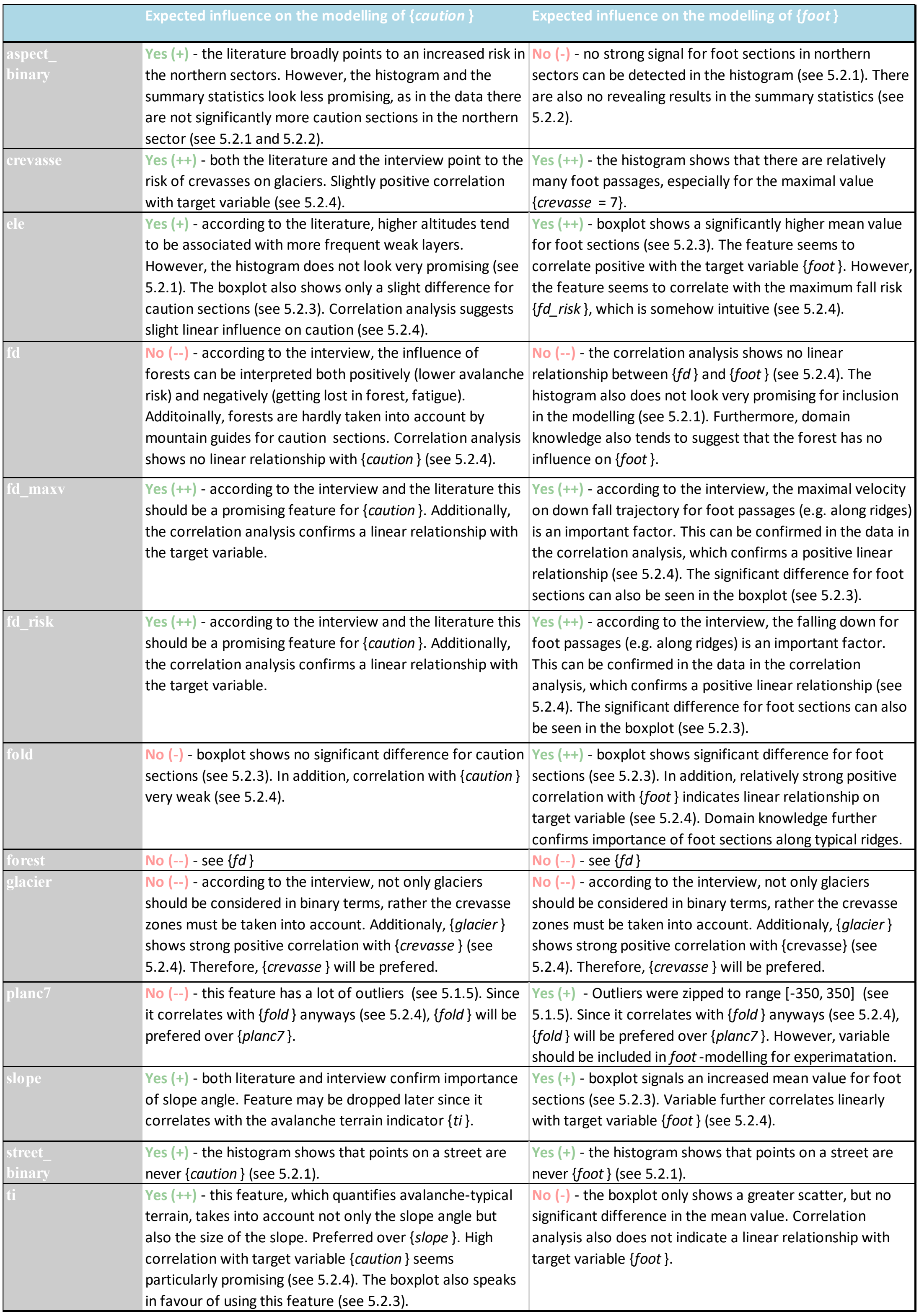
List of figures

List of tables

List of equations

Feature importance table

The following table belongs to chapter 5.4.3 and summarises the results of the explanatory data analysis and places them in the context of the literature knowledge. The 'most promising features' for *caution*- and *foot*-sections were also determined on this basis.

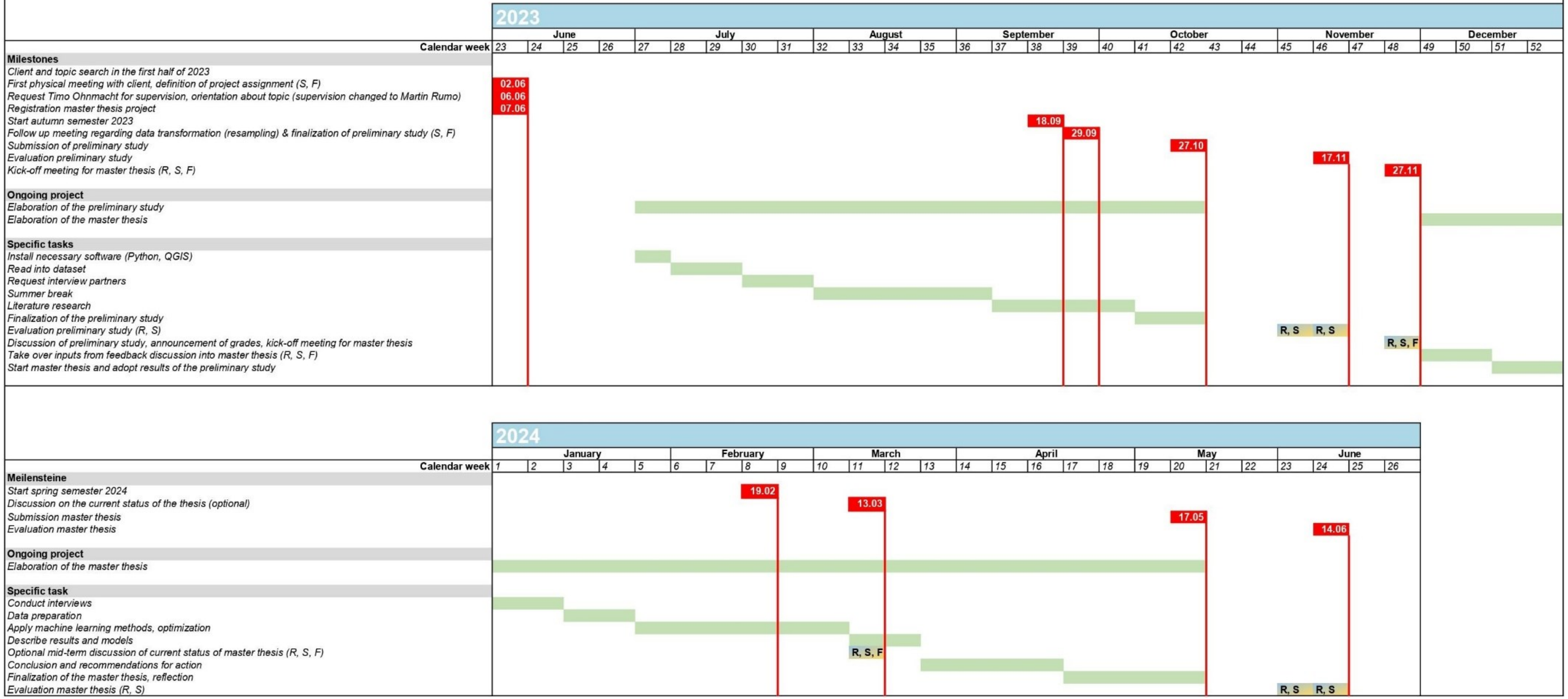


	Expected influence on the modelling of {*caution*}	Expected influence on the modelling of {*foot*}
aspect_ binary	Yes (+) - the literature broadly points to an increased risk in the northern sectors. However, the histogram and the summary statistics look less promising, as in the data there are not significantly more caution sections in the northern sector (see 5.2.1 and 5.2.2).	No (-) - no strong signal for foot sections in northern sectors can be detected in the histogram (see 5.2.1). There are also no revealing results in the summary statistics (see 5.2.2).
crevasse	Yes (++) - both the literature and the interview point to the risk of crevasses on glaciers. Slightly positive correlation with target variable (see 5.2.4).	Yes (++) - the histogram shows that there are relatively many foot passages, especially for the maximal value {*crevasse* = 7}.
ele	Yes (+) - according to the literature, higher altitudes tend to be associated with more frequent weak layers. However, the histogram does not look very promising (see 5.2.1). The boxplot also shows only a slight difference for caution sections (see 5.2.3). Correlation analysis suggests slight linear influence on caution (see 5.2.4).	Yes (++) - boxplot shows a significantly higher mean value for foot sections (see 5.2.3). The feature seems to correlate positive with the target variable {*foot*}. However, the feature seems to correlate with the maximum fall risk {*fd_risk*}, which is somehow intuitive (see 5.2.4).
fd	No (--) - according to the interview, the influence of forests can be interpreted both positively (lower avalanche risk) and negatively (getting lost in forest, fatigue). Additoinally, forests are hardly taken into account by mountain guides for caution sections. Correlation analysis shows no linear relationship with {*caution*} (see 5.2.4).	No (--) - the correlation analysis shows no linear relationship between {*fd*} and {*foot*} (see 5.2.4). The histogram also does not look very promising for inclusion in the modelling (see 5.2.1). Furthermore, domain knowledge also tends to suggest that the forest has no influence on {*foot*}.
fd_maxv	Yes (++) - according to the interview and the literature this should be a promising feature for {*caution*}. Additionally, the correlation analysis confirms a linear relationship with the target variable.	Yes (++) - according to the interview, the maximal velocity on down fall trajectory for foot passages (e.g. along ridges) is an important factor. This can be confirmed in the data in the correlation analysis, which confirms a positive linear relationship (see 5.2.4). The significant difference for foot sections can also be seen in the boxplot (see 5.2.3).
fd_risk	Yes (++) - according to the interview and the literature this should be a promising feature for {*caution*}. Additionally, the correlation analysis confirms a linear relationship with the target variable.	Yes (++) - according to the interview, the falling down for foot passages (e.g. along ridges) is an important factor. This can be confirmed in the data in the correlation analysis, which confirms a positive linear relationship (see 5.2.4). The significant difference for foot sections can also be seen in the boxplot (see 5.2.3).
fold	No (-) - boxplot shows no significant difference for caution sections (see 5.2.3). In addition, correlation with {*caution*} very weak (see 5.2.4).	Yes (++) - boxplot shows significant difference for foot sections (see 5.2.3). In addition, relatively strong positive correlation with {*foot*} indicates linear relationship on target variable (see 5.2.4). Domain knowledge further confirms importance of foot sections along typical ridges.
forest	No (--) - see {*fd*}	No (--) - see {*fd*}
glacier	No (--) - according to the interview, not only glaciers should be considered in binary terms, rather the crevasse zones must be taken into account. Additionaly, {*glacier*} shows strong positive correlation with {*crevasse*} (see 5.2.4). Therefore, {*crevasse*} will be prefered.	No (--) - according to the interview, not only glaciers should be considered in binary terms, rather the crevasse zones must be taken into account. Additionaly, {*glacier*} shows strong positive correlation with {crevasse} (see 5.2.4). Therefore, {*crevasse*} will be prefered.
planc7	No (--) - this feature has a lot of outliers (see 5.1.5). Since it correlates with {*fold*} anyways (see 5.2.4), {*fold*} will be prefered over {*planc7*}.	Yes (+) - Outliers were zipped to range [-350, 350] (see 5.1.5). Since it correlates with {*fold*} anyways (see 5.2.4), {*fold*} will be prefered over {*planc7*}. However, variable should be included in *foot*-modelling for experimatation.
slope	Yes (+) - both literature and interview confirm importance of slope angle. Feature may be dropped later since it correlates with the avalanche terrain indicator {*ti*}.	Yes (+) - boxplot signals an increased mean value for foot sections (see 5.2.3). Variable further correlates linearly with target variable {*foot*} (see 5.2.4).
street_ binary	Yes (+) - the histogram shows that points on a street are never {*caution*} (see 5.2.1).	Yes (+) - the histogram shows that points on a street are never {*foot*} (see 5.2.1).
ti	Yes (++) - this feature, which quantifies avalanche-typical terrain, takes into account not only the slope angle but also the size of the slope. Preferred over {*slope*}. High correlation with target variable {*caution*} seems particularly promising (see 5.2.4). The boxplot also speaks in favour of using this feature (see 5.2.3).	No (-) - the boxplot only shows a greater scatter, but no significant difference in the mean value. Correlation analysis also does not indicate a linear relationship with target variable {*foot*}.

Project plan

Project plan the master thesis

Topic:	***Artificial intelligence in ski touring*** *Prediction of "increased caution"- and "foot"-sections for ski touring routes with a machine learning approach*
Period	*01.06.2023 - 17.05.2024*
Participants	🟩 Claudio Furrer (F, Student) 🟦 Martin Rumo (R, Referat) 🟨 Günther Schmudlach (S, Skitourenguru.ch)



Milestones

Client and topic search in the first half of 2023

First physical meeting with client, definition of project assignment (S, F) 02.06

Request Timo Ohnmacht for supervision, orientation about topic (supervision changed to Martin Rumo) 06.06

Registration master thesis project 07.06

Start autumn semester 2023 18.09

Follow up meeting regarding data transformation (resampling) & finalization of preliminary study (S, F) 29.09

Submission of preliminary study 27.10

Evaluation preliminary study 17.11

Kick-off meeting for master thesis (R, S, F) 27.11

Ongoing project

Elaboration of the preliminary study

Elaboration of the master thesis

Specific tasks

Install necessary software (Python, QGIS)

Read into dataset

Request interview partners

Summer break

Literature research

Finalization of the preliminary study

Evaluation preliminary study (R, S) R, S R, S

Discussion of preliminary study, announcement of grades, kick-off meeting for master thesis R, S, F

Take over inputs from feedback discussion into master thesis (R, S, F)

Start master thesis and adopt results of the preliminary study

Meilensteine

Start spring semester 2024 19.02

Discussion on the current status of the thesis (optional) 13.03

Submission master thesis 17.05

Evaluation master thesis 14.06

Ongoing project

Elaboration of the master thesis

Specific task

Conduct interviews

Data preparation

Apply machine learning methods, optimization

Describe results and models

Optional mid-term discussion of current status of master thesis (R, S, F) R, S, F

Conclusion and recommendations for action

Finalization of the master thesis, reflection

Evaluation master thesis (R, S) R, S R, S

Interview transcript

Date: 25 February 2024

Location: Virtual meeting (teams)

Duration: 1 hour

Note: Interview was held in German and later translated into English.

(C) Interviewer: Claudio Furrer (Author of the master thesis)

(A) Interviewee: Andreas Eisenhut (Responsible for the SAC touring portal)

C: Hello Andreas. Thank you very much for your time. I will have to record the interview for use in my Master's thesis. Is that OK for you?

A: Yes, that's ok for me.

C: Then I'll share my screen now. You've written a little basic section on the cooperation between SAC and swisstopo. I've also looked at your PowerPoint, that you sent me, plus the Word. So it's effectively the case that until 2021 virtually all the tours come from the tour guides, i.e. the books, right? And you are now drawing in the routes more precisely? So 1:10'000?

A: To summarise briefly, in the past, until about two to four years ago, there were simply the two classic worlds: the Swiss Alpine Club (SAC) and swisstopo snow sports. Traditionally, the SAC simply produced the ski touring guides. Books that described the ski tours, and basically not directly marked on maps. On the other hand, swisstopo Schneesport, together with the Swiss Ski Association, has for many years, i.e. decades, put the most important routes, mainly from SAC literature, onto maps as a source. The swisstopo lead contains the classic ski touring maps, but the classic ski touring maps, and this was the case at swisstopo until two years ago, were standard in the past, and this is very important for you, they were rough at a scale of 50,000. The signatures, i.e. the distinction between foot terrain, increased caution or a solid line, were also done quite well. This is still partly the case today. So you can still see that pretty well in the data. I am now gradually improving the quality. About five years ago, the SAC started to create the SAC Touring Portal and had the requirement to adopt the geometries from swisstopo so that they would continue to fit together, but the editorial teams of the SAC Touring Portal and SAC Snowsports were separate until two years ago and the SAC Touring Portal therefore had outdated swisstopo geometries for a relatively long time and swisstopo has gradually started to build up a new quality. Until five years ago, the routes were simply drawn in an illustrator, without GIS in the background, and about four years ago they started to use QGIS directly up to the 25,000 and 10,000 meters as a basis. And I have now taken over the editorship from my predecessor with the requirement that updates should only be made at a common location in a common database. And from now on, the routes from swisstopo and the SAC tour portal, which also go to Skitourenguru, will fit together again. And only the tours from swisstopo are relevant for you. Günter has also only given you the routes that swisstopo publishes. There are also other routes in the SAC tour portal, but you are analysing the signatures from swisstopo. I think it is important in terms of the source that you rely exclusively on the data from swisstopo - and not from the SAC touring portal. The data from swisstopo Schneesport is based on the source of the SAC guide literature and close to the touring portal. It's a bit difficult, so it's not easy to understand.

C: Yes, I see…

A: Is a bit difficult…

C: (*laughs*)

A: (*laughs*) It's not easy, but...

C: Yes, I think you also drew a diagram on the PowerPoint showing how you imagine the process. I was able to roughly estimate it from this, but it's certainly easy for me to understand if you describe it again now.

A: I don't think this is relevant for you, but it is extremely important for you to understand that you have received data, namely the current swisstopo dataset, with the given differentiation of signatures. And many of these, or the distinction between where a signature such as increased caution begins, is on the old 50,000 accuracy for very many locations. And on the 50,000 map, whole sections are often assigned to a signature from branch to branch, and not based on location. So, for example, if you descend from a summit down a slope with avalanche danger or risk of falling, which really corresponds to increased caution in terms of content, and then this route continues over flatter terrain, down through the forest, to the next junction where another route joins. On the old 50'000 meters, the increased caution signature often continued right up to the junction.

C: Oh, okay. So rather generously drawn...

A: Very generously drawn.

C: So not really just in the area where it would really only apply. But just the whole section?

A: Exactly. On the one hand, the route location wasn't accurate enough to really show the area where you were actually going. That was a big problem, especially in the foot terrain, with the dotted lines. But a lot of things have already been adjusted there. The route position, if you are interested I can send you other data where you can see which data already has the new position quality and which still has the old indicative 50,000s. So you have a fuzziness in the data on the one hand because the position quality is not yet uniformly new, but in about 20% to 30% of the routes you still have the old indicative 50,000 position quality, and on the other hand you have the demarcation of the signatures, which is still very rough in the vast majority of cases.

C: Okay... Which is of course suboptimal for me...

A: That’s effectively suboptimal for you...

C: Think about it...

A: So you can pragmatically say that the foot terrain is much easier to get out automatically. The increased caution, on the other hand, will be extremely difficult. I can give you two more pointers here, which are decisive as to whether the work yields a good result or not. Maybe I’ll get to them later, I may not have described them in enough detail. The scales have just fallen from my eyes. But I think it’s best to go through all your questions first.

C: Yes, okay. Let’s do it that way.

A: Or maybe I can tell you in advance. I think it’s extremely important (*pause*) I’ll share my screen briefly anyway. So here you can see the 50,000 map as printed by swisstopo. Now that’s a super good example of increased caution. Can you see my mouse?

C: Yes

A: From Piz Tamül here is this descent variant. Of course, it’s absolute rubbish to draw it as a dashed line at the bottom of the descent. Because if anything, the avalanche slope up here is simply dashed, and the rest is not. And, in the old quality or quasi for the printed product of the ski tour map, which is becoming less and less important, as swisstopo almost no longer sells these physical maps, they are gradually working with digital first. There will also be a new digital product based on this data and it will be imperative that the dashed signature is only shown in those areas where there is really increased caution.

C: Yes, I see. That really is a problem. Hmm...

A: That’s just a...

C: (*Interrupts*) Just briefly, at this point, what’s the reason that this was historically so broad? Because otherwise it simply wouldn’t be visible on this scale?

A: (*Interrupts*) No...

C: Or you can’t draw it so precisely at 1:50,000?

A: The reason is that this data was clearly produced by a network of authors who were paid. Some of the geologists also have this, like the grain of your authoring career was that you used to be responsible for a map sheet at swisstopo. The authors had the power of interpretation like kings.

C: Laughs

A: And that of course led to extreme differences and here in this section, the focus of this tour is simply on the descent with increased caution compared to the ascent or the entire valley. If you only look at the valley, this is perhaps a tour that has an increased caution on the descent. But if you look at the whole of Switzerland and you want to do a calibration, that's a joke, of course. So you have a huge bias. And the authors basically have, well, you have an editor-in-chief, so to speak, which was my predecessor, who collected the information from the local authors, so to speak. And that's actually different for each individual, with different eyes, it's just put together. So if a route description says avalanche danger or more difficult to ski down, then he often did something like this (generous signature) as an interpretation.

C: Okay, I see.

A: Exactly, and I have... yes. Exactly, so to really go step by step, this example here is perhaps a bit complicated, but... I have to have a quick look... no... sorry... The sketch here, which I also sent you...

C: (*Interrupts*) Yes, exactly, I've seen it too

A: ... it abstracts something... which is also quite important for you to understand. The ski touring routes are in blue, do you see that?

C: Yes.

A: With the three different signatures. Either solid, dashed or dotted.

C: Exactly.

A: And then there are areas where the blue route goes along roads and paths, i.e. off-road. So either a road up through the forest, or a forest path up through the forest, or if it's a track from a ski slope and there's still a road and the route still goes straight up. Wherever the markings are yellow, as you can see

C: (*Interrupts*) Yes, exactly, I see.

A: Everywhere there, the route runs along swisstopo TLM. So there the route runs exactly along a path. And basically you can say that as long as a route runs on a TLM, i.e. on a path...

C: (*Nods*)

A: ...the route is generally not increased caution. There are very, very few exceptions, and I can or you probably have to separate them clearly. And, for example, you have a criterion where you look, you look if it runs through the forest or through bushes, does it tend to be increased caution or not. But from my point of view, you have to make a clear distinction, you only look at the route areas that don't run along a path.

C: (*Nods*) Yes.

A: Or in principle you would have to... once you look, you calculate the original routes, and once you calculate only those route areas that do not run along a TLM or are not yellow. And I can imagine that you will arrive at relatively different results. And what I wanted to say is that if you want to know whether forestation is relevant, then you should only look at those routes in the forest where the route does not run along a road.

C: (*Nods*)

A: Because if the route runs on a road in the forest, then the forest is not relevant. But if you put the bare forest on top of your data comparison without considering whether there is a road running through it or not, then it will definitely backfire.

C: Yes, that makes sense, I understand. So let me get this straight: I basically have to divide the forest into a forest without a hiking trail or forest road...

A: (*Interrupts*) Yes!

C: ... and the forest really as a forest where I walk through pathless terrain. I have to take that into account. That's what you mean, isn't it?

A: Exactly. You can formulate it from the forest, or you can formulate it from the route. You can also say that you only look at the sections of the route that are completely in open terrain. I've developed a model for this, I have a calculation model that shows you the individual TLM sections along the route.

C: Okay, that sounds good.

A: So I can simply provide you with the geodata, including the threshold that I used for this. Then you can simply click on the points that you have created along the route and select which ones are along the TLM. So which ones are yellow and which ones are not yellow.

C: So then I have a zero or a one, so to speak, whether it's on a road or not.

A: Exactly, exactly. So I can do that according to my criteria, which I differentiate, I just had to make this distinction because swisstopo sets relatively embarrassing criteria when routes are digitised (laughs).

C: (*laughs*)

A: Then I simply smoothed out these routes, but only where it wasn't yellow. Because if you were to smooth the hiking trail too, you want it to run exactly on the TLM.

C: Yes, I can see that.

A: Yes. That's why it's such a compromise with them. But you can, I'm convinced, or you have to, from my point of view, take that into account one-to-one.

C: Yes, that probably makes sense...

A: Because otherwise all the labour of love from your exciting work... it's just a gap that's simply there.

C: Absolutely, I can see that. That's a very good punt. And I think you've also shown the reverse case further down, as I saw. So if a path leads along a pass road...

A: In the PowerPoint, or where?

C: Yes, now I'm not sure where this illustration was.

A: But otherwise I'll send you this slide, I haven't sent it to you yet. I only just found it earlier. I also didn't know whether you still had the old product. It's important to understand what the data is and where the limits are.

C: Yes. Then let's go through the questions systematically. Exactly, you've actually described the sources relatively precisely. Ehm.

A: Exactly, so with the tour sources you can also be pragmatic and say that the source is the SAC guide literature. Plus the SAC tour portal.

C: But just to clarify again. In my case, it's simply the dataset from swisstopo...

A: (*Interrupts*) Exactly.

C: ... 1:50'000 with all the weaknesses you mentioned earlier (laughs)

A: Yes, exactly.

C: Then exactly, the main sources, so that's actually a similar question, were often mountain guides and experts in the past... and what did you mean by that, the calibration at swisstopo, that was then critically looked at again by swisstopo and checked whether something wasn't very crooked in the landscape, and so swisstopo made a few small adjustments again?

A: Exactly, the calibration is meant in such a way that only one person drew it at a time. And that is relatively relevant. In principle, you can say that the calibration, as the draughtsman, received the information from the authors, for example via screenshots or descriptions, and then the draughtsman took over this as an expert. And up until four years ago, only the 1:50'000 maps were processed in Adobe Illustrator and the illustrator only worked with curves, so to speak, where he could simply display the lines. So the line was then really simply indicative and often detailed in the reeds. This is not irrelevant for you, so it can generate relatively large errors. But in the meantime, about 80% of the routes are of good quality within Switzerland. I don't know, do you limit your work to Switzerland or... I wanted to say that too. So the perimeter that you calculate, there is... ehm... the location quality is much better within Switzerland than in neighbouring countries.

C: Okay... So yes, I'm already working with the data from Switzerland.

A: Well then, it probably makes pragmatic sense for you to restrict the calculation perimeter to Swiss territory.

C: Yes, exactly. That's actually the idea.

A: Exactly... You don't really need to look at the data from abroad, it's often so outdated... The whole Mont Blanc massif in particular is still in there, and no more has been done there since we started working with QGIS and the terrain has changed massively due to glacier retreat. So there, yes... Exactly.

C: Then. Point four, who is mainly involved in the classification process, you actually said that too. Mainly the mountain guides and the authors.

A: (*Nods*) Exactly.

C: And then the calibration afterwards. And what did you mean by the second point, will there be more feedback from the community later on?

A: We get all the relevant feedback from tourers on the road. In the swisstopo app, for example, you can see when a route is closed. Tourers can give feedback directly via the app. This applies not only to spontaneous events (e.g. rockfall), but also to well-founded feedback or criticism of the individual signatures. For example, where there is still a mismatch between swisstopo and the SAC tour portal. Compared to the past with the authors, the community has become very important. The SAC also has 111 sections that are very active, each SAC section has summer and winter tour leaders, and they give... they know which tour leaders are the most active, and the winter tour leader is instructed to pass on the feedback to us. With appropriate photos etc.

C: All right... and this feedback didn't exist before, did it?

A: No, exactly.

C: The mountain guide simply drew his map and then it remained relatively static.

A: Correct. And what you never see from the data due to increased caution, for example, are reports where there is a rockfall area due to permafrost. I will then record the area mentioned internally as an area and in future I imagine that most of the increased caution passages can be classified using your approach. And then I can selectively mark certain passages manually, such as rockfall hazards. But these are thoughts that I've just had.

C: Okay, I see. But exactly, rockfall areas are areas that you have to draw in, which cannot be extracted directly from a swisstopo layer?

A: Exactly.

C: Then to the sixth question, whether historical accident and avalanche events are also included in the classification of the sections in the current assessment of increased caution?

A: (*Interrupts*) That's a very cool point! I haven't thought about this enough yet and it's also exciting in my dialogue with Günter Schmudlach, I've been in close contact with him since 2015. Yes, I'm less active at this level, but I can imagine that this could be very, very important for your work.

C: Yes, in theory it could be a feature of the work. On the other hand, it's probably not such a good feature because of the dependency on accessing a slope. Just because an avalanche has not yet been recorded on a slope does not mean that it should be included in the modelling with less probability for increased caution. My main question was whether the authors have included this feature in the classification of individual passages in the past. I doubt whether it would really make sense to include the feature in the work.

A: Yes, exactly, I am not aware from the tour literature that historical accidents had an influence on the classification.

C: Okay. Then to the next question, what do you think are possible criteria for increased caution? You gave me the following criteria in advance: Pronounced avalanche and fall terrain, dangerous crevasse zones on glaciers, unavoidable crossings of reservoirs, unstable terrain such as risk of rockfall or falling rocks, heavily overgrown or unclear terrain. Regarding the crevasse zones on glaciers that you listed, this is presumably similar to the rockfall terrain, where you can't simply extract from a map or layer on swisstopo?

A: Yes, fortunately you can.

C: So there is a layer for this?

A: No, there is no ready-made layer for this. But I can show you briefly if you go to map.admin, for example, you can simply use the swissALTI3D Relief and there you have a resolution of half a metre. Above 2'000 meters, the detection method probably doesn't quite do it justice, below 2'000 meters you have LIDAR flights. Above that you have a different method, which I don't know exactly. But you can see the crevasse zones extremely well here at a glance. And you have these differences in the digital terrain model. On the swissTLM, you have the glacier surface itself as a polygon...

C: (*Interrupts*) Yes.

A: ... you can use this to cut out the elevation model of the glacier in this area. And there it is then simply a suitable, simple neighbourhood analysis in the GIS raster. You say, for example, if you calculate with a resolution of 5 meters or 10 meters, each raster cell within a radius of 10 or 50 meters looks at how much the exposure or the altitude or how large the variance is. So you can use simple tricks to derive a polygon from raster data, where you then have pronounced crevasses within the glacier.

C: Very cool! And can you estimate how static these crevasse zones are? I mean, the glaciers are shrinking and ...

A: (*Interrupts*) Yes, they are shrinking. But the data from swisstopo is available every three to six years and it's... so a crevasse zone... as long as there are still glaciers here, the crevasse zone remains there. So the glacier is disappearing, but the crevasses remain more or less in the same place. So the crevasse zones themselves don't really move with the glacier. So you can use simple tricks to automatically locate the crevasse zones very well.

C: Very good, okay. I wouldn't have thought that you could do that so well.

A: Yes, you can... And I could send it to you otherwise. I've already done this for Günter, albeit somewhat pragmatically and not quite up to date in terms of the data status, but I could send it to you otherwise.

C: Hey, that would be great! Because I'm not that much of a GIS expert myself, so I'd be very happy about that.

A: Yes of course, I'll put it in my diary.

C: So the dangerous part is not the glacier itself, but the crevasse zone in particular?

A: Absolutely! That's very important! It can't be the case that when you're skiing over a glacier, everything is heightened caution or not. So I will provide you with the grid that contains the areas that are dangerous in terms of crevasses. I'll provide you with a grid where you can see a progression of danger.

C: And just a quick reminder. The crevasse zones tend to be where the glacier flows over a ridge, and that's where it virtually breaks apart? Because there are tensile forces?

A: Yes.

C: Okay. Exciting...

A: I'll see what I can do. I can just give you the current layer, but it's relatively pragmatic.

C: And with regard to reservoirs: Why only reservoirs or rather reservoirs - and not simply lakes in a more general sense?

A: Yes, of course, general lakes too. But if you only look at lakes now, you can say that frozen lakes tend not to be increased caution. Lakes that are only sometimes frozen over are increased caution. In the case of lakes, regardless of reservoirs, if it is simply a lake, I have said that lakes above a certain altitude are probably or definitely frozen over. The smaller and higher up the lake, the more likely it is to be frozen over. Then it is no longer relevant for the signature. Around 90% of this is already covered by the current signatures. Tours should only cross a lake if there is no other option. For example, the Grimsel reservoir if you are travelling towards Lauteraar. In the case of reservoirs, we have often received feedback, including from reservoir operators, that the routes should be marked with a minimum of caution. This is because it becomes problematic when certain reservoirs release a lot of water in a very short time.

C: So it’s another level more dangerous than just a standard lake...

A: Yes. But from a gut feeling, you can probably disregard this at work. It’s never as important as, for example, crevasses, fall terrain or avalanche terrain. Never.

C: And then again about the risk of falling rocks. This is really not something that I can read from swisstopo, but something that has to be drawn in manually, as you do? So it will probably be something that I won’t be able to take into account in the work. But I think that’s okay.

A: Yes. I think so too.

C: And that’s right, now it occurs to me. That’s the special case I was talking about earlier. Here with the mountain pass road, which is classified as an increased level of caution. So if you’re on a road, but it can still be an increased level of caution.

A: And just now. The suggestion only occurred to me after I had already sent you the documents. Namely, you should consistently calculate the entire route network and only the sections that are not along roads. Of course, this is just a single case, but it’s a relatively long route. Or, for example, the criterion in the forest or not, where you have an extremely large number of routes that are on a road in the forest. And if you don’t differentiate that, then you don’t have to do the work. And here, in this case, it’s simply the case that this road is closed in winter and leads through avalanche terrain and it can therefore be very difficult to go back there in winter.

C: Okay, yes. I have to somehow find a way to manage that.

A: And that’s why he has only classified the rest of the large section here as an increased caution from the 50,000 product. And that’s why I think it makes sense to do the maths twice. Once with roads and once without roads.

C: I see, I see... So your direction of travel would be to exclude those sections that obviously run along a road but are an increased level of caution?

A: No, my approach would be to consistently exclude all ski tour sections along roads. Therefore, calculate all sections and then calculate all sections without the road sections. And then see where the differences are.

C: Okay... I'll have to get my head round this and also have a look with Günter. But I'll keep that in mind.

A: Yes, I'm curious. And Günter and I have different views on how far machine learning can go in this area. I have little experience with machine learning myself, but I am sceptical if the data is not processed in a reflective way. That's why this input is so important to me.

C: Yes, they are really very helpful. Because it's true, garbage in... garbage out...

A: Exactly...

C: I agree with your comments on steepness, exposure, altitude, forestation, risk of falling, slope size, curvature, crest and historical avalanche events. Only in the case of forestation. Why does it tend to become more dangerous with increasing forest cover? Because I would have said that the danger decreases because the risk of triggering an avalanche also decreases.

A: From the point of view of avalanche danger, I am absolutely on your side. If you are in dense forest, it is of course much less dangerous. But from a holistic point of view, increased caution...

C: (*Interrupts*) Where it's generally about the accident?

A: ... where it's generally about accidents, or how well you can manage your tour, if you're on a summit, you're doing a traverse and descending into an unknown valley in Ticino, for example, where you have to descend several hundred meters down a forest without a path, then there's a considerable risk here that you won't find the path...

C: (*Interrupts*) ... or suddenly find yourself in front of a rock face...

A: (*Interrupts*) ... yes, exactly!

C: Okay, I see.

A: Or the risk of injury, greater effort or disorientation also play a role. Depending on the snow conditions, it can become very critical in the steep forest.

C: Okay. But is it already labelled that way? So if I go to swisstopo and the route leads through extremely dense forest, without a path, is that increased caution today, or not?

A: No. If at all, then only very inconsistently from certain authors. If you take this into account, it would be interesting to only look at those sections of the route that are off the road.

C: Yes.

A: There it would be decisive whether you find something, so to speak, the steeper the longer, in the forest, the more dangerous or problematic...

C: Absolutely.

A: And if you don’t find anything, I wouldn’t be surprised.

C: Okay. And just in terms of walking a route, I don’t think it makes sense to include that as a feature. Because this is not a static feature. I was more interested in whether the authors had taken this into account in any way.

A: Exactly.

C: Do certain criteria have a particularly high influence on increased caution? You mentioned slope inclination and fall risk to me in advance Skitourenguru. And the reservoirs, if they have an influence, then only a small one.

A: Exactly. If I were to make a pragmatic classification, i.e. a classification without machine learning, then I would use the first three criteria. So slope gradient, fall risk from Skitourenguru and crevasse zones, I would define a threshold value for these three zones. The advantage of this would be that you could say in black and white in two sentences what the signature is.

C: Okay. Yes. It’s then really easy to justify why it’s like that. And not easy, we’ve done the maths and we don’t know why it’s like that, but it is.

A: Exactly. That’s why your work is so important. That we can increase acceptance there. Because I’m convinced of the potential that such work has. It would still be exciting if, independently of your work, we could create the mask as an expert opinion and then compare what can be achieved with your results. I would like your work to give me an indication that I can set the threshold value.

C: I see.

A: And if your work is so good, it may be that you don’t need to set the threshold at all. So if it turns out that these three features are the most important anyway, then I won’t have to do anything else. Then you can simply tell the users that the terrain indicator, the risk of falling and the crevasse zones are the most important. I also find it interesting whether the bare height above sea level is a relevant criterion. The higher the more dangerous.

C: Where I think Günter is very much of the opinion that this will have an influence, right?

A: Yes. And this is where we disagree. But I understand his point of view as well, but... yes.

C: But you know, if I do the maths, then the height is then combined with other features such as is it high and a glacier and not a road. And then I can also see to some extent that it can have a greater influence, or not.

A: Yes...

C: But that will be exciting, these are all things that I will certainly look at.

A: Cool.

C: Then just at the very end, about the foot passages. You've already described the criteria for this mainly as Günter's risk of falling and the steepness, because you have to walk up from a certain steepness.

A: Yes, exactly. I had another thought about it, but I can't think of it now.

C: Okay, otherwise you can just write to me if it occurs to you.

A: Yes. And as far as I know, Günter just took the risk of falling almost one-to-one.

C: Okay.

A: And according to the SAC, the foot passages are precisely described verbally and only marked for the relevant part. Whereas the more cautious passages are often simply marked from branch to branch. As mentioned at the beginning, most of the new markings for the increased caution are still in the old state and I do 99% of the new implementation automatically. Either with my approach or with your approach.

C: Okay.

A: And that's not my priority. It won't be done for another year or two. So we'll wait for your work first before we do anything else.

C: Okay, yes. Then the pressure is on (*laughs*).

A: No! There's not much pressure, you just do the best you can and that's valuable. But we're waiting for your work and that's important for you to know.

C: Hey okay, that's great. I think I've gained a lot of impressions about the data and that's important. That I know how they came about and with the knowledge you've given me today, I think the basis is hopefully good enough and that the right things are clear and I think that helped a lot. So thank you very much for your time.

A: With pleasure. If you have any questions, you can always get in touch. And I find it very exciting that something is going in this direction. And then at the latest at the presentation or opportunity to see these results. Or, if you are in contact with Günter, he is extremely inspiring and yes, then you decide for yourself whether you have questions for me during the work or not.

C: Yes, great. I would just email you if there are any more questions.

A: Good luck, have fun!

399 **C:** Thank you! Merci, merci. And have a good weekend, take care!

400 **A:** Thank you, you too.

Declaration of originality

The undersigned hereby declares that he or she

- wrote the work in question independently and without the help of any third party,
- has provided all the sources and cited the literature used,
- will protect the confidentiality interests of the client and respect the copyright regulations of Lucerne University of Applied Sciences and Arts.

……………………………………

Claudio Furrer

Lucerne University of Applied Sciences and Arts
Master of Science in Applied Information and Data Science (MScIDS)
Lucerne, May 10, 2024

Declaration of the use of Generative AI

The following declaration of the use of Generative AI must be filled in and included at the end of the master thesis. It is not relevant for the grading.

For the master thesis, I have used Generative AI to:
(please select all that apply)

☐ brainstorm new ideas

☐ generate new research hypotheses

☐ do scientific research

☐ summarize other research

☐ set up the research project / project design

☒ code (primarily debugging)

☐ write the manuscript

☐ edit text

☒ translate text

……………………………………

Claudio Furrer

Lucerne University of Applied Sciences and Arts
Master of Science in Applied Information and Data Science (MScIDS)
Lucerne, May 10, 2024