universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Application of Machine Learning to Weather-Triggered Hazards and Damages in Alpine Territory

verfasst von / submitted by

Georg A. Seyerl, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2020 / Vienna, 2020

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 614
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Meteorologie
Betreut von / Supervisor:	Priv.-Doz. Dr. Christoph Matulla

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Vienna, _____ _____

Date Signature

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

Wien, am _____ _____

Datum Unterschrift

Abstract

The application of tree-based machine learning methods in the field of hazard event prediction is challenging. The imbalanced binary classification task requires suitable countermeasures in order to enhance the models predictability of the minority class (hazard event). We estimate the potential of different techniques [Branco, Torgo, and Ribeiro, 2015], such as pre-processing, special purpose learning methods and post-processing, by applying them on random forests and gradient boosted trees under well known synthetic conditions. The special purpose learning methods outperform the pre- and post-processing approaches, whereto k-nearest neighbor difficulty analyses and various performance metrics for ten synthetic data sets are evaluated. Following up on Enigl et al., 2019 we analyse the yearly sum of hazard events for seven categories and expanded the Austrian hazard event space by "time independent" features derived from terrain, soil, vegetation and geological data. The data is further filtered for slide events on which we perform a difficulty analysis for different geological domains. Thus, the analysis reveals the degree of imbalance and difficulty, at which comparable synthetic data sets tend to be in limbo of viability. Nevertheless, an extensive modeling approach for the hazard category slide is performed using scaled random forests (SXRF) and scaled gradient boosted trees (SXGBT), both implemented in the XGBoost framework [T. Chen and Guestrin, 2016]. Whereat, the term "scaled" refers to the fact that weights are balanced in the cost function for minority and majority instances in relation to their global imbalance ratio. SXRF is outperformed by SXGBT and 5-fold cross validation scores indicate the robustness of the model with consistent sensitivity scores of approximately 0.75 and areas under the receiver operator curve of approximately 0.80. The most important numeric features are "slope" (considering correlated features), "minimum distance to street", "minimum distance to nappe boundary" and the "topsoil physical property bulk density". The most important binary one-hot encoded categorical features are the landcover class "Woody" and the geological domains "Austroalpine Units" and "Siliclastic Rocks". Non-linear tree-based machine learning methods may further improve data-driven models for susceptibility mapping of spatially non-persistent hazard events. Nevertheless, they depend heavily on the quality of the underlying feature and hazard event data.

Zusammenfassung

Die Anwendung von entscheidungsbaumbasierten Machine Learning Methoden im Umfeld der Extremereignisvorhersage ist herausfordernd. Imbalanzierte binäre Klassifikationsaufgaben bedingen einer passenden Gegenmaßnahme um die Vorhersagequalität des Modells in Bezug auf die Minderheitsklasse (Extremereignis) zu verbessern. Wir evaluieren das Potential unterschiedlicher Techniken [Branco, Torgo und Ribeiro, 2015] wie preprocessing, special purpose learning methods und post-processing durch ihre Anwendung auf random forests und gradient boosted trees. K-nearest neighbor Schwierigkeitsanalysen und unterschiedliche Validierungsmetriken wurden für zehn synthetische Datensätze ausgewertet und liefern die besten Ergebnisse für special purpose learning Methoden. Bezugnehmend auf Enigl u. a., 2019 analysieren wir die jährlichen Summen von Extremereignissen für sieben Kategorien und erweitern den Schadensdatensatz mit 'zeitlich unabhängigen' Prädiktoren aus Terrain-, Vegetations-, Boden- und Geologiedaten. Der gesamte Datensatz wird auf Rutschungsereignisse reduziert und eine Schwierigkeitsanalyse für unterschiedliche geologische Domainen wird durchgeführt. Die Analyse zeigt den hohen Grad des Ungleichgewichtes und der Schwierigkeit welche, bei vergleichbaren synthetische Datensätzen, die Grenze valider Auswertemöglichkeiten darstellen. Ein umfangreicher Suszeptibilitätsmodellierungsversuch für die Ereigniskategorie Hangrutschung wird mit skalierten random forest (SXRF) und skalierten gradient boosted tree (SXGBT) Modellen aus dem XGBoost Framework [T. Chen und Guestrin, 2016] durchgeführt. Hierbei bezeichnet der Term 'skaliert' die balancierte Gewichtung der Minoritäts- und Majoritätsklasse in der Kostenfunktion unter Berücksichtigung ihres globalen Ungleichgewichtes. SXRF liefert schlechtere Ergebnisse als SXGBT. Die 5-fach Kreuzvalidierung von SXGBT mit konsistente Sensitivitäten von $\sim$ 0.75 und den Flächen unter den Grenzwertoptimierungskurven von $\sim$ 0.8 deuten auf die Robustheit des Modells und der gewählten Prädiktoren hin. Die wichtigsten numerischen Prädiktoren sind 'Hangneigung' (unter Berücksichtigung korrelierter Prädiktoren), 'Distanz zur nächstgelegenen Straße', 'Distanz zur nächstgelegenen geologischen Grenzschicht' und der Oberflächenparameter 'Bodendichte'. Die wichtigsten binären one-hot kodierten kategorischen Prädiktoren sind die Landbedeckungsklasse 'Wald' und die geologischen Domänen 'Austroalpine Einheit' und 'Siliciklastika'.

Contents

List of Figures

1 Introduction

Recently colleagues from the Institute of Meteorology and Geophysics at the University of Vienna have successfully contributed to the development of forward-looking strategies in dealing with climate-change-induced extreme events and damages, which are reportedly on the increase in most parts of the world [EASAC, 2018]. They have established the so called 'event space' (currently the largest Austrian database on landslides, floodings and heat spells), linkages between weather-sequences and threat events, ensembles of downscaled projections pertaining to different pathways of mankind and a procedure supporting decision-makers in sustainable planning [Enigl et al., 2019]. These and in fact most climatological studies rely on data analysis and transformation methods. Therefore, in this study we analyze the applicability of Machine Learning (ML) to this field of research, whereby the above-mentioned studies provide a test bed for our undertaking.

We focus mainly on the past. Hence, on the 'event space' made up of more than 20.000 landslides, flows and flooding events in different categories observed throughout seven decades. From there, the data is further filtered for landslide events and a k-nearest neighbor difficulty analysis is performed on multiple geological domains. After those analyses, we perform an extensive slide susceptibility mapping approach including geological domains as one-hot-encoded features. The best performing model is then analyzed in terms of it's feature importance. In order to tackle those tasks, we merge the damage data with 'time independent' geology, lithology, soil, vegetation and auspicious terrain properties obtained and derived from the Austrian Geological Institute (GBA), a digital elevation model and the European soil database.

However, when dealing with 'event space' data several problems occur. To begin with, the nature of extreme events results in highly skewed/unbalanced data sets. Put differently, we compare sparse occurrences of extreme events to high numbers of non-events over the observed regional and temporal period. The complexity of a classification/prediction task increases further as trigger mechanisms are highly variable (e.g. weather, snow melt, freezing/thawing, terrain, soil) for different types of hazardous events. Another

problem for machine learning algorithms occurs due to missing observations in the event space, especially during the early period when communication technology was not naturally at a fingertip.

In order to estimate some of the above mentioned problems, we want to elaborate the applicability of ML algorithms under specific conditions. Confidence in algorithms and methods is generally enhanced in case their findings gained under known conditions (e.g. synthetic data) match expectations.

1.1 Literature

To examine the applicability of ML approaches a literature review was carried out. Publications dealing with the problem of imbalanced domains throughout the last decade were favored and analyzed for data sampling and modelling techniques. Additionally, literature dealing with hazard susceptibility mapping and prediction was reviewed for the damage event categories slides and floods.

1.1.1 Machine Learning

There is a broad selection of introductory literature on ML algorithms but two outstanding books are tackling the task from different perspectives, the practical introduction by Gareth et al., 2019 and the more theoretical introduction by Hastie, Robert, and J. Friedman, 2017. Those introductions offer a basic knowledge on different types of supervised and unsupervised learning methods, which are going to be described in more detail within chapter 3.

An in depth review of different approaches to tackle the problems of imbalanced domains for classification and regression tasks was conducted by Branco, Torgo, and Ribeiro, 2015. They propose a general definition for imbalanced data distributions. Therefore, they define a user preference bias where a subset of the target variable assigns more importance in the performance of approximation than the rest of the target variable. If there is

a high discrepancy in the number of examples for different subsets and the model/evaluation is insensitive to the user preference bias, it is most likely a problem of imbalanced data sets. Besides the necessity for definitions of different performance metrics, Branco, Torgo, and Ribeiro, 2015 constitute four modelling strategies categories to address the problem: Data Pre-processing, Special-purpose Learning Methods, Prediction Post-processing and Hybrid Methods.

The data pre-processing approach re-sampling [Estabrooks, Jo, and Japkowicz, 2004; Batuwita and Palade, 2010] is performed through changes in the data distribution in order to reflect the user preference bias before the actual model is trained. Although it is difficult to find the optimal new data distribution, re-sampling can be applied to any standard learning tool. The most common approaches are random under- and over-sampling, where data is either removed from the majority class or added to the minority class. Over-sampling might be applied by duplicating existing minority data or by adding synthetically generated data to the minority class [Chawla et al., 2002]. Under-sampling methods can be combined with ensemble techniques such as boosting [Schapire, 2003] and bagging [Breiman, 1996]. Weiss and Provost, 2003 showed that perfectly balanced classes are not an ideal sampling strategy, consequently a different approach named roughly balanced bagging was developed by Hido, Kashima, and Takahashi, 2009 and extended for multi-class problems by Lango and Stefanowski, 2018.

Special-purpose learning methods modify the existing machine learning algorithms in order to account for the user preference bias. This approach is more delicate than re-sampling and the implementation depends highly on the classification method used in the first place. For tree-based methods used throughout this work, the idea is to increase the loss for wrong predictions of the minority class [C. Chen, Liaw, and Breiman, 2004]. Consequently, the model is forced to put more weight on the classification of the minority class.

Another method to account for imbalanced data after the actual classification task is prediction post-processing. Although there are more possibilities, we only want to mention the threshold method. If a classifier provides the possibility to predict probabilities of class memberships, one can tune the

classification threshold in order to generate new classifiers based on the probability and the chosen threshold.

In this study, we are going to use a special-purpose learning method and prediction post-processing in combination with random forests [Breiman, 2001] and gradient boosted trees [J. H. Friedman, 2001], which are described in chapter 3 and implemented in the C++ framework xgboost [T. Chen and Guestrin, 2016].

1.1.2 Slides

Landslide susceptibility is defined as the probability of a slope failure calculated from a limited set of pre-defined features [Guzzetti et al., 2005]. The frequencies and the magnitudes of slide events remain undetermined.

Carrara et al., 1991 used geographic information system (GIS) techniques in combination with discriminant analysis to produce landslide susceptibility maps and proposed it as a cost-effective method with drawbacks for regions where physical based modelling is not possible due to missing geotechnical data. Other traditional methods in literature include weights of evidence [Dahal et al., 2008; Regmi, Giardino, and Vitek, 2010] and logistic regression [Ohlmacher and J. C. Davis, 2003, Ayalew and Yamagishi, 2005].

With the rise of machine learning methods, they were also applied in the field of land slide susceptibility mapping. Marjanović et al., 2011 compared the traditional method logistic regression with the machine learning methods Support Vector Machines (SVM) and Decision Trees for a mountainous region in north Serbia. The results indicate that SVM scores best in all evaluation metrics.

Micheletti et al., 2014 used adaptive support vector machines, random forests and AdaBoost in order to evaluate the feature importance of landslide susceptibility for the Swiss canton Vaud. They analyzed the applicability of their methods for two different types of landslides. The accuracy of the model and therefore also the quality of the susceptibility map were high for shallow landslides. For deep-seated landslides the model performance was low due to the lack of characterizing features.

Another tackling problem mentioned above is the incompleteness of the hazard event inventory, which is used to train and verify the model. Steger et al., 2017 analyzes the effect of incomplete data catalogues on synthetic and real data in Lower Austria using binary logistic regression (generalized linear model). The study shows that good evaluation metrics for a model do not necessarily indicate realistic susceptibility maps. They suggest the use of non-spatially assessed and spatially assessed cross validation in order to verify the consistency of the model.

Petschko et al., 2014 also conducted a case study for Lower Austria. In contrast to generalized linear models (GLM), the generalized additive model (GAM) used in this study replaces the linear function of each co-variate with an empirical function. The study shows high variability in model performance for different spatial domains. Nevertheless, the following features are important in all modelling domains in descending order of importance: slope, catchment height, euclidian distance to nappe boundaries, topographic wetness index, convergence index, curvature. In contrast to other studies Petschko et al., 2014, also provides a relative uncertainty index of the predicted susceptibility map.

Comparing traditional and machine learning methods for different regions in Lower Austria, Goetz et al., 2015 used GLM, GAM, weights of evidence, support vector machine, random forest classification and bootstrap aggregated trees. The latter two seem to perform better, although the modelling performance does not differ significantly for all of them. The feature importance varies based on different geomorphic conditions. Nevertheless, the analysis shows that slope angle, surface roughness and plan curvature are belong the most important features throughout different geo-morphological model domains.

1.1.3 Floods

As climate changes, so does the risk for floods [Milly et al., 2002]. The effect of climate change on the global water cycle results in a change of flood risk. By analyzing the risk of great floods in large basins, Milly et al., 2002 shows

a steady increase of flood events during the 20th century, which will further increase onward until 2100.

Mosavi, Ozturk, and Chau, 2018 provide an overview of machine learning approaches for flood prediction. They differentiate between short (warning systems) and long (policy objective) term predictions based on the lead-time of the prediction, whereat literature on spatial flood prediction was excluded. Their conclusion of analyzing more than 6000 articles highlights four general outcomes in order to improve the prediction quality: use of hybrid models, combining machine learning and conventional statistical/physical methods; data decomposition techniques to enhance data quality; use of ensemble methods and improved optimizer algorithms.

In contrast to slide events, floods are spatially more consistent and we can map endangered areas from past events. This shifts the research interests towards prediction of events and intensities instead of spatial susceptibility mapping. In Austria flood risk zone analysis maps based on past flood events are available and contribute to spatial flood prediction. Extending the 'event space' with additional data on catchment areas, gauge height time series and runoff data might improve its usability within the context of flood predictions.

1.2 Thesis Organization

Chapter 2 describes the data used during this study. Chapter 3 examines a difficulty analysis approach, various tree-based machine learning methods and model performance metrics. The theoretical descriptions are accompanied by their practical application on synthetically generated data and are therefore put into the context of imbalanced classification problems. Chapter 4 covers the results of the landslide susceptibility mapping approach, whereas chapter 5 continues on it's discussion and outlook.

2 Data

Data used in this study are of different kind. We use well-defined synthetic data to evaluate the performance of machine learning algorithms for comparable problems. Afterwards, we use merged damage data in combination with time independent data, which includes features derived from digital elevation models and satellite products as well as geological and physical soil properties.

2.1 Synthetic

In order to evaluate the performance of algorithms and pre-processing methods for (unevenly distributed) classification problems we generate synthetic binary classification data sets. Following up on Guyon, 2003, we create clusters of normally distributed points around the vertices of a n-dimensional hypercube. Here n represents the number of informative features. The side length of the hypercube defines the position of the clusters in the n-dimensional feature space and therefore also the separability of the classes. Two clusters per class for a binary classification problem would result in 4 clusters on 4 different vertices. At a later point we are going to use different combinations of cluster numbers and side lengths to simulate data sets of different complexity.

Figure 2.1 illustrates an example of a binary classification problem (n=3) with two clusters per vertex for each class and random noise added to the data. The bimodal distribution of features indicates two clusters per vertex. Only the first three of six features are informative. The fourth feature is a redundant linear combination of the first three features. The fifth feature is a repeated informative feature and the sixth feature represents random noise.

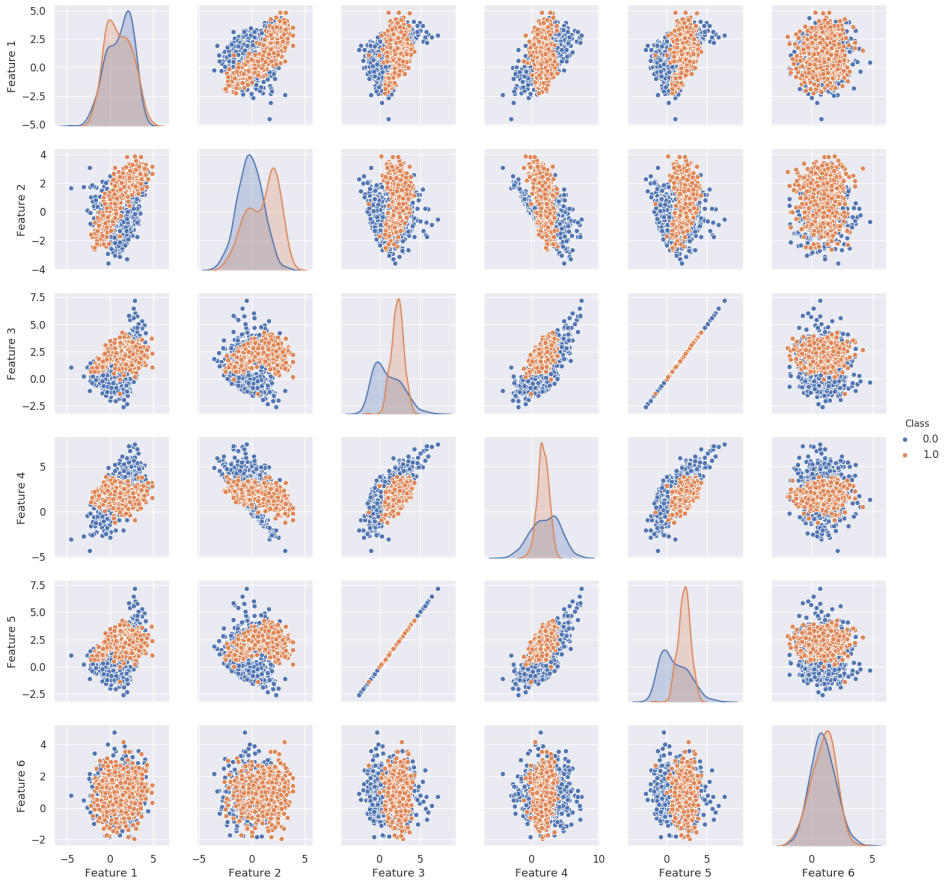

Figure 2.1: Pairplot of synthetic binary classification problem for six features.

With this synthetic data set baseline, various feature selection and classification algorithms are going to be analyzed. In order to check the performance in case of highly unbalanced data sets, we are going to unbalance the data sets by removing different percentage thresholds of data points.

2.2 Damage

Following up on Enigl et al., 2019, the damage data in this study consists of records collected by three different Austrian federal agencies, the Austrian Service for Torrent and Avalanche Control ('Wildbach- und Lawinenverbauung', 'WLV'), the Geological Survey of Austria ('Geologische Bundesanstalt', 'GBA') and the Austrian National Weather and Geophysical Service ('Zentralanstalt für Meteorologie und Geodynamik','ZAMG'). ZAMG's data set is called VIOLA (Violent Weather Assessment), whereas WLV and GBA named their data sets after the institution. VIOLA was derived from archived media reports on weather-driven damage-processes. The data is structured into following event categories: floodings, flash floods, slides, flows, falls and others.

Table 2.1 depicts those damage event categories. Each category lists the total number of events (square brackets) and the fractions from which data sources it originates. Based on the division of the underlying processes, VIOLA dominates the categories flash floods and heats, whereas floods and flows are dominated by the WLV data set. Slides and falls consist of all three data sources, but more than 60 % of the fall records are found in the GBA data set. There is a total number of 23777 damage events occurring on 5344 days between January 1959 and December 2017. If we accumulate the

Table 2.1: Total number of damage events (square brackets) and fractions of data sources in percent for each damage category.

Category	Source	Fraction [%]
falls [1123]	GBA	63.75
	VIOLA	18.34
	WLV	17.89
flash floods [9228]	VIOLA	99.82
	WLV	0.17
floods [6869]	VIOLA	7.62
	WLV	92.37
flows [2942]	GBA	2.58
	VIOLA	25.49
	WLV	71.92
heat [284]	VIOLA	100.00
others [603]	GBA	28.85
	WLV	71.14
slides [2728]	GBA	46.88
	VIOLA	42.04
	WLV	11.07

top 100 days with the highest numbers of events we get 6017 events, representing more than a fourth of the whole data set.

In the context of data fraction and damage event occurrences, it is important to note that GBA and WLV data are point located damage events, suggesting a high spatial resolution. Under different conditions, VIOLA is derived from newspaper reports after the actual damage event happened and goes back to 1948. This indirect acquisition of data results in a reduced spatial quality where the size of polygons varies heavily and depends on the degree of accuracy in the newspaper report. Hence, the spatial resolution is lower compared to WLV and GBA. In addition, single damage events may consist of multiple polygons in VIOLA. In the course of data pre-processing, events with multiple polygons, e.g. different districts or federal states of Austria, are split into multiple events of the same category but with different spatial extent and are therefore increasing the number of total damage events in VIOLA.

Figure 2.2 shows the sum of damage occurrences over all categories per year for the three data sources. All sources are steadily increasing in data volume starting from 1990 onward. This might not necessarily indicate an increase in damage events in general during the last decades, but may also be a result of modernized methods in data acquisition, e.g. cell phones and satellite imagery. The GBA in particular seems to have intensified its data acquisition in the late nineties. In the period between 1959 and 1990 the pearson correlation coefficient for the detrended WLV and VIOLA data is 0.7442. For the period between 1990 and 2017 the correlation coefficient drops to 0.5240, although high peaks of damage

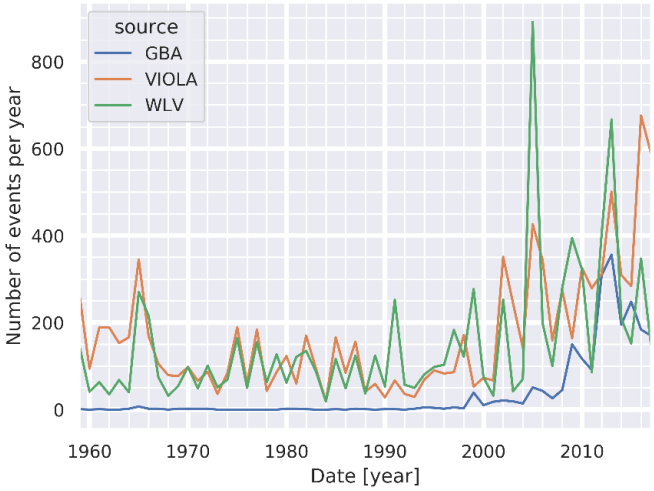

Figure 2.2: Total hazardous event sum per year

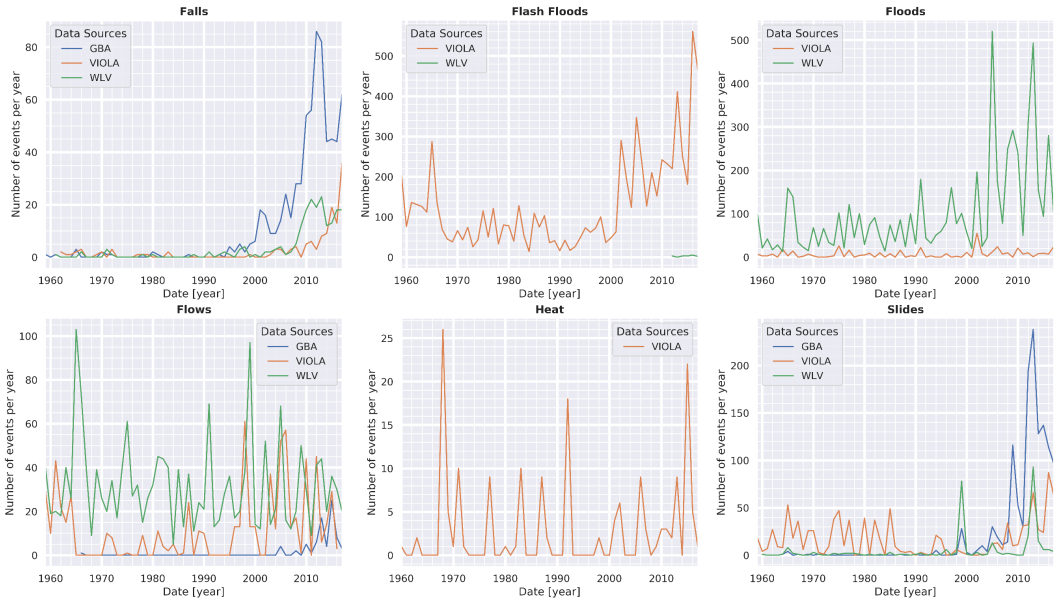

Figure 2.3: Total hazardous event sum per year for three data sources and six hazard categories.

events occur in both data sets during specific years. One possible explanation for this could be the interconnection between flash floods and floods in the first place. Heavy precipitation events, surface water and thunderstorms from VIOLA as the underlying processes in flash floods do also increase the possibility of flood events from WLV. Another explanation for some peaks during specific years is the manifestation of surpassing flood events in Austria, e.g. 1965, 1999, 2002, 2005, 2012, 2013.

When taking a look at the different damage subcategories it can be seen that falls and slides are increasing from 2000 onward with high peaks in 2012 and 2013 (see Figure 2.3). However, there is a decline in slide events for VIOLA between 1990 and 2000. The accumulated number of floods and flash floods sum up to about two thirds of the whole data set and are therefore dominating the total amount of damage events. The temporal trend in flows and heat events is not as strong as for the rest of the categories.

Within this chapter we analyze the number of damage events per data source. It is unquestionable that there is missing data. It also depends on

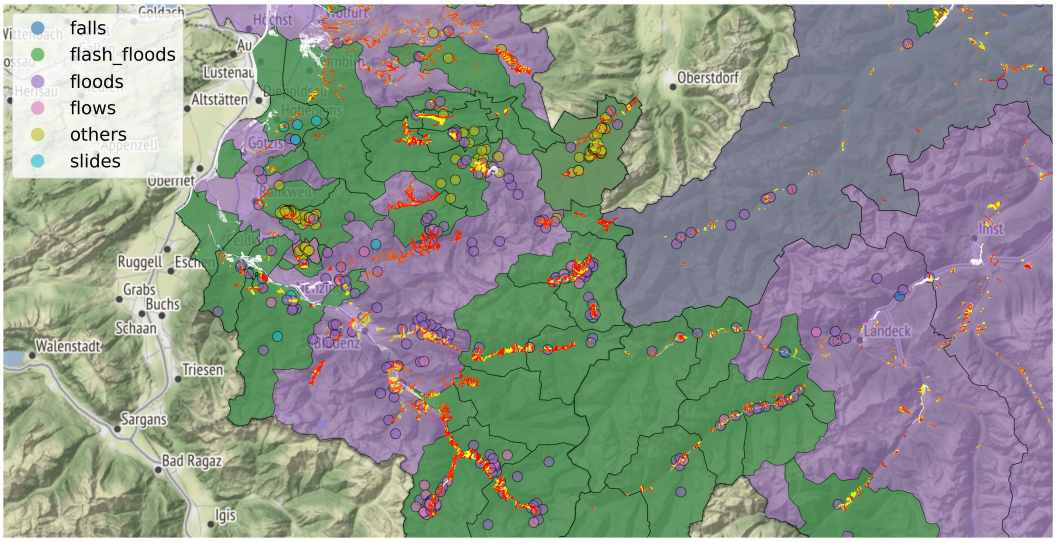

Figure 2.4: West Austria damage events between 2005/08/20 and 2005/08/23. Damage polygon (VIOLA) and damage point (WLV, GBA) data for six damage categories. Endangered zones (yellow, red). 30, 50 and 100 yearly flood zones (white).

the data acquisition methods of the collecting institute on how small single events are split up. There is no intensity of a damage event. Although, high numbers of damage events and big areas of polygons indicate clusters of outstanding damage events.

Figure 2.4 depicts all events during the period between 2005/08/20 and 2005/08/23 for the western part of Austria. The yellow and red areas indicate known endangered zones from WLV, whereas the white areas are the 30, 50 and 100 yearly flood zones. The damage event polygons represent the VIOLA data set, the dots are the WLV and GBA events. The majority of flood, flash flood and flow events lie in proximity to the endangered zones. For the whole area of Austria a total amount of 600 hazardous events were registered during this specific period of three days.

2.3 Terrain

Throughout this thesis we use the oe3d digital elevation model (DEM) to derive terrain based indices. With a resolution of one arcsecond the oe3d DEM is a quality controlled combination of local and regional data sets based on a least square method [Rechenraum, 2014]. It is dominated by the ASTER data set. However, if the quality index of ASTER is too low, SRTM and BEV data are used to increase the quality in those regions. Low performances of the ASTER data set near water surface areas and their corrections require the use of a watermask. The oe3d initiative uses openstreetmap data to correct for the water artifacts in the elevation model. The calculation of terrain derived indices requires us to re-project the oe3d data set from WGS84 (EPSG:4326) to Austrian Lambert (EPSG:31278) using bilinear resampling.

To calculate aspect, slope, curvatures (9th order polynom) and the Topographic Wetness Index (TWI) we use SAGA GIS
[Zevenbergen and Thorne, 1987; Böhner and Selige, 2006]:

- The aspect represents the direction of the slope in radians. Beginning in an eastward direction, the aspect increases counterclockwise, e.g. North is $\pi/2$ and South is $3\pi/2$. Yet, East is represented by 2π, because an aspect value of zero indicates undefined flat areas with a slope of zero.
- The slope variable describes the steepness of the slope in radians inclination compared to a horizontal flat surface.
- Profile and tangential curvatures are the curvatures in the direction of the steepest slope and in the direction of the contour tangent. They are expressed as $1/\text{meters}$, where convex forms result in positive values and concave forms are represented by negative values.
- The topographic wetness index,

  $$\text{TWI} = \frac{\ln(a)}{\tan(\beta)}$$

  combines the local upslope contributing area (specific catchment area) a and the local slope β in radians [Böhner and Selige, 2006]. High values of TWI represent areas with high potential of runoff generation.

The computation of the Terrain Ruggedness Index, the Topographic Position Index and the Roughness was done using gdaldem:

- The Terrain Ruggedness Index (TRI) is a measure of the local variation in terrain about a central pixel [Wilson et al., 2007]. Using the annotation from Wilson2007 for a 3x3 window we get

$$\begin{aligned}\text{TRI} =&(|z_{-1,1} - z_{0,0}| + |z_{0,1} - z_{0,0}| + |z_{1,1} - z_{0,0}| + |z_{-1,0} - z_{0,0}| \\ &+ |z_{1,0} - z_{0,0}| + |z_{1,-1} - z_{0,0}| + |z_{0,-1} - z_{0,0}| + |z_{1,-1} - z_{0,0}|)/8\end{aligned}$$

- The Topographic Position Index (TPI) compares the elevation of a central pixel to the mean elevation of a specified neighborhood around it [Andrew D. Weiss, 2001]. A positive TPI value indicates a central pixel with higher elevation than the average of its surrounding pixels (ridges). A negative TPI value on the other hand represents a central pixel with lower elevation than the average of its surroundings (valley). TPI values near zero are either areas of constant slope or flat areas.
- Roughness is defined as the difference between the maximum and minimum elevation for cells within a 3x3 neighborhood around the central pixel [Wilson et al., 2007].

The above mentioned indices depend on the scale of the DEM. It is not always advisable to use the highest possible resolution. The research topic itself sets the bounds, whether it is necessary to describe e.g. ridges and valleys on a detailed scale or if it is adequate to only describe main features on a rough scale. Weather-triggered hazards and damage events are of different scales and catchment areas. Nevertheless, in this thesis we use the high resolution of the oe3d data set for the calculation of the indices. The resolution is coarsened after the calculation using the average value within a 5 x 5 window.

2.4 JRC Soil Data

Topsoil physical and hydraulic properties were retrieved from the European Soil Data Centre (ESDAC) [Panagos et al., 2012].

In this thesis we use European topsoil physical properties with a resolution of 500 meters for the following parameters: clay, silt and salt content; coarse fragments; bulk density; USDA soil textural class; available water capacity. [Ballabio, Panagos, and Monatanarella, 2016] used the European top soil database (LUCAS) in combination with satellite data products from the MODIS sensor and fitted regression models to create spatially homogeneous maps of soil properties. An error estimation was conducted using cross validation and showed normalized errors between 4 and 10 %. The LUCAS topsoil data itself is derived from in-situ data in combination with satellite derived data, such as CORINE land cover and the Shuttle Radar Topography Mission (SRTM) DEMand (slope, aspect and curvature).

Furthermore, soil water information may act as important feature in slide modelling. Based on the European pedotransfer functions (PTF) [Tóth et al., 2015] the ESDAC provides maps of hydraulic soil properties on a 1 km grid resolution for Europe. They used rules PTF01, PTF07, PTF10 and PTF13 to predict saturated water content (THS), water content at field capacity (FC), water content at wilting point (WP) and saturated hydraulic conductivity (KS) based on modified FAO texture classes.

2.5 Others

Besides physical and hydraulic soil properties, the most important factors concerning slides and flows are geological and geomorphological conditions [Varnes, 1984]. Therefore, we use the Austrian geological data set provided by GBA. This data set is based on the geological map "Metallogenetische Karte" (F. Ebner, L. Weber) and has a scale of 1:500000. It provides geological information for 61 classes, whereat those classes are themselves grouped in three supergroups (tektonic, chronostratigraphic). In addition, the data set

provides separated lithological information as qualitative feature with 33 classes.

The Land Information System Austria (LISA) provides information regarding the status and development of Austria's land cover [Banko et al., 2014]. The data is available on a 10 meter spatial resolution and was derived from Sentinel-2A L2A data collected between 2015/12/08 and 2017/05/21. The categorical feature consists of eight classes: sealed areas, non-vegetated unsealed surfaces, water, snow and ice, woody, herbaceous permanent, herbaceous periodically, reeds.

Although some information with respect to vegetation is captured within the LISA land cover classification, the normalized vegetation index (NDVI) is a widely used numeric feature for landslide susceptibility mapping. Thus, we use a NDVI derived from Sentinel 2 data between August and September 2016. The data set was provided by ZAMG on a 10 meter spatial resolution.

Based on the experience of the data collecting institutes the potential for slide events is enhanced near streets. Therefore, we use openstreetmap data with more than 1.8 million streets in Austria to derive the distance of a grid cell to the nearest street. The distance was calculated within a 1 km window around the grid point, whereas for windows without any streets we set the distance to NaN.

Risk zone assessment data is only used for visualization and was not included in the model.

Table 2.2 lists all features and highlights the features used during the modelling process.

Table 2.2: Feature short names, long names and types grouped by data sources

Short Name	Long Name	Type
oe3d_aspect	aspect	numeric
oe3d_slope	slope	numeric
oe3d_twi	topographic wetness index	numeric
oe3d_tri	topographic roughness index	numeric
oe3d_tpi	topographic position index	numeric
oe3d_profilecurv	profile curvature	numeric
oe3d_plancurv	plan curvature	numeric
oe3d_tancurv	tangential curvature	numeric
oe3d_loncurv	longitudinal curvature	numeric
oe3d_gencurv	general curvature	numeric
jrc_ts_clay	topsoil clay content	numeric
jrc_ts_sand	topsoil sand content	numeric
jrc_ts_silt	topsoil silt content	numeric
jrc_ts_coarse_frag	topsoil coarse fragments content	numeric
jrc_ts_texture	USDA soil textural classes	numeric
jrc_ts_bulk_density	bulk density	numeric
jrc_ts_awc	available water capacity	numeric
jrc_hy_ths	saturated water content	numeric
jrc_hy_fc	water content at field capacity	numeric
jrc_hy_wp	water content at wilting point	numeric
jrc_hy_ks	saturated hydraulic conductivity	numeric
geo_bnd_distance	distance to geological boundary	numeric
geo_legtext	geology legend text	factor
geo_lithol	lithology classification	factor
geo_ueber1	geology supergroup one	factor
geo_ueber2	geology supergroup two	factor
geo_ueber3	geology supergroup three	factor
geo_age	geology age	factor
lisa_landcover	land cover classification	factor
cop_ndvi	normalized vegetation index	numeric
osm_street_distance	distance to street (open street map)	numeric

3 Methods

This chapter provides an overview of the basic concepts of the supervised learning methods used throughout this work. The applicability of the methods is demonstrated on different synthetic data sets (see chapter 2.1).

Section 3.1 describes a data difficulty analysis approach using a k-nearest neighbor method. The basics of supervised learning methods are described in Section 3.2. This section includes linear logistic regression and feature selection methods as well as cross validation, bagging and boosting. Section 3.3 describes tree-based supervised machine learning methods used for regression and classification tasks, whereas the latter was needed for binary classification problems described in this work. In order to assess the quality of a model and its prediction, several performance metrics are described in section 3.4

3.1 Data Difficulty Analysis

Following Napierala and Stefanowski, 2016 we want to quantify difficulty factors of the data. Those factors may arise from differences in class overlapping and fragmentation or the presence of outliers and noise in the data. We perform a k-nearest neighbor (k-NN) analysis of the local neighborhood of all minority labeled data points.

With the class information of the neighboring data points the minority examples can be grouped based on the ratio of minority to majority neighbors. If there is a ratio of 5:0 or 4:1 the example is safe, hence, five or four neighboring data points have the same label as the minority example. 3:2 or 2:3 is a borderline example. 1:4 is a rare examples and 0:5 is an outlier without any neighbors of its own minority class.

In order to perform a k-NN analyses it is necessary to choose k and a distance metric. Napierala and Stefanowski, 2016 showed that the identified types of examples remain more or less stable for $k >= 5$. Thus, we chose $k = 5$. We do not use qualitative features in this analysis and therefore do

Table 3.1: Ten synthetically generated data sets [Guyon, 2003], their defined creation properties (separability, number of clusters), global imbalance ratio (IR) and the number of minority examples per difficulty group.

#	Separability	Clusters	IR	# Safe	# Borderl.	# Rare	# Outlier
1	1.2	2	3.9	3728	215	27	88
2	1.2	2	8.6	1733	195	45	109
3	1.2	2	65.0	93	49	36	125
4	2.6	2	8.6	1985	0	2	95
5	0.6	2	65.0	0	21	47	235
6	0.1	4	99.0	0	2	21	177
7	1.6	2	99.0	62	48	33	57
8	0.8	2	99.0	4	24	45	127
9	1	2	99.0	30	27	35	108
10	0.2	2	99.0	0	2	23	175

not use the HVDM metric [Wilson et al.(1997)]. We chose the Minkowski distance metric,

$$D(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

with $p = 2$. Thus, resulting in the euclidean distance which is a proper metric for quantitative features (real valued vector spaces) [Batista and Silva, 2009].

Table 3.1 lists the data creation settings, the difficulty analysis and the global imbalance ratio (IR) of majority to minority class members. All ten data sets consist of 20000 samples and 16 features. As described in chapter 2.1 there are three informative features, a linearly combined feature and a repeated feature. Features six to sixteen are random noise features. The first three data sets have the same separability and two clusters per vertex but different IRs and 1 % of the majority class values are wrongly classified as minority class. Data set 4 has a high separability and two clusters, whereas data set 5 (two clusters) and 6 (four clusters) have low class separability. Data set 7 to 10 were created with different random states. For all data sets the difficulty analysis is plausible in the context of the given settings during the data creation process. Although data set 6 and 10 have almost identical difficulty

analysis estimations, they differ in the number of clusters per class. This fact implies a higher difficulty of data set 6 as there are double the amount of normal distributed sample clusters on every vertex of the three dimensional hypercube.

3.2 Elements in Supervised Learning

The general idea of supervised learning [Gareth et al., 2019; Hastie, Robert, and J. Friedman, 2017] can be explained by looking at three key concepts: model, parameter and objective function [T. Chen and Guestrin, 2016]. $x_i \in \mathbb{R}^m$ represents a vector of m features for the i-th sample (n total number of samples).

Model

The model describes how the algorithm makes predictions based on the given input data x_i (features). A simple model is the linear model, where the prediction is generated using a linear combination of the input data and $\hat{y}_i$ is the predicted score:

$$\hat{y}_i = \sum_j w_j x_{ij}$$

Parameter

The parameters Θ are learned from the training data x_i considering the objective function below.

For a linear model Θ is a vector where each entry represents the weight of one feature in the linear model: $\Theta = \{w_j | j = 1, \ldots, m\}$

Objective function

The objective function (also called loss or cost function) measures the performance of the model based on the given parameters.

$$\text{Obj}(\Theta) = L(\Theta) + \Omega(\Theta)$$

The terms on the right-hand side are called training loss L and regularization term Ω. The training loss indicates how well the model fits the data:

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i)$$

Square and logistic loss terms:

$$\begin{aligned} \text{Square Loss:} \quad & l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2 \\ \text{Logistic Loss:} \quad & l(y_i, \hat{y}_i) = y_i \ln\left(1 + e^{-\hat{y}_i}\right) + (1 - y_i) \ln\left(1 + e^{\hat{y}_i}\right) \end{aligned}$$

The regularization term might be interpreted as the complexity of the model and is also used to penalize complex models in order to prefer simpler models when minimizing the objective function.

L1 and L2 regularization terms:

$$\begin{aligned} \text{L1:} \quad & \Omega(w) = \lambda \|w\|_1 = \lambda \sum |w_j| \\ \text{L2:} \quad & \Omega(w) = \lambda \|w\|^2 \end{aligned}$$

where λ is as a tuning parameter and controls the impact of the regularization term on the objective function.

The following subsection uses these concepts to summarize attributes of the feature selection methods ridge and lasso.

Feature Selection

Unrelated features may not be associated with the response of the model but increase the model's complexity unnecessarily. In order to evaluate the importance of features we use the shrinkage methods ridge and lasso. These approaches fit linear models on all features, but the coefficients of the features are shrunk towards zero. The shrinkage of coefficients tends to result in a big decrease of variance and a small increase in bias. Thus, it reduces the mean squared error [Gareth et al., 2019].

For the logistic version of lasso and ridge, a generalized linear model is used with a logistic link function:

$$\text{logit}(p_i) = \ln \frac{p_i}{1 - p_i} = \sum_j w_j x_{ij}$$

where p_i represents the probability of the i-th sample being positive.

Logistic Ridge

The above generalized linear model is used in case of ridge feature selection with a logistic loss and L2 regularization in the objective function:

$$\text{Obj}(w) = \sum_{i=1}^{n} y_i \ln \left(1 + e^{-w^T x_i}\right) + (1 - y_i) \ln \left(1 + e^{w^T x_i}\right) + \lambda \|w\|^2$$

Logistic Lasso

The above generalized linear model is again used, but with a logistic loss and L1 regularization in the objective function which results in the lasso feature selection approach:

$$\text{Obj}(w) = \sum_{i=1}^{n} y_i \ln \left(1 + e^{-w^T x_i}\right) + (1 - y_i) \ln \left(1 + e^{w^T x_i}\right) + \lambda \|w\|_1$$

Synthetic Feature Selection

To estimate λ a k-fold cross validation is performed on the data. A typical value is $k = 10$, it is one of the most often chosen values for k throughout the analyzed literature.

Ridge regression has one major disadvantage compared to lasso regression. It involves all features and only shrinks coefficients towards zero. If λ is sufficiently big, Lasso and its L1 regularization, on the other hand, forces coefficients to become zero.

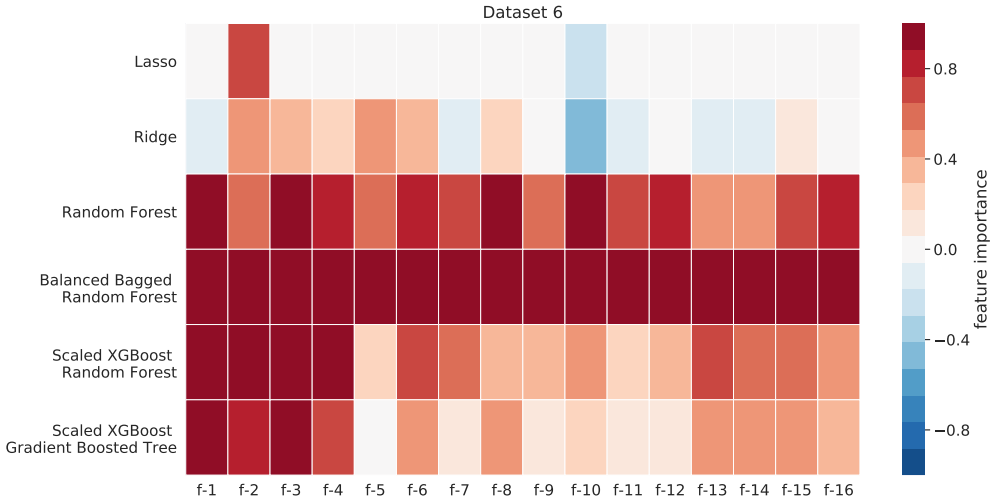

Figure 3.1: Logistic (Ridge and Lasso) and non-linear (tree-based methods) scaled feature importance analysis for the difficult synthetic data set 6 (table 3.1).

Even though a direct comparison of linear and non-linear feature importance is not reasonable, the interpretability of non-linear tree-based model performance estimation (section 3.3) benefits from a consistent feature importance analysis of different methods. For tree-based models, feature importance cannot be analyzed independently of other features and the tree's increase in purity due to a features split (gain due to split) is used to assess its importance.

For the sake of synopsis, linear feature selection methods and tree-based feature importance analysis are put together in figures 3.1 and 3.2. Analyzing the feature importance of the synthetic imbalanced data sets, the linear approaches (lasso and ridge) and the non-linear tree-based approaches perform well for the synthetic data sets 1 to 4 and identify features one, two and four as informative features. Data sets 5, 6, 9 and 10 have a lower separability and are therefore more difficult.

Figure 3.1 shows the scaled feature importance analysis of the most difficult data set 6. Lasso and ridge cannot identify the informative features and are "conservative" in a way not to put high weight on random features. Random Forest (RF) and Balanced Bagged Random Forest (BBRF) are not

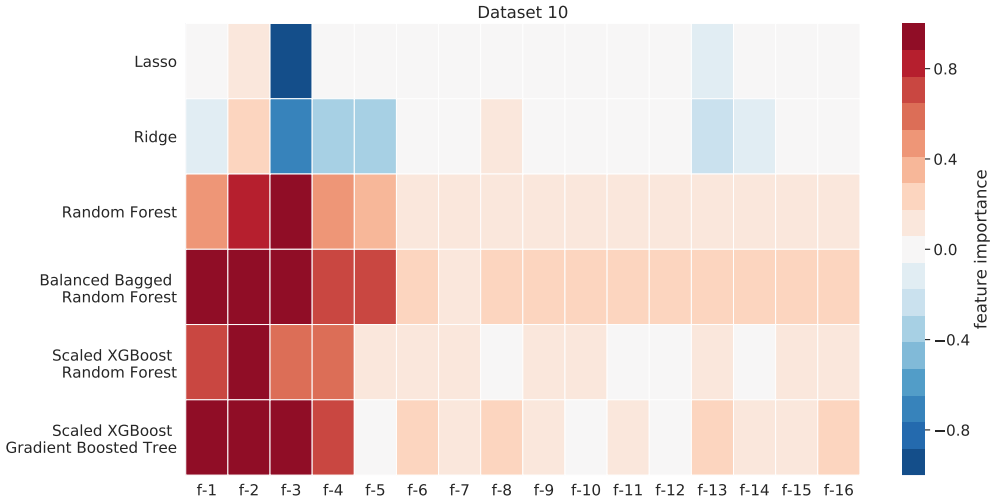

Figure 3.2: Logistic (Ridge and Lasso) and non-linear (tree-based methods) scaled feature importance analysis for the difficult synthetic data set 10 (table 3.1).

able to identify informative features but "gain" from all features instead. Scaling the minority class in the objective function with Scaled XGBoost Random Forest (SXRF) and Scaled XGBoost Gradient Boosted Trees (SXGBT) improves the performance of the model, where the highest gain arises from the first four features. Nevertheless, all tree-based models include random features and extreme caution must be exercised when interpreting feature importance for imbalanced data sets. The feature importance analysis of data set 10 (figure 3.2) is more consistent and all tree-based methods gain most from features one to four.

Cross Validation

A performance estimation of a trained machine learning model is important. The chosen performance metric and its difference between train and test data set estimates the performance of predictions on unknown data and acts as indicator on how well the model generalizes. A simple approach is to split the data, e.g. 3/4 of the data to train the model and 1/4 of unseen data for the validation part.

A more advanced approach is the k-fold cross validation. We split the data in k randomly chosen non overlapping groups of equal size. Every single group acts as validation data set, whereas the rest of the data is used to train the model. With this approach we get k performance estimations.

In a nested cross validation we use the $k-1$ training groups from the "outer" cross validation and perform an additional "inner" cross validation. The inner cross validation is used to tune the hyper parameters of the chosen machine learning method. The best performing model setting combination from the inner cross validation is then applied on the outer cross validation test group.

3.3 Tree-Based Methods

A separation of the predictor space into multiple non-overlapping regions can be visualized in form of a tree. Trees are simple and interpretable, but they are outperformed by more sophisticated ML methods such as Support Vector Machines and Generalized Additive models. To improve the accuracy of tree-based models, ensemble methods such as random forest and boosting were developed. Multiple trees are trained and combined in order to improve the skill of the model at the cost of interpretability. Tree-based methods are used for classification and regression tasks.

Decision Trees

When growing a tree the objective is to divide the predictor space in regions such that the predicted values (mean of region, majority class of region) minimize the square loss (regression trees) or the gini index (classification trees). It is computational not possible to consider all combinations of splits. Hence, recursive binary splitting is often used to grow a tree. Although there might be a better combination to improve the gain of the tree in a futures split, this top down approach minimizes the objective recursively at each particular step without looking ahead and starts at the top of the tree where all observations belong to one region.

Such grown trees tend to become complex and often overfit on the training data which in turn results in poor performances on test data. Growing a complex tree and selecting a subtree afterwards is called tree pruning. Gareth et al., 2019 describes cost complexity pruning as one method to select a subtree.

Bagging

Bootstrap aggregation (bagging) as a general purpose method is also useful in order to reduce the variance of tree-based models. The idea is to create an ensemble of decision trees on bootstraped training data. In other words, we draw "new" data sets with replacement from the original data (bootstrap) and train unpruned decision trees, where each tree has high variance. Averaging or majority voting the results (predictions) of those trees reduces the variance of the ensemble model [Breiman, 1996; Gareth et al., 2019].

Random Forest

Features with high predictive information may lead to correlated trees (similar tree structure) in the bagging approach. The random forest ensemble method addresses this downside of the bagging approach and decorrelates the trees by using a subset of features for the estimation of the best split at each node [Breiman, 2001]. If we choose all features as a "subset", the random forest approach would be equal to the bagging approach.

Boosting

Compared to bagging and random forest, boosted trees are not grown independently but sequentially. This implies that information from previous trees is considered when growing a new tree. The bootstrap method is not used when building a boosted model and instead of one large tree many small trees (stumps) are fitted on the data. Therefore, boosted models learn

slower and do not tend to overfit the training data as fast as other methods [J. H. Friedman, 2001].

We use the XGBoost framework developed by T. Chen and Guestrin, 2016 and follow their notation for a data set with n examples and m features $D = \{(\mathbf{x}_i, y_i)\}$ ($|D| = n, \mathbf{x}_i \in \mathbb{R}^m, y_i \in \mathbb{R}$). The **model** for K additive tree functions is

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i), \quad f_k \in \mathbb{F}$$

where $\mathbb{F} = \{f(\mathbf{x}) = w_{q(x)}\}(q : \mathbb{R}^m \to T, w \in \mathbb{R}^T)$ represents the tree space and q maps an example to a leaf index of the tree. f_k is an independent tree structure where each tree has T leaves and w_i represents the score of the i-th leaf. There is no such score in a decision tree. The functions representing the trees are the **parameters** $\Theta = \{f_1, f_2, \ldots, f_k\}$.

The **objective function**

$$\text{Obj}(\Theta) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$$

with $\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$ contains functions in the tree ensemble model and therefore needs to be trained in an additive manner (boosting):

$$\text{Obj}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)\right) + \Omega(f_t)$$

where $\hat{y}_i^t$ represents the prediction of the i-th instance at t-th iteration.

T. Chen and Guestrin, 2016 use a second order Taylor approximation and remove the constant terms, which leads to a simplified objective function at the iteration step t

$$\tilde{\text{Obj}}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_t)$$

where $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ are loss function gradient statistics.

By further defining an instance set of the j-th leaf $I_j = \{i|q(\mathbf{x}_i) = j\}$, they compute an optimal weight of leaf j and derive a gini equivalent but more general scoring function:

$$\tilde{\text{Obj}}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T$$

The learning procedure of f_k now only depends on g_i and h_i. To circumvent the problem of infinite possible tree structures q, the tree is grown greedily by adding one branch after another at the best split. The best split in xgboost is determined using the exact greedy algorithm or an approximation of it. For the exact greedy algorithm the data is sorted by features and a linear scan from left to right is performed in order to find the best split. For the approximation of it, the data does not need to be sorted and does not have to fit into memory entirely.

3.4 Model Quality Assessment (Metrics)

The evaluation of the model performance for highly skewed data sets is exacerbated. Many standard metrics result in misleading conclusions as they do not account for the imbalance of the classes. During this section we want to list the most commonly used metrics in the context of skewed data distributions. Many of them are derived from the confusion matrix (table 3.2) which consists of the numbers of correct predictions (true positive, true negative) and wrong predictions (false positive, false negative) of the model.

Accuracy is defined as $(\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$

Precision (positive predicted value) is defined as $\text{TP}/(\text{TP} + \text{FP})$

Fallout (false positive rate) is defined as $\text{FP}/(\text{FP} + \text{TN})$

Sensitivity (true positive rate or recall) is defined as $\text{TP}/(\text{TP} + \text{FN})$

Specificity (true negative rate) is defined as $\text{TN}/(\text{TN} + \text{FP})$

Table 3.2: Binary confusion matrix: True Positive (TP), False Positive (FP), False Negative (FN), True Negative (TN)

		True Class		
		Positive	Negative	Total
Predicted Class	Positive	TP	FP	TP+FP
	Negative	FN	TN	FN+TN
	Total	TP+FN	FP+TN	

F1-Score is defined as the harmonic mean of precision and sensitivity $2TP/(2TP + FP + FN)$

The precision and the f1-score are important metrics for a model's performance estimation in the context of imbalanced data domains. If the minority class is scaled in the objective function, the model is forced to predict higher probabilities more often and false positive samples are therefore high. In combination with highly skewed class distributions this fact results in very low scores.

If we vary the discrimination threshold of our model, the resulting values of fallout (x-axis) and sensitivity (y-axis) can be plotted against each other as the so called receiver operator curve. The area under the receiver operator curve (**AUC**) is one value and measures the models ability to correctly classify the outcome.

Instead of fallout and sensitivity, the **AUCPR** is calculated as the area under the precision (x-axis) recall (y-axis) curve. It is more sensitive to imbalanced data than the AUC. J. Davis and Goadrich, 2006 showed differences between the curves and demonstrates that optimizing the AUC does not guarantee to optimize the AUCPR.

Figure 3.3 displays bar plots of 10-fold cross validated AUC scores for three different model strategies on our synthetic data sets. An AUC score of 0.5 would be a random guess and an AUC score of 1 would be a perfect predicting model. The random forest tree depths are 5 and the feature's subset used to estimate the best split at each node is 80 percent. BBRF uses bootstrap with replacement, whereas SXRF bootstraps without replacement, meaning that every sample may occur 0 or 1 times in a training sample. For gradient boosted trees, tree depths are in general smaller and therefore

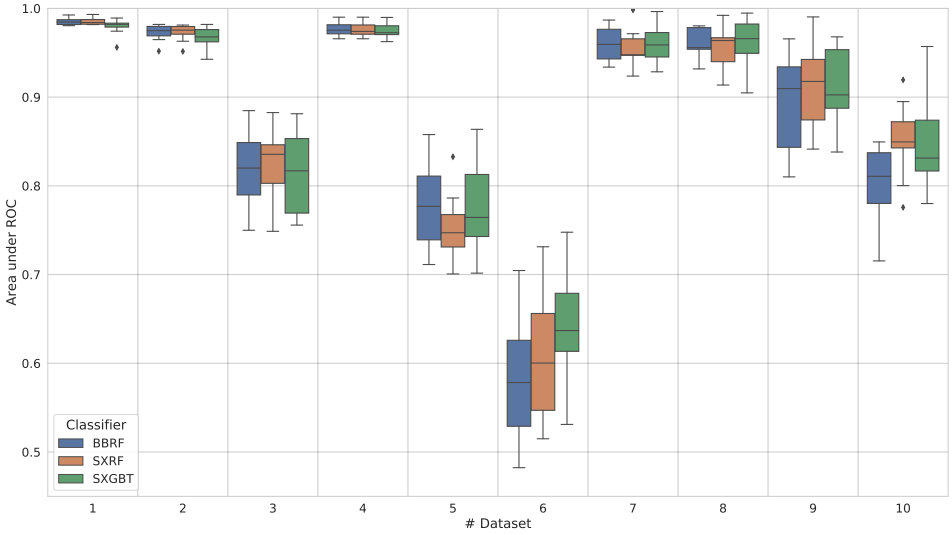

Figure 3.3: Area under Receiver Operator Curve for 10 synthetic data sets (table 3.1) using three different model strategies. Balanced Bagged random forest (BBRF), Scaled XGBoost Random Forest and Scaled XGBoost Gradient Boosted Trees (SXGBT)

SXGBT has a tree depth of 3. SXRF and SXGBT scale the minority class with the imbalance ratio in the objective function. All models perform in a similar range for each data set and it is not possible to constitute a best model approach.

The most difficult data set 6 has the lowest score followed by data sets 3, 5 and 10. Data set 3 has an original unbalance ratio of 99 but 1 % of the majority class examples are marked as minority examples and therefore reduces the analyzed IR to 65. The AUC is not sensitive to imbalanced data and data set 7 and 8 have a missleading high score. The amount of false positive values for data set 8 is up to five times the amount of true positives, on the contrary data set 2 has 1/10 false positives compared to the amount of true positives.

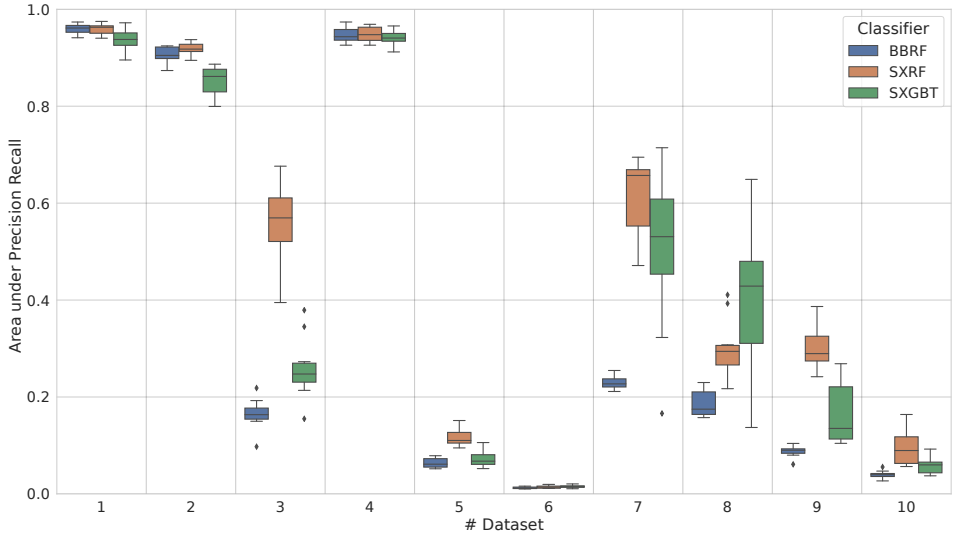

Figure 3.4: Area under Precision Recall Curve for 10 synthetic data sets (table 3.1) using three different model strategies. Balanced Bagged random forest (BBRF), Scaled XGBoost Random Forest and Scaled XGBoost Gradient Boosted Trees (SXGBT)

Figure 3.4 shows bar plots of 10-fold cross validated AUCPR scores for the synthetic data sets using the same three model strategies. This score is more sensitive to imbalance ratio of the data. The prediction for data set 8 is clearly not as good as it is for data set 2. We also see a difference in the predictions between the different model strategies, whereat scaled random forest performs better in almost every data set. Especially for data set 3 with the flipped majority class examples SXRF clearly outperforms balanced bagged random forests and scaled gradient boosted trees.

4 Results

Throughout this study we investigate three main objectives to assess the potential of tree-based algorithms for slide susceptibility mapping in Austria. At first, we perform a difficulty analysis on multiple geological domains in order to estimate the purposes of a model approach beforehand. Secondly, we perform an extensive modelling approach including geological domains as one-hot-encoded features. Last but not least, we conduct a feature importance analysis on the best performing model run.

As mentioned in section 2 the oe3d derived indices are coarsened by averaging 5 x 5 grid cells, resulting in an approximate spatial resolution of 100 meters. The center of this 100 meter grid cell is then used to homogenize the rest of the features using a nearest neighbour approach. Grid points within a 100 m buffer around slide events are marked as slides. For the following difficulty analysis we use numeric features only (table 2.2).

Table 4.1 depicts the resulting difficulty analysis for geological supergroup one defined by the GBA. The spatial distribution of the supergroup and the analyzed slide events are illustrated in figure 4.1. Imbalance ratios range from 453 up to 10619. Safe samples are only present within the "penninic units". The biggest domain in terms of slide events and spatial extent is the "austroalpine unit" with almost 5 times the amount of outliers compared to the sum of borderline and rare samples. On the contrary, the smallest domain, "periadriatic intrusive rocks", contains only five outliers. The most difficult domain seems to be the "penninic unit" followed by "helvetic zone", "austroalpine units", "quaternary" and "tertiary basins". The spatial distribution shown in figure 4.1 and the high imbalance ratio for the "bohemian massif" indicate small numbers of slide events throughout the domain.

Furthermore, the outcome of this difficulty analysis is put into context with the synthetic data experiments in chapter 3. It shows that the difficulty of the slide event data set lies somewhere between the most difficult data set 6 and the second most difficult data set 10 of the synthetic data, but with higher imbalance ratios. For the synthetic experiments this is the transition zone where the model results become doubtful and in contrast to the synthetic

Table 4.1: Difficulty Analysis for slide events grouped by supergroup one from the geological data set (GBA). Imbalance ratio (IR) and grouped slide samples based on the ratio of minority to majority neighbours in feature space: Safe (5:0, 4:1), Borderline (3:2, 2:3), Rare (1:4), Outlier (5:0)

Supergroup One	IR	# Safe	# Borderl.	# Rare	# Outlier
Quaternary	1449.7	0	9	78	600
Helvetic Zone i.g.	453.2	0	2	20	65
Austroalpine Units	1606.8	0	24	282	1391
Penninic Units	1282.7	3	12	86	270
Tertiary Basins	1830.9	0	1	70	505
Periadriatic Intrusive	1125.0	0	0	0	5
Southalpine Units	2128.7	0	2	4	21
Bohemian Massif	10619.9	0	0	4	53

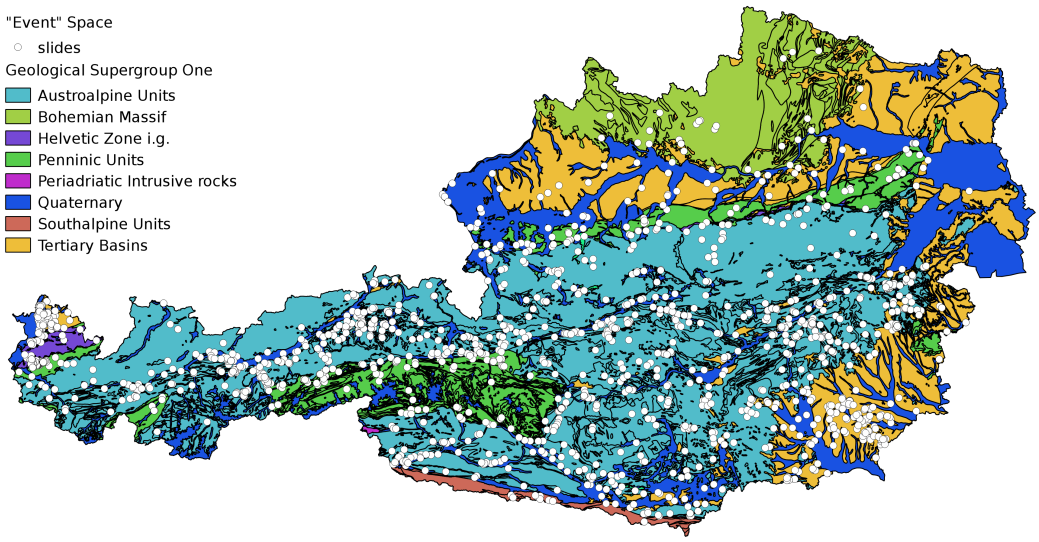

Figure 4.1: Spatial distribution of geological supergroup one (GBA) and slide events

data set, we do not know if our features truly contain characterizing elements of slide processes. Having said that, the following part depicts training and evaluation of a single domain model for Austria, including the geological factor variables geo_ueber1 and geo_lithol as one-hot-encoded features.

We conduct the modeling approach of slide events using scaled random forest and scaled gradient boosted tree models. A 5-fold cross validation is performed in order to train the models and estimate their performance metrics on hold out sets. The gradient boosted trees are grown with tree depths of 3 and 5, whereat for random forest the trees are deeper with a depth of 15. For all models the imbalance ratio of the data set is used in order to scale the minority class in the objective function. Due to limited hardware resources, we do not use nested cross validation to tune the hyper-parameter within each cross validation step. Instead, the same model parameter setting was used throughout a model run.

Table 4.2: Performance metrics of 5-fold cross validation for different model strategies: Scaled XGBoost Gradient Boosted Trees (SXGBT), Scaled XGBoost Random Forest (SXRF). The number in brackets represents the tree depth of the classifier.

CV	Classifier	Accuracy	Sensitivity	Specificity	AUC	AUCPR
0	SXGBT (3)	0.8978	0.6311	0.8979	0.7645	0.002338
1	SXGBT (3)	0.8970	0.6752	0.8971	0.7861	0.002603
2	SXGBT (3)	0.8937	0.6334	0.8938	0.7636	0.002266
3	SXGBT (3)	0.8999	0.6548	0.9000	0.7774	0.002533
4	SXGBT (3)	0.8934	0.6334	0.8935	0.7635	0.002261
0	SXGBT (5)	0.8548	0.7446	0.8549	0.7997	0.002353
1	SXGBT (5)	0.8196	0.7862	0.8196	0.8029	0.002100
2	SXGBT (5)	0.8842	0.7429	0.8843	0.8136	0.002897
3	SXGBT (5)	0.8201	0.7943	0.8201	0.8072	0.002142
4	SXGBT (5)	0.8839	0.6942	0.8840	0.7891	0.002571
0	SXRF (15)	0.8851	0.7446	0.8853	0.6040	0.001959
1	SXRF (15)	0.8836	0.7425	0.8838	0.6011	0.001922
2	SXRF (15)	0.8795	0.7351	0.8796	0.5906	0.001810
3	SXRF (15)	0.8823	0.7401	0.8824	0.5977	0.001883
4	SXRF (15)	0.8825	0.7431	0.8827	0.6034	0.001915

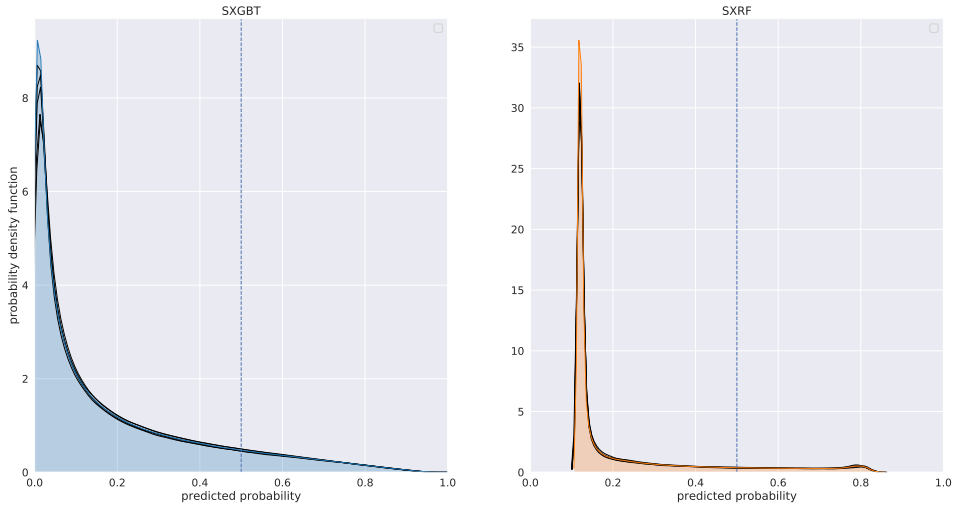

Figure 4.2: Probability density estimation of class slides for the models SXGBT(5) and SXRF(15). The blue vertical dashed line indicates the class discrimination threshold at 0.5. Values bigger than this threshold are classified as slide events. The total imbalance of the dataset is 6393347 non slide events and 3507 slide events.

Table 4.2 depicts performance scores for three model runs. The numbers in brackets indicate the tree depth of the classifier. For our slide classification task, gradient boosted trees outperform random forests. As mentioned in section 3.4, false positive samples are high due to the scaling of the minority class in the objective function and result in very low precision and therefore also AUCPR scores. The total imbalance ratio of 1823 is also visible in the equality of accuracy and specificity due to high numbers of majority samples dominating them both. Averaging over all cross validations, the model SXGBT with tree depth 5 and a learning rate of 0.3 classifies approximately 3/4 of all slide events correct. The similarity of performance metrics across all CV iterations indicates the robustness of the approach.

Furthermore, the similar shapes in probability density function over all CV iterations for SXGBT(5) and SXRF(15) in figure 4.2 underline this assumption. The SXGBT(5) model has its peak in probability density near 0 and decreases continuously until 1. Whereas SXRF(15) has its peak near 0.1 and decreases until 0.7 and has another small peak near 0.8. Although accuracy and sensitivity scores are of similar range for both models, the predictive

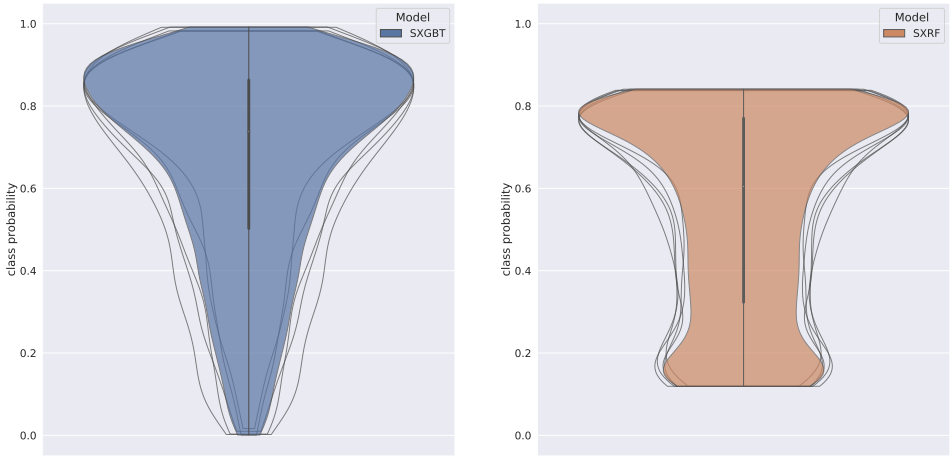

Figure 4.3: Gaussian kernel density estimation and boxplot of predicted class probability for 3507 grid cells lying within a 100 meter perimeter to the observed slide events. The grey lines indicate the 5 cross validations, whereas the filled violinplot combines all CVs in one kernel density estimation.

performance for the minority class is much better in SXGBT(5) and results in higher AUC scores.

Figure 4.3 shows gaussian kernel density estimations and a boxplot for the predicted class probabilities of 3507 grid cells lying within a 100 meter perimeter to the observed slide events. The grey lines are the 5 cross validations, whereas the filled violinplot combines all CVs. Overall, the SXGBT(5) model outperforms SXRF(15). The median of SXGBT(5) lies at approximately 0.75 and the lower/upper quartiles at around 0.5/0.86. SXRF(15) on the other hand, has its median at approximately 0.6 and its lower/upper quartiles at around 0.33/0.76. The maxima in probability density at 0.8 and 0.2 (figure 4.2) are still visible, although we only consider cells with positive slide events.

Before discussing the feature importance, we want to take a look at the feature correlation matrix (see figure 4.4). Due to the DEMs impact on the calculation of various features, it is important to analyze the linear correlations between them. Slope is predominant and many terrain based indices such as TWI and TRI are derived from it. The topsoil property coarse fragmentation also shows a high correlation of 0.7 with respect to slope.

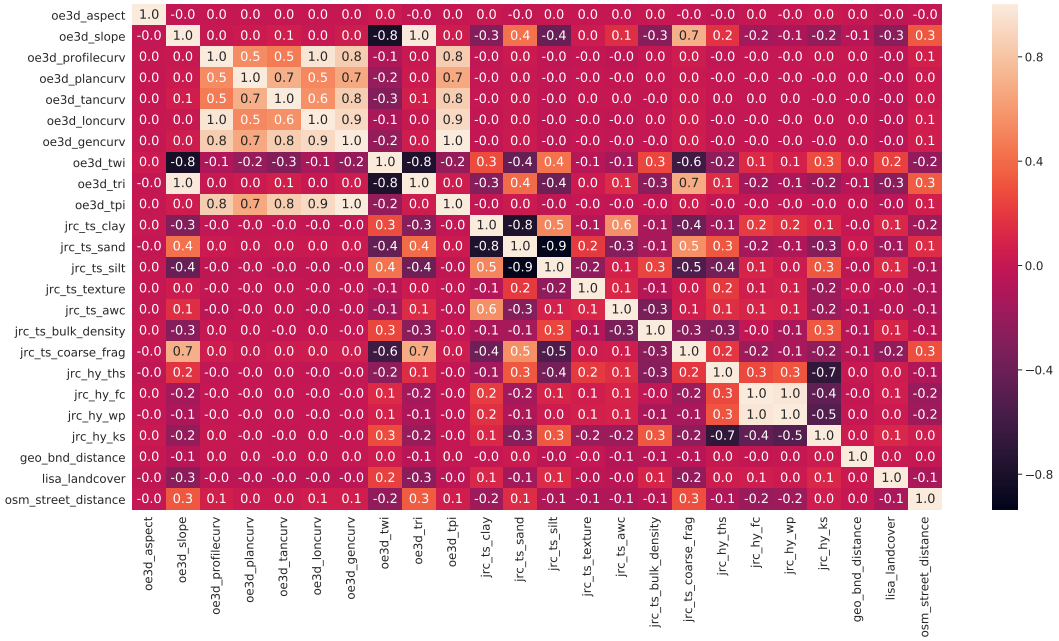

Figure 4.4: Heatmap showing the linear feature correlation matrix

All curvature features show high correlations between each other and are highly correlated with the topographic position index. Strong correlations are also present between the jrc topsoil parameters sand and silt as well as between the hydraulic parameters water content at field capacity and water content at wilting point.

Although there is an impact of linear correlation on the feature importance analysis, the non-linear nature of tree-based methods, nevertheless, justifies the consideration of all variables. However, it must be taken under consideration that when the model splits on one feature, it reduces the information of the correlated feature and therefore also reduces its potential gain.

The feature importance analysis depicts the 20 most important features of model SXGBT(5). When comparing binary with numeric features it can be seen that their average gain per split and their total gain differ vastly. Binary features, on the one hand, may have a very high gain per split as they can either be ”1” or ”0”. However, they can only be used once per branch in the model. Numeric features, on the other hand, can be used several times

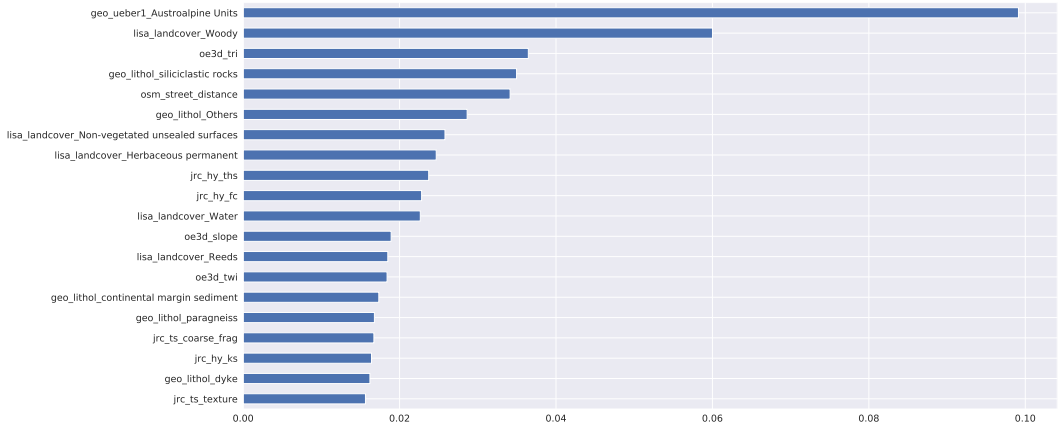

Figure 4.5: Feature importance analysis: model's average gain per split for the topmost 20 features.

in the model. Hence, their total gain is higher than the total gain of binary features.

Figure 4.5 shows the average gain per split for the first 20 features. The accumulated feature importance values sum up to 1. The binary features ”austroalpine units” and the landcover class ”woody” are followed by the numeric feature ”topographic roughness index”. The latter has a strong correlation with ”slope”, indicating that ”slope” is also among the most important features. The fourth most important feature is ”siliclastic rocks”, which further sub-divides the ”austroalpline units”. Moreover, it seems that ”street distance” is an extremely important feature. However, it is not clear whether the feature is in fact important or simply a result of artifacts from the data collection process (further explained in section 5).The sixth most important feature is ”lithology others”. It is noteworthy that when combining ”austroalpine units” and ”lithology others” only ”bohemian massif” and ”penninic units” are left over, where hardly any damage events occur.

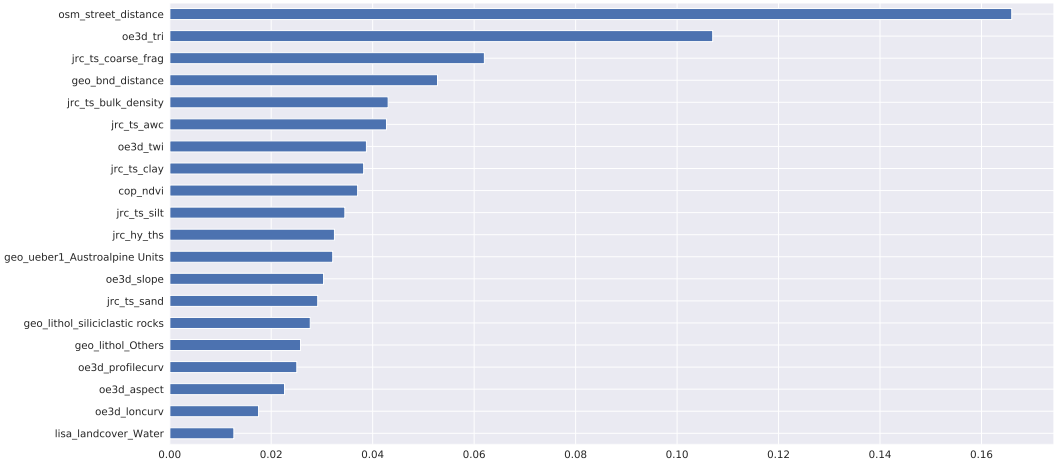

Figure 4.6: Feature importance analysis: model's total gain considering every split in all trees for the topmost 20 features.

Figure 4.6 depicts the total gain over all splits for the 20 most important features. Numeric features are more important in terms of total gain resulting in the first 11 features being of this kind. The most important feature is "street distance" followed by "topographic roughness index" and "topsoil coarse fragmentation" (both being correlated with "slope"). The fourth most important feature is "distance to nappe boundary" which is the only geological feature. The following two features "bulk density" and "available water content" are both physical topsoil properties. Thus, among the 6 most important features three of them are physical topsoil properties. Again considering the linear correlations between slope and the features "oe3d_tri", "jrc_ts_corase_frag" and "oe3d_twi", slope is clearly the most dominating factor in landslide susceptibility mapping.

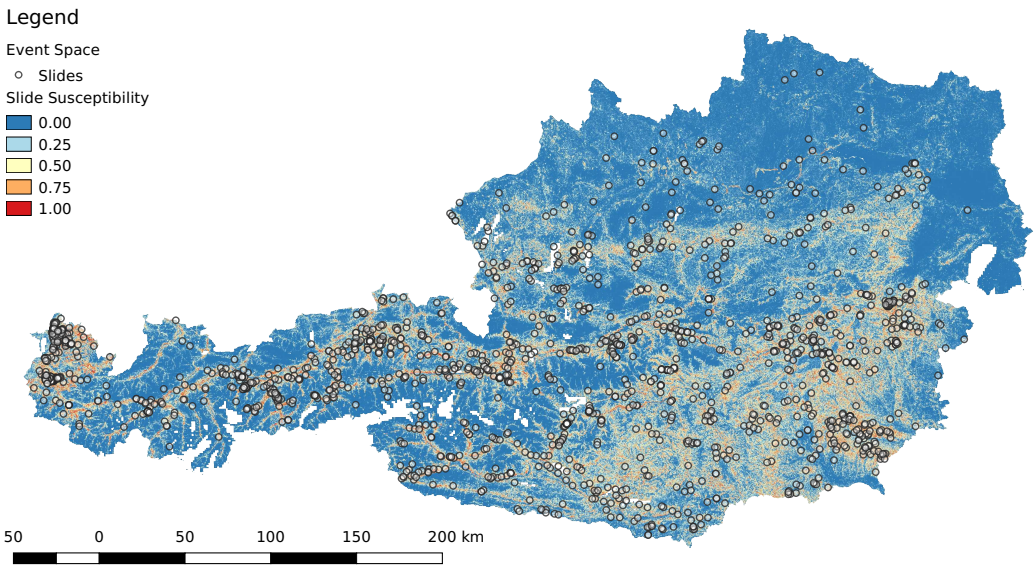

Figure 4.7: 'Event space' slide records and slide susceptibility map from model SXGBT(5)

Figure 4.7 illustrates the 'event space' slide records and the predicted probabilities of the SXGBT(5) model. Although lots of slide records are obviously missing in the flysch zone, the model predicts a correct high potential in the whole region. This is unfortunately not the case for the southern alpine units where slide probability is underestimated due to missing records. Generally, the south-eastern part of Austria including Styria and Carinthia seems to be the most uncertain model area, as probabilities are often between 0.25 and 0.75. The geological area austroalpine unit shows a more disjunctive picture with very high probabilities near slide hot-spots in steep terrain and low probabilities elsewhere.

5 Discussion and Outlook

Although, the difficulty analysis reveals the high complexity of slide susceptibility mapping tasks, non-linear tree-based models show promising results in terms of robustness and validation scores. In combination with characterizing features, they also allow for a correct classification of slide sub processes by assigning them to different leaves of the trees.

The dominance of the DEM in various features and its impact on land slide susceptibility is ubiquitous. Nonetheless, for slide probabilities over 50 % the potentially endangered areas and the imbalance ratio of the data set are reduced and the model does not only reproduce the Austrian topography. Adjusting the scaling factor in the cost function for the minority class results in a trade-off between classification performance and imbalance ratio for probabilities over 50 percent. Put differently, the model misses more slides but also reduces the total area of highly endangered zones if we reduce the weight of the minority class in the cost function.

Besides model properties, the high importance of terrain attributes and the reduction of the spatial resolution through coarsening may also limit the model's performance. On the other hand, the spatial information of the damage data might be of varying quality. Hence, coarsening can reduce the effect of erroneous spatial information in the damage event records. Throughout the study we averaged the terrain derived indices within a 5x5 window, but taking the maximum or minimum could also be an alternative to averaging and might improve the model's performance. The averaged indices show strong linear correlations between themselves but are kept due to possible non linear relations and are accounted for during the feature importance analysis.

Another problem arises as all geological features are qualitative features. The only exception is the distance to the closest geological boundary. The qualitative variables have up to 60 categories and may double the amount of dimensions in the feature space if one-hot-encoding is performed. A hazardous centered geological hierarchy and enhanced geological maps in terms of resolution would be of great value for this type of model approach.

Other feature related problems may also occur related to artifacts based on the event collection process. As an example, streets may act as a barrier layer and therefore force more water to drain into the ground. This could explain the high gain of the model from street distance data. On the other hand, the gain could also stem from the fact that it is necessary to take action if a street is affected by a land slide. Hence, the land slide is registered because the street has to be repaired and insurances have to pay the damage, which is not the case if a landslide occurs somewhere far from infrastructure.

An uncertainty estimation would be the next consequent step to improve the reliability of our model approach and to evaluate the model's predictive quality. One approach could be the use of cross validated bootstrap samples, where new data sets are drawn with replacement from the original data set. Cross validation is then performed on the so created bootstrap samples. Confidence intervals in every grid point can then be estimated with those bootstrap predictions. Due to limited hardware resources during the thesis this task will remain an open issue.

Moreover, time-dependent slide triggering data was not included either, as it increases the total imbalance ratio even further. Nevertheless, a predictive approach demands its inclusion into the feature space. A first step could include only weather data over Austria on slide event days.

In the field of non-linear damage event prediction, tree-based models have a high potential and may contribute to various problems on different scales in climate science. Nevertheless, those data driven approaches depend highly on the underlying data and collecting processes are not yet centralized. Therefore, chances are high that there are more slide records available in different archives. With new generations of high resolution satellite products, the quality of damage data catalogues will improve dramatically and it would be of great value to include spatial information on single damage events in internationally homogeneous catalogues.

Acknowledgements

Thank you:

To mom and dad, for being my lighthouses and anchors, shaping me to become who I am now.

To Katharina, cause it feels like home every day.

To Noah, for all the passionate phone calls and for the ear you lend me when it really matters.

To my sister and brother, for never falling apart no matter how different we are.

To Valerie, for all the memories we share, walking together through a formative time.

Bibliography

Andrew D. Weiss (2001). "Topographic position and landforms analysis." In: *jennessent.com*. URL: http://www.jennessent.com/downloads/TPI-poster-TNC%7B%5C_%7D18x22.pdf (cit. on p. 14).

Ayalew, Lulseged and Hiromitsu Yamagishi (2005). "The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan." In: *Geomorphology* 65.1-2, pp. 15–31. ISSN: 0169-555X. DOI: 10.1016/J.GEOMORPH.2004.06.010. URL: https://www.sciencedirect.com/science/article/abs/pii/S0169555X04001631 (cit. on p. 4).

Ballabio, Cristiano, Panos Panagos, and Luca Monatanarella (2016). "Mapping topsoil physical properties at European scale using the LUCAS database." In: *Geoderma* 261, pp. 110–123. ISSN: 0016-7061. DOI: 10.1016/J.GEODERMA.2015.07.006. URL: https://www.sciencedirect.com/science/article/pii/S0016706115300173 (cit. on p. 15).

Banko, Gebhard et al. (2014). "Land Information System Austria (LISA)." In: Springer, Dordrecht, pp. 237–254. DOI: 10.1007/978-94-007-7969-3_15. URL: http://link.springer.com/10.1007/978-94-007-7969-3%7B%5C_%7D15 (cit. on p. 16).

Batista, Gustave and Diego Furtado Silva (2009). "How k-Nearest Neighbor Parameters Affect its Performance." In: *Journal of Applied Sciences* 14.2, pp. 171–176. ISSN: 18125654. DOI: 10.3923/jas.2014.171.176 (cit. on p. 19).

Batuwita, Rukshan and Vasile Palade (2010). "Efficient resampling methods for training support vector machines with imbalanced datasets." In: *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8. ISBN: 978-1-4244-6916-1. DOI: 10.1109/IJCNN.2010.5596787. URL: http://ieeexplore.ieee.org/document/5596787/ (cit. on p. 3).

Böhner, Jürgen and Thomas Selige (2006). "Spatial prediction of soil attributes using terrain analysis and climate regionalisation." In: *SAGA*

- *Analysis and Modelling Applications* 115.January 2002, pp. 13–27. ISSN: 14712288. DOI: 10.1186/1471-2288-4-5 (cit. on p. 13).

Branco, Paula, Luis Torgo, and Rita Ribeiro (2015). "A Survey of Predictive Modelling under Imbalanced Distributions." In: pp. 1–48. arXiv: 1505.01658. URL: http://arxiv.org/abs/1505.01658 (cit. on pp. iv, v, 2, 3).

Breiman, Leo (1996). "Bagging predictors." In: *Machine Learning* 24.2, pp. 123–140. ISSN: 0885-6125. DOI: 10.1007/BF00058655. URL: http://link.springer.com/10.1007/BF00058655 (cit. on pp. 3, 26).

Breiman, Leo (2001). "Random Forests." In: *Machine Learning* 45.1, pp. 5–32. ISSN: 08856125. DOI: 10.1023/A:1010933404324. URL: http://link.springer.com/10.1023/A:1010933404324 (cit. on pp. 4, 26).

Carrara, A. et al. (1991). "GIS techniques and statistical models in evaluating landslide hazard." In: *Earth Surface Processes and Landforms* 16.5, pp. 427–445. ISSN: 01979337. DOI: 10.1002/esp.3290160505. URL: http://doi.wiley.com/10.1002/esp.3290160505 (cit. on p. 4).

Chawla, N. V. et al. (2002). "SMOTE: Synthetic Minority Over-sampling Technique Nitesh." In: *Machine Learning Group Universite Libre de Bruxelles Belgium* 16.1, pp. 732–735. ISSN: 10769757. DOI: 10.1613/jair.953. arXiv: 1106.1813 (cit. on p. 3).

Chen, Chao, Andy Liaw, and Leo Breiman (2004). "Using Random Forest to Learn Imbalanced Data." In: *Discovery* 1999, pp. 1–12 (cit. on p. 3).

Chen, Tianqi and Carlos Guestrin (2016). "XGBoost." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. New York, New York, USA: ACM Press, pp. 785–794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785. URL: http://dl.acm.org/citation.cfm?doid=2939672.2939785 (cit. on pp. iv, v, 4, 20, 27).

Dahal, Ranjan Kumar et al. (2008). "GIS-based weights-of-evidence modelling of rainfall-induced landslides in small catchments for landslide susceptibility mapping." In: *Environmental Geology* 54.2, pp. 311–324. ISSN: 0943-0105. DOI: 10.1007/s00254-007-0818-3. URL: http://link.springer.com/10.1007/s00254-007-0818-3 (cit. on p. 4).

Davis, Jesse and Mark Goadrich (2006). "The relationship between Precision-Recall and ROC curves." In: *Proceedings of the 23rd international conference on Machine learning - ICML '06*. New York, New York, USA: ACM Press, pp. 233–240. ISBN: 1595933832. DOI: 10.1145/1143844.1143874. URL: http://portal.acm.org/citation.cfm?doid=1143844.1143874 (cit. on p. 29).

EASAC (2018). "Extreme weather events in Europe. Preparing for climate change adaptation: an update on EASAC's 2013 study." In: *Report n.22* March, pp. 1–8. URL: www.easac.eu (cit. on p. 1).

Enigl, Katharina et al. (2019). "Derivation of canonical total-sequences triggering landslides and floodings in complex terrain." In: *Advances in Water Resources*. ISSN: 03091708. DOI: 10.1016/j.advwatres.2019.04.018 (cit. on pp. iv, v, 1, 9).

Estabrooks, Andrew, Taeho Jo, and Nathalie Japkowicz (2004). "A Multiple Resampling Method for Learning from Imbalanced Data Sets." In: *Computational Intelligence* 20.1, pp. 18–36. ISSN: 0824-7935. DOI: 10.1111/j.0824-7935.2004.t01-1-00228.x. URL: http://doi.wiley.com/10.1111/j.0824-7935.2004.t01-1-00228.x (cit. on p. 3).

Friedman, Jerome H. (2001). *Greedy Function Approximation: A Gradient Boosting Machine*. DOI: 10.2307/2699986. URL: https://www.jstor.org/stable/2699986 (cit. on pp. 4, 27).

Gareth, James et al. (2019). *Introduction to Statistical Learning with Applications in R*. Vol. 11. 4, pp. 1–235. ISBN: 9781461471370. DOI: 10.2200/s00899ed1v01y201902mas024 (cit. on pp. 2, 20, 21, 26).

Goetz, J.N. et al. (2015). "Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling." In: *Computers & Geosciences* 81, pp. 1–11. ISSN: 0098-3004. DOI: 10.1016/J.CAGEO.2015.04.007. URL: https://www.sciencedirect.com/science/article/pii/S0098300415000904 (cit. on p. 5).

Guyon, Isabelle (2003). "Design of experiments of the NIPS 2003 variable selection benchmark." In: *NIPS 2003 workshop on feature extraction and feature selection* (cit. on pp. 7, 19).

Guzzetti, Fausto et al. (2005). "Probabilistic landslide hazard assessment at the basin scale." In: *Geomorphology* 72.1-4, pp. 272–299. ISSN: 0169-555X. DOI: 10.1016/J.GEOMORPH.2005.06.002. URL: https://www.sciencedirect.com/science/article/abs/pii/S0169555X05001911 (cit. on p. 4).

Hastie, Trevor, Tibshiran Robert, and Jerome Friedman (2017). "Elements Of Statistic Learning." In: ISSN: 03436993. DOI: 10.1007/b94608. arXiv: arXiv:1011.1669v3 (cit. on pp. 2, 20).

Hido, Shohei, Hisashi Kashima, and Yutaka Takahashi (2009). "Roughly balanced bagging for imbalanced data." In: *Statistical Analysis and Data Mining* 2.5â6, pp. 412–426. ISSN: 19321864. DOI: 10.1002/sam.10061. URL: http://doi.wiley.com/10.1002/sam.10061 (cit. on p. 3).

Lango, Mateusz and Jerzy Stefanowski (2018). "Multi-class and feature selection extensions of Roughly Balanced Bagging for imbalanced data." In: *Journal of Intelligent Information Systems* 50.1, pp. 97–127. ISSN: 15737675. DOI: 10.1007/s10844-017-0446-7 (cit. on p. 3).

Marjanović, Miloš et al. (2011). "Landslide susceptibility assessment using SVM machine learning algorithm." In: *Engineering Geology* 123.3, pp. 225–234. ISSN: 0013-7952. DOI: 10.1016/J.ENGGEO.2011.09.006. URL: https://www.sciencedirect.com/science/article/abs/pii/S0013795211002195 (cit. on p. 4).

Micheletti, Natan et al. (2014). "Machine Learning Feature Selection Methods for Landslide Susceptibility Mapping." In: *Mathematical Geosciences* 46.1, pp. 33–57. ISSN: 18748953. DOI: 10.1007/s11004-013-9511-0. URL: http://link.springer.com/10.1007/s11004-013-9511-0 (cit. on p. 4).

Milly, P. C. D. et al. (2002). "Increasing risk of great floods in a changing climate." In: *Nature* 415.6871, pp. 514–517. ISSN: 0028-0836. DOI: 10.1038/415514a. URL: http://www.nature.com/articles/415514a (cit. on p. 5).

Mosavi, Amir, Pinar Ozturk, and Kwok-wing Chau (2018). "Flood Prediction Using Machine Learning Models: Literature Review." In: *Water* 10.11, p. 1536. ISSN: 2073-4441. DOI: 10.3390/w10111536. URL: http://www.mdpi.com/2073-4441/10/11/1536 (cit. on p. 6).

Napierala, Krystyna and Jerzy Stefanowski (2016). "Types of minority class examples and their influence on learning classifiers from imbalanced data." In: *Journal of Intelligent Information Systems* 46.3, pp. 563–597. ISSN: 0925-9902. DOI: 10.1007/s10844-015-0368-1. URL: http://link.springer.com/10.1007/s10844-015-0368-1 (cit. on p. 18).

Ohlmacher, Gregory C. and John C. Davis (2003). "Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas, USA." In: *Engineering Geology* 69.3-4, pp. 331–343. ISSN: 0013-7952. DOI: 10.1016/S0013-7952(03)00069-3. URL: https://www.sciencedirect.com/science/article/abs/pii/S0013795203000693 (cit. on p. 4).

Panagos, Panos et al. (2012). "European Soil Data Centre: Response to European policy support and public data requirements." In: *Land Use Policy* 29.2, pp. 329–338. ISSN: 0264-8377. DOI: 10.1016/J.LANDUSEPOL.2011.07.003. URL: https://www.sciencedirect.com/science/article/abs/pii/S0264837711000718 (cit. on p. 15).

Petschko, H. et al. (2014). "Assessing the quality of landslide susceptibility maps - Case study Lower Austria." In: *Natural Hazards and Earth System Sciences* 14.1, pp. 95–118. ISSN: 15618633. DOI: 10.5194/nhess-14-95-2014 (cit. on p. 5).

Rechenraum (2014). "Der oe3d Datensatz." In: URL: www.oe3d.at (cit. on p. 13).

Regmi, Netra R., John R. Giardino, and John D. Vitek (2010). "Modeling susceptibility to landslides using the weight of evidence approach: Western Colorado, USA." In: *Geomorphology* 115.1-2, pp. 172–187. ISSN: 0169-555X. DOI: 10.1016/J.GEOMORPH.2009.10.002. URL: https://www.sciencedirect.com/science/article/abs/pii/S0169555X09004279 (cit. on p. 4).

Schapire, Robert E. (2003). "The Boosting Approach to Machine Learning: An Overview." In: Springer, New York, NY, pp. 149–171. DOI: 10.1007/978-0-387-21579-2_9. URL: http://link.springer.com/10.1007/978-0-387-21579-2%7B%5C_%7D9 (cit. on p. 3).

Steger, S. et al. (2017). "The influence of systematically incomplete shallow landslide inventories on statistical susceptibility models and suggestions for improvements." In: *Landslides* 14.5, pp. 1767–1781. ISSN: 16125118. DOI: 10.1007/s10346-017-0820-0 (cit. on p. 5).

Tóth, B. et al. (2015). "New generation of hydraulic pedotransfer functions for Europe." In: *European Journal of Soil Science* 66.1, pp. 226–238. ISSN: 13652389. DOI: 10.1111/ejss.12192 (cit. on p. 15).

Varnes, David J. (1984). *Landslide hazard zonation : a review of principles and practice*. 3. Unesco, p. 63. ISBN: 9231018957. URL: https://trid.trb.org/view/281932 (cit. on p. 15).

Weiss, Gary M. and Foster Provost (2003). "Learning when training data are costly: The effect of class distribution on tree induction." In: *Journal of Artificial Intelligence Research* 19, pp. 315–354. ISSN: 10769757. DOI: 10.1613/jair.1199 (cit. on p. 3).

Wilson, Margaret F.J. et al. (2007). *Multiscale terrain analysis of multibeam bathymetry data for habitat mapping on the continental slope*. Vol. 30. 1-2, pp. 3–35. ISBN: 0149041070129. DOI: 10.1080/01490410701295962 (cit. on p. 14).

Zevenbergen, Lyle W. and Colin R. Thorne (1987). "Quantitative analysis of land surface topography." In: *Earth Surface Processes and Landforms* 12.1, pp. 47–56. ISSN: 01979337. DOI: 10.1002/esp.3290120107. URL: http://doi.wiley.com/10.1002/esp.3290120107 (cit. on p. 13).