Predicting the risk of accidents for downhill skiers

MARCO DALLAGIACOMA
Predicting the risk of accidents for downhill skiers

Marco Dallagiacoma

Master of Science Thesis

ICT Innovation
School of Information and Communication Technology
KTH Royal Institute of Technology
Stockholm, Sweden

15 August 2017

Examiner: Šarūnas Girdzijauskas
Abstract

In recent years, the need for insurance coverage for downhill skiers is becoming increasingly important. The goal of this thesis work is to enable the development of innovative insurance services for skiers. Specifically, this project addresses the problem of estimating the probability for a skier to suffer injuries while skiing.

This problem is addressed by developing and evaluating a number of machine-learning models. The models are trained on data that is commonly available to ski-resorts, namely the history of accesses to ski-lifts, reports of accidents collected by ski-patrols, and weather-related information retrieved from publicly accessible weather stations. Both personal information about skiers and environmental variables are considered to estimate the risk. Additionally, an auxiliary model is developed to estimate the condition of the snow in a ski-resort from past weather data. A number of techniques to deal with the problems related to this task, such as the class imbalance and the calibration of probabilities, are evaluated and compared.

The main contribution of this project is the implementation of machine learning models to predict the probability of accidents for downhill skiers. The obtained models achieve a satisfactory performance at estimating the risk of accidents for skiers, provided that the needed historical data for the target ski-resorts is available. The biggest limitation encountered by this study is related to the relatively low volume and quality of available data, which suggests that there are opportunities for further enhancements if additional (and especially better) data is collected.
Sammanfattning

Under senaste åren har behovet av försäkringsskydd för utförsäkare vuxit sig större och blivit viktigare än någonsin. Målet med detta examensarbete är att möjliggöra utveckling av innovativa försäkringar för skidåkare. Projektet tar specifikt upp problemet med att uppskatta sannolikheten att en skidåkare skadar sig under skidåkning.


Projektets huvudsakliga bidrag består av genomförandet av maskininlärningsmodeller att förutsäga sannolikheten för olyckor för utförsäkning skidåkare. Den erhållna modellen uppnår en tillfredsställande prestanda på uppskatta risken för olyckor där skidåkare är involverade, förutsatt att de historiska uppgifter som behövs för skidorterna är tillgängliga. Den största begränsningen som denna studie har stött på är relaterad till de relativt låga volymer och kvaliteten på tillgänglig data, vilket tyder på att det finns möjligheter för ytterligare förbättringar om ytterligare (och särskilt bättre) data samlas in.
Acknowledgements

This thesis was developed during an internship at Motorialab s.r.l., in the context of a double degree master’s programme offered by EIT (European Institute of Innovation and Technology). I would like to thank Prof. Šarūnas Girdzijauskas for accepting the role of examiner at KTH, Prof. Keijo Heljanko for accepting the role of supervisor at Aalto University, and Amira Soliman El Hosary for being my supervisor at KTH and for her valuable recommendations.

I would like to express my gratitude to Riccardo De Filippi for giving me the opportunity to work on this project, and to the rest of the team at Motorialab (Luca, Ale, Shamar, Andrea) for supporting me during this time. I would also like to thank the MPBA unit of the Bruno Kessler Foundation for their invaluable support. In particular, I would like to thank Andrea Gobbi for his patience and the many useful discussions we had during these months, and Cesare Furlanello for his valuable advice. I also owe my gratitude to Illy for his help in the early phases of the project, and Azra for being a fantastic Swedish translator.

I am also very thankful to all my friends, both in Italy and around the world, for sharing with me those wonderful years.

I want to express my gratitude to my family, in particular to my mom and my dad, for their constant love and moral support. A special thanks goes to my sister Giulia, for her encouragement and for being a perfect example of what can be achieved by studying and working hard. I would also like to thank my grandparents, for their constant and unfailing support for all these years. Finally, I want to thank Elena, for her continuous encouragement and for standing by me in all situations.
# Contents

1 Introduction ...................................................... 1
   1.1 Skiing safety .............................................. 1
   1.2 Insurance for alpine skiers .............................. 1
      1.2.1 Insurance rate-making ............................. 2
   1.3 Research question ......................................... 3
   1.4 Purpose .................................................. 4
   1.5 Contribution ............................................. 4
   1.6 Delimitations ............................................ 4
   1.7 Outline .................................................. 4

2 Related Work .................................................. 7
   2.1 Risk factors .............................................. 7
   2.2 Prediction of skiing injuries ............................ 8
   2.3 Analysis of skiers’ activities .......................... 10

3 Machine Learning Background ................................. 13
   3.1 Missing data ............................................. 13
      3.1.1 EM Imputation ....................................... 14
   3.2 Classification ........................................... 15
      3.2.1 Classification models ............................... 15
         3.2.1.1 Logistic Regression ........................... 15
         3.2.1.2 Random Forest ................................. 16
         3.2.1.3 Gradient Boosted Trees ....................... 17
         3.2.1.4 Feedforward neural networks ................. 18
      3.2.2 Models tuning ....................................... 18
         3.2.2.1 Hyperparameters optimization ................ 18
         3.2.2.2 Feature selection .............................. 19
         3.2.2.3 Cross validation .............................. 19
   3.3 Probability estimation .................................. 20
      3.3.1 Calibration of probabilities ...................... 20
         3.3.1.1 Platt Scaling ................................. 21
3.3.1.2 Isotonic Regression ........................................ 21
3.4 Dealing with unbalanced datasets .................................. 22
  3.4.1 Random Undersampling .............................................. 22
    3.4.1.1 Probability estimation with Random Undersampling .................. 22
  3.4.2 Random Oversampling .............................................. 23
  3.4.3 SMOTE .......................................................... 23
  3.4.4 Balanced Bagging .................................................. 23

4 Methodology ............................................................. 25
  4.1 Problem formalization ................................................ 25
  4.2 Data preparation .................................................... 25
    4.2.1 Classification of snow condition .................................... 27
  4.3 Risk models ........................................................ 28
  4.4 Analysis of behavior of skiers ....................................... 28
  4.5 Evaluation metrics .................................................. 29
    4.5.1 Classification ..................................................... 29
    4.5.2 Probability estimation ........................................... 30
      4.5.2.1 Discrimination .................................................. 30
      4.5.2.2 Calibration ..................................................... 32
  4.6 Tools and frameworks .............................................. 33
    4.6.1 Sklearn .......................................................... 33
    4.6.2 Amelia ........................................................... 33
    4.6.3 Keras .......................................................... 33

5 Feature Engineering ..................................................... 35
  5.1 Available data ...................................................... 35
    5.1.1 Ski-lift runs ..................................................... 35
    5.1.2 Ski accidents .................................................... 37
    5.1.3 Weather ........................................................ 38
  5.2 Obtaining dataset of skiing-sessions .................................. 38
    5.2.1 Length of time slots ............................................ 38
    5.2.2 Combining ski-lift runs and accident reports ..................... 39
  5.3 Estimating the condition of the snow ................................ 40
    5.3.1 Data properties .................................................. 40
    5.3.2 Experiments setup ............................................... 40
    5.3.3 Results .......................................................... 41
  5.4 Imputation of missing data .......................................... 43
## Contents

6 Risk prediction .......................................................... 45
   6.1 Features .................................................................. 45
   6.2 Risk model .............................................................. 46
      6.2.1 Experiments setup ........................................... 46
      6.2.2 Results .......................................................... 47
   6.3 Probability calibration ................................................ 48
      6.3.1 Experiments setup ........................................... 50
      6.3.2 Results .......................................................... 51
   6.4 Features relevance .................................................... 54
   6.5 Applicability to new ski-resorts ................................... 55

7 Analysis of skiers’ activities .............................................. 57
   7.1 Retrieving skipass-id of injured skiers ......................... 58
      7.1.1 Technical approach ......................................... 58
      7.1.2 Experiments and results ................................... 59
   7.2 Behavior and risk ..................................................... 60
      7.2.1 Experiments and results ................................... 61
         7.2.1.1 Tiredness ............................................... 62
         7.2.1.2 Behavior ................................................. 64
         7.2.1.3 Discussion of results ................................. 65

8 Discussion and conclusion ............................................... 67
   8.1 Discussion ............................................................ 67
      8.1.1 Methodology .................................................. 67
      8.1.2 Data preparation .............................................. 68
      8.1.3 Risk model ..................................................... 68
      8.1.4 Analysis of skiers’ activities ............................... 69
      8.1.5 Challenges ..................................................... 69
   8.2 Conclusion ............................................................ 70
   8.3 Future work .......................................................... 71
   8.4 Ethical considerations ............................................... 71

Bibliography ...................................................................... 73

A Details on configuration of models .................................. 77
   A.1 Snow condition model ............................................. 77
   A.2 Risk models .......................................................... 78
Chapter 1

Introduction

1.1 Skiing safety

Downhill skiing is a popular winter sport and a key tourism resource in the Alps. The number of people enjoying downhill skiing every year is estimated to be 200 million worldwide [1].

While skiing is not considered more dangerous than other popular sports, such as football [2], the risk of injuries for skiers is still significant. The risk of accidents is influenced by many variables, including environmental conditions, traffic on the slopes, experience and ability of the skier, etc. Over the last decades the incidence of injuries among skiers followed a downward trend, with the frequency decreasing from approximately 5 to 8 accidents per 1000 skier-days in the 1970s to approximately 2 to 3 accidents per 1000 skier-days in the 2000s [3]. This decline in incidence is mostly related to the evolution of the equipment used by skiers, along with stricter laws and regulations for skiers and generally improved safety conditions in ski resorts.

To increase the safety for skiers, the Bruno Kessler Foundation (FBK) developed SicurSkiWeb, a platform that provides ski resorts with ICT operative tools to collect and analyze data about interventions of ski-patrols and ski-accidents. In ski resorts where this system is used, all the interventions of ski patrols to aid injured skiers are recorded in a spatial database, along with detailed information about the accident and the skier(s) involved. SicurSkiWeb debuted in year 2009, and today it is used by 19 distinct ski areas in the Italian Alps.

1.2 Insurance for alpine skiers

Although the risk of sustaining injuries while skiing has considerably decreased over time, the need for insurance coverage for skiers is becoming more important
in recent years, for a number of reasons. First, the rescue and first aid service, that was once provided for free in most ski resorts, is no longer provided free of charge in an increasing number of ski areas, in an effort to cut operating costs for the ski resorts [4]. Moreover, some regions (e.g., the Piedmont region in Italy [5]) have recently introduced laws that make it mandatory for skiers to be covered by an insurance policy.

In the Italian Alps region, the market of insurance for downhill skiers is currently dominated by traditional insurance services, that provide coverage for a desired period of time (e.g., a week, or the full season) at a fixed price. In alternative, a daily insurance plan can usually be bought for a relatively low cost when buying the skipass (i.e., the card required to access ski-lifts).

This project aims to enable the development of innovative insurance products for skiers. From a broad perspective, the idea is to provide users with a flexible insurance service, providing offers tailored to their risk profile, and providing insurance coverage only for the time they need it, with a pay-per-use approach. The interface for this service could be provided by a mobile application, allowing skiers to buy insurance coverage for the desired period of time before they start to ski. The advantages over current solutions would include both a potentially lower price and a more convenient and interactive interface to the service.

From a technical perspective, the goal of this project is to estimate the risk for a skier to sustain injuries while practicing downhill skiing. The estimation of the risk of accidents could then be used to tailor the insurance service for the user.

### 1.2.1 Insurance rate-making

Insurance is traditionally provided by an insurer, that sells a contract (the insurance policy) to their customers (policyholders) in exchange for money (premium). Basically, an insurance policy is a promise of the insurer to indemnify the customer in case a specific event happens. When an event covered by the insurance contract happens, the customer can make a demand (claim) to the insurer for indemnification according to the insurance policy. The amount of money that the insurer gives as compensation to the claimant is called loss.

In order to be profitable, insurers need to sell insurance plans for a premium that is higher than the expected costs they will sustain. The costs for insurance companies are mainly represented by losses and underwriting costs (i.e., the expenses needed to provide their service, excluding the losses). The process of determining the optimal rate for an insurance policy is called rate-making.

One of the most crucial tasks of rate-making is to determine the pure premium, which is the expected loss associated with an insurance policy. The estimation of losses is traditionally based on two random variables [6]:
### 1.3. Research Question

- **Loss frequency**: the amount of times a loss occurs in a specific period of time. In other words, the probability that a loss happens.

- **Loss severity**: the expected entity of a loss, given that a loss occurred.

A simple method to estimate the pure premium considering these two variables is to use the following equation:

\[ E[l | x] = E[l | y = 1, x] P[y = 1 | x] \]  

(1.1)

where \( x \) represents the profile of the user (and potentially other relevant variables), \( y \) is a binary variable that represents the occurrence of an accident, and \( l \) represents the loss.

With this approach, known as the **frequency/severity method**, the first term (i.e., the severity part) corresponds to a regression problem, while the second term (i.e., the frequency part) can be addressed as a probabilistic classification problem.

This project addresses the problem of estimating the probability of accidents, i.e., the **frequency** term in Equation 1.1.

### 1.3 Research question

The main research question of this thesis is:

*How can we perform a reliable estimation of the probability for skiers to sustain injuries, by relying on data that is commonly available for ski-resorts?*

This thesis addresses this problem by proposing a methodology to clean and aggregate the relevant data, and to use it to train machine-learning models to predict the risk of accidents for skiers. The goal is to perform this estimation on a personal (i.e., per-skier) basis, considering both personal information about skiers and external (e.g., environmental) variables.

Injuries sustained by skiers can vary by type and entity. In this project, skiers are considered as “injured” if they were involved in accidents that required the intervention of ski patrols and first aid services.

The proposed methodology is required to be applicable to new ski resorts with minimal effort, therefore all the data used to train the model should be easily obtainable for new ski-resorts as well. For this reason, this project relies on data that is commonly available for most ski-resorts, namely the reports of skiing accidents (collected by the *SicurSkiWeb* platform), the history of ski-lift runs (retrieved from the ski-lift infrastructure of the ski-resort) and weather-related information obtained from publicly accessible sources.
1.4 Purpose

A reliable estimation of the risk of sustaining injuries would be useful in a number of scenarios. As mentioned above, the main use-case considered in this thesis is the development of innovative insurance solutions for skiers. A practical example of a product that may benefit from this is a service that allows skiers to purchase insurance coverage for a period of time when they start to ski or shortly before, with offers tailored to the estimated risk profile of users.

In addition to the insurance use-case, the estimation of risk of accidents could intuitively be used for other purposes as well, such as to increase awareness and educate skiers about skiing safety.

1.5 Contribution

The main contributions of this thesis are:

1. A machine-learning methodology to predict the probability for skiers to sustain injuries, based on information that is commonly available to ski-resorts; and

2. An assessment of the potential to study the activities and behavior of skiers and to use this information to better estimate their risk of accidents.

1.6 Delimitations

This thesis addresses the problem of estimating the probability of accidents for skiers. While this estimation would mostly be useful in the context of insurance rate-making, this project does not address insurance-specific problems, such as market regulations and other problems specific to the rate-making task.

Moreover, this project is focused on the development of machine-learning models for the prediction of risk for skiers, and it does not aim to develop the full infrastructure needed to deploy it in a real-world scenario (e.g., the API to interface with the system from the client-side, and the infrastructure to retrieve the necessary data in real-time).

1.7 Outline

This thesis is organized as follows.
1.7. OUTLINE

- Chapter 2 introduces the past research that is relevant to this project, in particular in regards to risk factors for skiers and machine-learning techniques to study the risk of accidents.

- Chapter 3 introduces the machine-learning techniques used to develop the risk models and to address some of the challenges relative to this project.

- Chapter 4 describes the methodology employed to develop this project and the metrics used to evaluate the results.

- Chapter 5 details the available data and the work done to obtain the dataset used to train the models.

- Chapter 6 details the experiments performed to develop probabilistic classifiers to predict the risk of accidents for skiers, and evaluates them.

- Chapter 7 discusses the possibility to analyze the activities of skiers (from their history of ski-lift runs) in order to obtain a number of behavior-related metrics that could potentially be used to improve the accuracy of the estimation of risk.

- Finally, Chapter 8 discusses the results obtained, the future work that can be done to improve the model, and some ethical considerations related to this project.
Chapter 2

Related Work

This chapter provides a review of the relevant literature for this project. Specifically, it introduces a number of epidemiological studies on skiing injuries, and studies related to the analysis of data relevant for this project.

2.1 Risk factors

In order to effectively estimate the risk for skiers to sustain injuries, it is first necessary to understand what are the most important risk factors that influence the probability of sustaining injuries. A number of epidemiological studies have been performed to analyze the causes and patterns in skiing-related injuries. This section introduces some of the most relevant studies about risk factors for alpine skiers, and summarizes their results. It is worth to highlight that, as mentioned above, the patterns and incidence of skiing-related injuries have drastically changed over the last decades, therefore the most dated studies may be less relevant to the current situation than recent ones.

A case-control study [7] on skiing accidents during the 1984/1985 season in the Netherlands evaluated a number of personal and environmental risk factors for skiers. The study was based on a case sample of 572 accidents, with a control group of 576 skiers. According to this study, the risk of accidents was lower for people who reported to be moderately rested and for people who reported to have fear of accidents. Similarly, the risk was lower with poor visibility, in the presence of clouds and when the perceived temperature was cold. Conversely, the risk was higher when the slopes contained icy spots.

A more recent case-control study [8] analyzed the injuries sustained by skiers in eight major Norwegian ski-resorts during the 2002 season, in order to evaluate the influence of a number of potential risk factors. The data about injured skiers was collected from reports of ski patrols, while data about the uninjured control
group was obtained by interviewing skiers at the entry of the bottom main ski-lift at each resort. This study was focused on personal variables, such as the age, gender, nationality and skiing ability of skiers. According to the results of this study, the probability of sustaining injuries is higher among beginners, children, adolescents, skiers of non-Nordic nationality and people who practice snowboarding.

Another study [9] performed a survey to study the risk factors focused on young snowboarders. The survey was done on 2745 students participating in winter sport programs organized by Austrian schools, with a mean age of participants of 15 years. Only students who practiced snowboarding at least once were asked to fill the questionnaires. The data collected regarded the demographics, experience level, equipment, snowboard riding habits and associated injuries. Results of this study show that beginners were at a higher risk of accidents. Moreover, the study showed that students who reported previous sports-related injuries were more at risk of sustaining new snowboard-related injuries, suggesting that the attitude toward risk-taking may influence the probability of sustaining injuries for snowboarders. Additionally, the study analyzed how the risk of accidents was affected by the condition of the snow and by the time of the day. Results show that the risk was the highest on hard snow, and the lowest on groomed and deep snow. Finally, according to the study the highest frequency of injuries was observed during the middle of the morning and in the afternoon.

Furthermore, a research [10] analyzed the usage of ski lifts by skiers together with reports of sustained injuries, in order to determine the impact of the traffic (i.e., number of skiers) and of the time of the day on the rate of accidents. According to the results of this research, the time of the day has a fairly important influence on the rate of accidents, with 11-13 and 15-17 being the time slots with the highest rates of accidents. Regarding the traffic on the slopes, the study observed a small relation between the number of skiers in a ski-resort and the rate of accidents, probably caused by the higher probability of colliding with other skiers.

To summarize, a number of studies have been performed to analyze the relation between a number of factors and the risk to sustain injuries for skiers. A number of notable risk factors were identified, including the experience of skiers, the age, time of the day, attitude toward risk, and condition of the snow.

### 2.2 Prediction of skiing injuries

A small number of studies applied data-mining and machine-learning techniques over skiing-related data combined with other relevant information (e.g., weather data) in order to predict the risk of injuries in a ski-day. While none of these
2.2. Prediction of skiing injuries

studies focused on the same task as this thesis work, the problems they address and the approaches they used are relevant for this project.

A recent study [11] proposed a neural-network model to predict the number of severe injuries that occur in a ski-resort each day. The model is trained on a number of time-related (i.e., day of year, day of month, day of week) and weather-related (i.e., minimum temperature, snow depth, precipitations) features, along with the foreseen affluence of skiers in the ski-resort. A relatively simple model was developed, consisting of a Feedforward Neural Network with a single hidden layer of 15 neurons. The data used to train and test the model consisted of 181 samples, one for each day of the 2013-2014 ski season in a Norwegian ski-resort. By running the model on test data, and comparing the obtained results with the true values (i.e., the true number of injured skiers), the study reported a Mean Squared Error of 0.003, which can be considered an excellent result. However, it is worth to highlight that the test dataset consisted of 27 samples only, which could potentially lead to a skewed evaluation of the performance.

Another study [12] proposed different models to predict the daily injury risk for a ski-resort. The objective of the study was to predict two variables: first, whether there will be injuries during a day, and second whether the number of injuries will be higher than average. The estimation of risk was based on variables related to the traffic of skiers, such as the number of skiers in the area and the number of ski-lift runs, and environmental variables, such as the wind speed, the cloudiness and the average temperature. Three different methodologies were employed and compared. First, a data-mining approach was used, training a number of machine-learning models (e.g., decision trees, k-nearest neighbors, etc.) on the data. Then, the results obtained with the data-mining approach were compared with the results obtained by two qualitative multi-attribute models, the first developed manually (i.e., not automatically derived from training data) with the help of field experts, and the second developed with a hybrid approach (defined by the paper as enhanced expert modeling), taking into consideration the results obtained with the data-mining approach when developing the qualitative multi-attribute model. The results obtained by this study show that estimating whether an accident will occur during one day is a difficult task, due to the uncertainty associated with injuries (as they mostly occur by chance). A better accuracy was achieved when predicting whether the number of accidents will be higher than average. In this case, the data-mining approach achieved an accuracy of 81%, while the multi-attribute models achieved an accuracy of 66% for the basic one and 75% for the “enhanced” (i.e., hybrid) one.

The results of those studies show that machine-learning and data-mining techniques can be successfully used to assess the risk of accidents for downhill skiers, achieving better results than models manually created by experts in skiing injuries. Specifically, information about the affluence of skiers and the weather
conditions appear to be relevant for the estimation of risk of accidents. The goal of the discussed studies was to predict the risk of injuries for a ski-resort during a ski-day. The main difference between these studies and this thesis is that this project aims to perform an estimation of the risk of accidents on a personal basis (i.e., per-skier), by also considering personal information about the skiers. In addition, this thesis aims to perform a more granular prediction of the risk, by performing the estimation on shorter periods of time, thus taking into account the change of risk at different times of the day (as discussed in Section 2.1) and the fact that skiers often ski only for a portion of the day (e.g., only the morning).

### 2.3 Analysis of skiers’ activities

Most of the ski-resorts regulate the access to their ski-lifts using skipass cards, hence recording all the movements of skiers through ski-lifts. A small number of studies have been performed to analyze this data in order to study the flow of skiers in a ski-area and their behavior on the slopes. As mentioned in Section 2.1, past literature suggests that the behavior of skiers and their skiing experience may influence their risk of sustaining injuries. Therefore, the ability to analyze the behavior of skiers from data collected by the ski-lifts infrastructure could potentially be useful in order to predict the risk of accidents for skiers more accurately.

A study [10] addressed the problem of estimating the flow of traffic on the slopes by analyzing the data about usage of lifts by skiers. The aim of the project was to build maps of the risk of accidents, by normalizing the number of accidents on each slope with the estimated number of skiers that skied on that slope. Estimating the traffic on the different slopes using only data regarding the usage of ski-lifts is a difficult task, since usually there is not a direct relation between ski-lifts and slopes (in other words, from a ski-lift it is often possible to reach many slopes, and from a slope it is often possible to reach many ski-lifts). To address this problem, this study relied on a set of constraints provided by the manager of the ski-resort, indicating the approximate frequency at which skiers take each possible slope after using a ski-lift. Additional experiments to overcome this limitation were performed, by estimating the most probable slope that a skier took after a ski-lift run by analyzing the time it took for the skier to reach the successive ski-lift. However, this last experiment did not achieve reliable results compared to the constraints-based one, suggesting that it is not possible to reliably determine the path that a skier used to go from a ski-lift to another one simply by analyzing the travel time.

Another study [13] performed an analysis of the skiing traffic focused on the behavior of groups of skiers. The aim of this research was to identify groups of
2.3. Analysis of skiers’ activities

skiers (e.g., small groups, large groups, skiing courses, athletes, etc.) by analyzing their usage of ski-lifts, and to study the relationship between the groups of skiers and their behavior (e.g., their speed).

Those studies show that it is difficult to obtain precise information about the movements and activities of skiers solely from their history of ski-lift runs. However, they suggest that the patterns of usage of ski-lifts (e.g., the travel times) of different skiers may provide interesting information about their behavior.
Chapter 3

Machine Learning Background

This chapter introduces the algorithms and techniques used to develop the predictive models discussed in this thesis. Section 3.1 introduces the problem of missing data and some well-known techniques used to deal with it. Section 3.2 introduces a number of classification models used in this project, along with techniques used to tune their configuration. Section 4.3 provides an overview on probabilistic classification, and it introduces a number of techniques to improve the probabilities estimated by a model. Finally, Section 3.4 discusses the problems caused by class-imbalance, along with a number of techniques used to improve the performance of models in the presence of this problem.

3.1 Missing data

In the field of statistics and machine-learning, it is common to deal with datasets where some information is occasionally missing. Learning from incomplete data can be difficult, and most supervised-learning algorithms cannot deal with missing values, as they need a complete and consistent dataset to be trained.

A number of approaches can be taken to deal with missing data. The simplest solution is to exclude samples with missing information (a.k.a. complete case analysis). Another method, denoted as imputation, consists in replacing the missing data with new values. However, both complete case analysis and data imputation (depending on the imputation technique used) can lead to biased results, and the choice of imputation method can affect the performance of a predictor.

When dealing with missing information, it is first important to understand the nature of the missingness of data. Missing data can be classified as three different types [14]:

- **Missing Completely At Random (MCAR)** when the incomplete rows are
a random subsample of the dataset. In this case, the missingness of a value does not depend on any variable.

- **Missing At Random** (MAR) when the missingness of information is influenced by other observed variables. In this case, incomplete rows do not represent a random subsample of the complete dataset.

- **Missing Not At Random** (MNAR) when the probability that an observation is missing depends on the (potentially missing) value of the observation itself. For example, if in a survey high income people are less likely to report their income than the rest of the population, the “income” variable is considered MNAR.

It is possible to test whether data is not MCAR by using Little’s MCAR test \([15]\), a statistical test where the null hypothesis is that data is missing completely at random. However, there are no deterministic methods to test if data is MAR or MNAR, since the information needed is missing. Generally, data is assumed to be MAR or MNAR after an analysis of the patterns of missingness of data, and by relying on domain knowledge about the data.

The choice of optimal methodology to deal with missing data depends mainly on the nature of missingness of the data. In case of MCAR, complete case analysis (i.e., ignoring incomplete samples) often works well and does not introduce bias in the dataset. However, it results in a smaller dataset, thus potentially leading in the loss of useful information. Other simple approaches to deal with MCAR data are to replace missing values with the mean (or mode, in case of discrete variables), or with a value taken from other (complete) records.

However, those approaches are not optimal when used with MAR or MNAR data, as they do not account for the factors that caused the missingness of information. While there is not an universal method to deal with MNAR data (since information that influences the missingness of data is unavailable), a number of approaches have been proposed to work with MAR data. The general concept is to replace the missing values with new values estimated from the observed values.

### 3.1.1 EM Imputation

A popular method to perform imputation of MAR data is the Expectation-Maximization (EM) algorithm \([16]\). The EM algorithm enables the estimation of parameters of models with incomplete data. In other words, it is a generalization of Maximum Likelihood Estimation (MLE) to the case of latent (incomplete) variables. The parameters estimated with the EM algorithm can then be used to create a regression model to perform the imputation of missing data.
Considering a multivariate normal distribution, and given an initial set of estimated parameters $\hat{\mu}_0$ (mean vector) and $\hat{\Sigma}_0$ (covariance matrix), the EM algorithm computes two iterative steps until convergence:

- Expectation step: for each sample $y_i$, replace missing values $y_{i\text{ miss}}$ with the conditional expectation of the missing data given the observed data and the estimated parameters $\hat{\theta}_k = (\hat{\mu}_k, \hat{\Sigma}_k)$.

- Maximization step: given the dataset obtained in the Expectation step, obtain maximum likelihood (ML) estimates for the new parameters $\hat{\theta}_{k+1} = (\hat{\mu}_{k+1}, \hat{\Sigma}_{k+1})$.

The algorithm stops when the estimated parameters $\theta_{k+1}$ are essentially equal to the previous estimation $\theta_k$.

The number of variables used to perform imputation affects the reliability and computational cost of the EM method, since the number of parameters that EM needs to estimate increases with the number of variables used. J. Graham [17] suggests that, for large datasets ($N > 1000$), the number of variables used for imputation of missing values should not be greater than 100.

### 3.2 Classification

Classification is the process of applying a “class” label to an observation. Some classification models are able to estimate the probability for a sample to belong to each class, in addition to the plain label. These classification models are defined as probabilistic classification models.

This section introduces a number of well-known classification models used in this project, as well as some popular techniques used to tune the configuration of the models.

#### 3.2.1 Classification models

##### 3.2.1.1 Logistic Regression

Logistic regression [18] is a widely used classification model. Given a set of samples $(x_i, y_i)$, where $y_i$ is a dichotomous dependent variables that can be reduced to $y \in \{0, 1\}$, logistic regression predicts the probability of $y_i$ to be positive $P(y_i = 1 \mid x_i)$. Logistic regression defines the natural logarithm of the odds (a.k.a. logit) of an event as:

$$
\ln \left( \frac{P(y_i = 1 \mid x_i)}{1 - P(y_i = 1 \mid x_i)} \right) = b + w^T x_i \quad (3.1)
$$
where \( w \) is a vector of weights \( w \in \mathbb{R}^n \) and \( b \) is the intercept.

From Equation 3.1 it is possible to obtain the probability \( P(y_i = 1|x_i) \) by applying the sigmoid function \( \sigma(x) \) on the natural logarithm of the odds, as follows:

\[
P(y_i = 1 \mid x_i) = \frac{1}{1 + e^{-(b + w^T x_i)}} = \sigma(b + w^T x_i)
\]

The parameters of a Logistic Regression model can be estimated by maximizing the log-likelihood function:

\[
\mathcal{L}(b, w) = \sum y_i \ln P(y_i = 1 \mid x_i) + (1 - y_i) \ln (1 - P(y_i = 1 \mid x_i))
= \sum y_i \ln \sigma(b + w^T x_i) + (1 - y_i) \ln (1 - \sigma(b + w^T x_i))
\]  

(3.2)

In Logistic Regression models, overfitting can be limited by applying either a L1 regularization term (Lasso) or a L2 regularization term (Ridge).

An advantage of using Logistic Regression is that it allows to interpret the results in a fairly transparent way. Since Logistic Regression models estimate the natural logarithm of the odds (logit) of an event, each coefficient \( w_j \) represents the change in the logit for each unit change in the predictor \( x_j \). While interpreting the change in logit can be unintuitive, it is possible to exponentiate the weights vector in order to obtain the contribution of each variable to the odds ratio. This enables a relatively easy interpretation of the contribution of each feature to the probability of \( y \) to be positive.

### 3.2.1.2 Random Forest

Random Forest [19] is an ensemble learning method for classification and regression tasks, based on a multitude of Decision Trees. The idea of Random Forests is to reduce the variance problem that often affects Decision Tree models by combining a number of trees using the bagging technique. Each tree is trained on a bootstrap sample (i.e., random sampling with replacement) of the total dataset. In addition, each tree in the forest uses a subset of the total features to determine the best splits (a.k.a. feature-bagging). This limits the possibility of some strong features to be selected by most (or all) the trees, that would result in correlated trees and thus reduce the benefits of bagging.

Once all the trees in the forest are trained, it is possible to obtain the results of classification by performing majority vote on the classifications done by the single trees. In addition to classification, the trees of the model can emit a probabilistic output, calculated as the fraction of samples of a particular class in the leaf. A Random Forest model can estimate probabilities by averaging the probabilistic output emitted by each tree. Formally, given a set of trees \( t_1, \ldots, t_N \) belonging to a
3.2. CLASSIFICATION

forest, and given a sample $x_i$, it is possible to obtain a probability estimate as:

$$P(y_i = 1 \mid x_i) = \frac{1}{N} \sum_{i=1}^{N} t_i(x_i)$$

Random Forest models provide a good degree of interpretability of their outputs. Given a prediction done by a Random Forest model, it is possible to decompose it in order to obtain the contribution to that result from each feature.

3.2.1.3 Gradient Boosted Trees

Gradient Boosting is an ensemble learning technique for classification and regression tasks. The basic idea is to sequentially fit a number of weak models in order to minimize an arbitrary loss function $\mathcal{L}$. A popular choice for the base estimator is to use Decision Trees with a fixed depth (in order to maintain a low variance).

Given a set of training samples $x$ and a weak base learner $h(x; \theta)$, the output of a Gradient Boosting model is determined as:

$$F(x) = \sum_{m=0}^{M} \beta_m h(x; \theta_m)$$

where $M$ is the number of estimators used in the Gradient Boosted model, and $\theta_m$ is the configuration of the base estimator at the iteration step $m$.

Given an arbitrary differentiable loss function $\mathcal{L}$, the ensemble of models is obtained with an iterative approach. At each iteration step, a weak base estimator is fit on the data, defined as:

$$F_m(x) = F_{m-1}(x) + \beta_m h(x; \theta_m)$$

The parameters $\beta_m$ and $\theta_m$ are obtained by minimizing the loss function:

$$\arg\min_{\beta, \theta} \sum_i \mathcal{L}(y_i, F_{m-1}(x_i) + \beta h(x_i; \theta))$$

For probabilistic classification tasks, a popular choice of loss function is the deviance function. Basically, with this approach the estimators are interpreted as the logit transform:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-2F(x)}}$$

and the loss function is set to the negative log-likelihood:

$$\mathcal{L}(y, F_m(x)) = y \ln P(y = 1 \mid x) + (1 - y) \ln(1 - P(y = 1 \mid x))$$
Friedman [20] proposed a variation of Gradient Boosting where at each step the base estimator is trained on a random sample (without replacement) of the training dataset. According to Friedman, this method improves the efficiency and accuracy of Gradient Boosting, by incorporating randomization into the training procedure of the model. This variation of Gradient Boosting models is denoted as Stochastic Gradient Boosting.

Similarly to Random Forest models, it is possible to decompose the outputs of a Gradient Boosted Trees model into the contributions from each feature.

### 3.2.1.4 Feedforward neural networks

Feedforward neural networks [21] are artificial neural networks where the nodes form a directed graph. The network is composed of an input layer, one or more hidden layers and an output layer. Each layer, except for the input layer, is composed of a number of neurons with a nonlinear activation function, and it is fully connected to the following layer. Each node (neuron) in the network is connected with a weight $w_{ij}$ to every node in the following layer. The output of a neuron with activation function $g(x)$ is defined as:

$$o = g(w^T x + b)$$

where $x$ is the output of neurons from the previous layers.

Some of the most popular choices for the activation function are the sigmoid (logistic) function $\sigma(x) = \frac{1}{1+e^{-x}}$, the hyperbolic tangent function $tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ and the rectifier function $relu(x) = \max(0,x)$.

The parameters of a Feedforward neural network are determined using the backpropagation technique, minimizing a loss function $L$. For binary classification, a popular loss function is the binary cross entropy, defined as:

$$L(\theta) = -\frac{1}{n} \sum_{i=1}^{n} [y_i \ln p_i + (1-y_i)\ln(1 - p_i)]$$

where $i$ indexes the samples and relative labels, and $p_i$ represents the estimated probability of sample $i$ to belong to the positive class.

### 3.2.2 Models tuning

#### 3.2.2.1 Hyperparameters optimization

Hyper-parameters are parameters that are not directly learnt within estimators. The hyperparameters of a model influence its performance, therefore it is necessary to perform a tuning step to find the optimal values. Two well-known techniques for this task are grid-search and random-search.
Grid-search consists in defining a number of possible values for each parameter, and then training the model with all the possible combinations of the considered parameters. The quality of each set of parameters is evaluated by testing the model on a validation set, and evaluating the results with a pre-defined metric (e.g., auROC, accuracy, etc.). The set of parameters that achieves the best performance is then used to train the final model.

Since grid-search performs an exhaustive search (i.e., it tests all the possible combinations of parameters), running it on a large parameters space can take a long time. Random-search [22] addresses this problem by randomly creating \( n \) sets of parameters from pre-defined distributions of values. The \( n \) parameter in random search can be tuned to define a trade-off between accuracy of the optimization of hyperparameters and time required to perform the search.

In this project, grid-search is used to find the optimal parameters when the considered parameters space is limited, and it is thus possible to perform an exhaustive search in a reasonable time. Otherwise, random-search is used.

### 3.2.2.2 Feature selection

Feature selection is the process of selecting a subset of the total features to be used in the learning phase of a model. This process is important for many reasons. First, it simplifies the model by reducing the features space, and it reduces the time required to train a model. Moreover, by removing the less relevant features, it enhances the generalization of the models, reducing the risk of overfitting.

In some models, such as Logistic Regression, it is possible to perform feature selection by applying an L1 penalty term (Lasso), which forces the coefficients of the less relevant features to be set to 0.

Otherwise, there are a large number of techniques that can be used for this purpose. A simple approach is to use statistical tests (e.g., the \( \chi^2 \) test) to select the features that have the strongest relationship with the output variable.

A more complex approach is to recursively remove attributes, evaluating the performance of the model at each step and finally keeping the set of features that achieved the best performance. The evaluation is done with a pre-defined scoring function (e.g., accuracy or auROC) on a validation set. At each round, the variables to be removed are chosen according to their contribution to the prediction of the target attribute (i.e., the variables with the smallest coefficient, or with the lowest importance, depending on the model used).

### 3.2.2.3 Cross validation

In order to perform hyperparameters optimization (and for a number of other tasks as well), it is necessary to use a validation dataset, i.e., a dataset which
is independent both from the training dataset (used to train the model) and the test dataset (used to assess the performance of the final model). Obtaining a validation set from the train set, however, reduces the number of samples that can be used for learning the model.

To address this problem, it is possible to use cross-validation [23]. In the most basic approach, called $k$-fold cross-validation, the training dataset is split into $k$ partitions. Then, for each of the $k$ partitions, a model is trained on the other $k-1$ “folds” and evaluated on the remaining data, which acts as the validation dataset. The results obtained in the $k$ rounds of cross-validation are then combined (e.g., by averaging or voting).

Compared to simply using a validation dataset, cross-validation has the advantage that it does not require to split the training data, but it is also more computationally expensive, as it requires to fit $k$ models instead of 1.

### 3.3 Probability estimation

Probability estimation is the task of estimating a probability distribution over a number of classes for a new observation. When the target output is binary, probability estimation is the task of estimating the probability of a sample to be positive.

All the classifiers introduced in Section 3.2 can provide a probabilistic output in addition to a discrete one. The probabilistic output of a classifier can usually be interpreted as a ranking function (i.e., the output should be higher when the sample is more likely to belong to the target class), but it is not guaranteed to represent reliable probabilities. While some classifiers (e.g., Logistic Regression) generally emit values that can be be interpreted as conditional probability estimates, other classification models often need a successive probability calibration step in order to predict good probabilities.

#### 3.3.1 Calibration of probabilities

As mentioned above, the binary classification models used in this project can provide probabilistic outputs, i.e., they emit a score representing how likely is a sample to belong to the positive class. In order to predict the class label for a sample, it is usually sufficient to apply a threshold on the scores obtained (e.g., at 0.5), and to label as positive all the samples that are assigned a score higher than the threshold. However, this score is not guaranteed to represent a conditional probability.

A good probabilistic classifier should emit scores such that, for example, among the samples to which it assigned a probabilistic score close to 0.7,
3.3. Probability Estimation

approximately 70% actually belong to the positive class. However, some models tend to push the probability estimates away from the margins (i.e., away from 0 and 1), while other models tend to push the predicted probabilities to the margins, closer to 0 and 1.

The ability of a model to produce a good probabilistic output is denoted as calibration of the model, and the process of correcting those distortions in classification models is known as probability calibration. Two well-known techniques to perform probability calibration of classification models are Platt Scaling and Isotonic Regression.

3.3.1.1 Platt Scaling

Platt scaling [24] was first introduced as a method to obtain probability estimates in the context of Support Vector Machine (SVM) models, but it is applicable as a calibration technique to a multitude of other models as well. The approach of Platt scaling is to pass the outputs of a classifier through a sigmoidal function.

More specifically, let $g$ be a classifier, and let $x \in X$ be the inputs for such classifier. Using Platt Scaling, the calibrated probabilities are obtained as

\[
P(y = 1 \mid x) = \frac{1}{1 + \exp(Ag(x) + B)}
\]

where $A$ and $B$ are two parameters learnt by the algorithm. $A$ and $B$ are learnt using Maximum Likelihood Estimation (MLE) from a dedicated validation dataset or, in alternative, using Cross Validation. Gradient descent is used to find the parameters $A$ and $B$ that maximize the log-likelihood:

\[
\sum_i y_i \log P(y_i = 1 \mid x_i) + (1 - y_i) \log (1 - P(y_i = 1 \mid x_i))
\]

3.3.1.2 Isotonic Regression

Platt Scaling works well with some models, but it can be unreliable in other cases. Isotonic Regression [25] was proposed as a more general alternative for probability calibration. Instead of relying on the sigmoid function, it uses a more generic isotonic function. Given the predictions of a model $g$ and the true targets $y_i$, isotonic regression makes the assumption that $y_i = m(g(x_i)) + \epsilon_i$ where $m$ is an isotonic (i.e. monotonically increasing) function. Thus, the goal of isotonic regression is to find the function $\hat{m}$ such that $\hat{m} = \arg \min_z \sum (y_i - z(g(x_i)))^2$

According to A. Niculescu-Mizil et al. [26], Platt scaling is usually more reliable than Isotonic regression when the training data is scarce and when the distortion in probability estimation is sigmoid-shaped, while Isotonic regression tends to be
more powerful in other cases, since it can correct any monotonic distortion and it is not limited to the sigmoidal one.

### 3.4 Dealing with unbalanced datasets

In binary and multi-class classification problems, it is common to work with unbalanced datasets. A dataset is said to be unbalanced when some classes are largely more represented than other ones. Many standard classification models can have difficulties in managing unbalanced datasets, since the fact that a class is less represented than other ones may make it more difficult for the model to generalize the behavior of the minority (i.e., less represented) class. This can result in classifiers that tend to classify all the samples as majority class, or to produce skewed probability estimates [27].

A common approach to address this issue is to obtain a balanced dataset by either under-sampling the majority classes, or by over-sampling the minority ones. The following sections introduce a number of popular methods to deal with unbalanced datasets in classification and probability estimation problems.

#### 3.4.1 Random Undersampling

Random Undersampling [28] consists in reducing the number of samples in the majority classes by removing observations at random. In addition to the fact that it solves the aforementioned problems, this technique has the advantage of speeding up the training phase (as it reduces the size of the training dataset). However, using Random Undersampling can lead to a loss of valuable information regarding the majority class, as a (potentially large) number of samples of majority class are removed from the training dataset. In practice, it has been shown that undersampling is reliable when the minority class has an adequate number of samples.

#### 3.4.1.1 Probability estimation with Random Undersampling

In the context of probability estimation, performing undersampling on the training dataset results in skewed probability estimations, as the samples contained in the undersampled dataset do not represent the real distribution of observations (i.e., the majority class is underrepresented). However, there are methods to deal with this problem and get an adjusted probability estimate from a classifier trained on a randomly undersampled dataset [29].

Consider a classification problem on an unbalanced dataset, where the negative class is overrepresented, and undersampling is performed on the negative class
3.4. DEALING WITH UNBALANCED DATASETS

Figure 3.1: Random undersampling

to reduce the class imbalance. Let $\beta$ be the probability of selecting a sample of the negative class with undersampling. Let $p_s$ be the posterior probability calculated after undersampling is performed and let $p$ be the posterior probability calculated on the original dataset. $p$ can be obtained from $p_s$ as follows:

$$p = \frac{\beta p_s}{\beta p_s - p_s + 1}$$

(3.3)

3.4.2 Random Oversampling

Random Oversampling [28] decreases the class imbalance by increasing the number of samples in the minority classes by replicating them. Compared to random undersampling, random oversampling has the advantage that it does not lose information about the majority classes, but at the same time it brings new problems. First, it increases the size of the training dataset, which translates in longer training times for the classifier. Second, it increases the risk of overfitting the minority class, thus obtaining biased outputs.

3.4.3 SMOTE

Similarly to Random Oversampling, SMOTE [30] reduces the class imbalance in the dataset by introducing new samples of minority class. However, instead of introducing duplicate records, SMOTE interpolates between samples of the same class (from the $k$ nearest neighbors) in order to produce new, synthetic samples.

3.4.4 Balanced Bagging

As mentioned above, Random Undersampling can lead to a loss of valuable information regarding the majority class by removing samples of that class from
the training dataset. To address this problem, Wallace et al. [31] propose an alternative approach based on the bagging [32] technique.

This technique, denoted as *Balanced Bagging*, consists in drawing a number of balanced bootstrap datasets from the full dataset, and training an ensemble of predictors over the different datasets. The results of the predictors are then merged (e.g., by averaging them) to obtain a single output.

In Balanced Bagging models, the calibration of probabilities is performed over each single learner composing the ensemble, instead of doing it once on the full models.

Figure 3.2: Balanced bagging
Chapter 4

Methodology

This chapter provides a formal definition of the problem addressed by this thesis, and it gives a high-level overview of the methodologies used to develop the models and the evaluation metrics used to assess their performance. Details regarding the experiments done are provided in Chapter 5 for tasks related to the preparation of the dataset, and in Chapter 6 for the creation and evaluation of the models.

4.1 Problem formalization

Given a skier profile \( p \in P \), a ski-resort \( r \in R \) and a time period \( t \in T \), the objective of this project is to estimate the probability for skier \( s \) to sustain injuries while skiing in the ski-resort \( r \) during the time period defined by \( t \).

More formally, the goal is to estimate the probability \( P(y = 1 \mid p, r, t) \) where \( y \) is a binary variable with value of 1 if the skier suffers injuries during the time period defined by \( t \), and value of 0 otherwise.

This problem can be approached as a probabilistic classification problem, where a sample represents a skier in a ski-resort for a fixed period of time, and the label indicates whether the skier was involved in a skiing accident during the considered time-slot. In this project, these samples are defined as skiing sessions. Due to the rarity of accidents, the dataset is heavily unbalanced, with negative samples (i.e., skiing sessions with no accidents) being vastly more frequent than positive samples.

4.2 Data preparation

As detailed above, the goal of this project is to estimate the risk for each skier to sustain injuries during time-slots of fixed length. To train models for this purpose, it is first necessary to obtain a dataset representing the activities of skiers during
the considered slots of time. In other words, the goal is to obtain a dataset where each sample represents a skier in a ski-area for a slot of time, with a number of relevant features (e.g., personal info about skier, environmental variables, etc.) and a label indicating whether the skier was involved in an accident during the period of time considered. As mentioned above, in this project such sample is defined as skiing session. Formally, a skiing session is defined as a tuple $(X_i, y_i)$, where $x_1, ..., x_n \in X_i$ are variables relative to the session (e.g., age and gender of the skier, temperature, etc.), and $y_i$ is a binary label indicating whether the skier sustained injuries during the session.

To build this dataset, data from different sources is transformed and combined. An approximation of the number and demography of skiers present at any time in a ski-resort is obtained by analyzing the history of ski-lift runs. Data regarding injuries is collected from reports of accidents from the SicurSkiWeb system. Additionally, other sources of data are used to retrieve additional relevant features, such as weather information.

In practice, the dataset is obtained with the following steps, also represented in Figure 4.1:

1. Combine the history of ski-lift runs and reports of accidents in order to obtain, for each day and time-slot, an approximation of the population of skiers in the ski-resort. In other words, obtain a dataset of skiing sessions by estimating the presence of skiers and the incidence of injuries from the available data.

2. Include additional features, either by processing the available data or by including new data from external sources, with a focus on information relevant for the risk of accidents (as discussed in Section 2.1). More specifically, information about weather (e.g., temperature and precipitations) is retrieved.
from publicly accessible external sources, and data relative to the affluence of skiers is approximated from the number of ski-lift runs. Furthermore, according to literature, a variable that impacts the risk of accidents for skiers is the condition of the snow. Since this information is currently not easily available for many ski-resorts, we develop a simple classification model to estimate it from past weather information.

3. Fill missing values. The personal information of skiers is retrieved from their ski-pass information. However, some types of ski-passes (e.g., promotional tickets) do not hold information about the owner, therefore the age and gender of the skier are occasionally unknown. This problem is addressed by replacing missing values in the process of imputation of missing values.

These three steps produce a dataset that follows the requirements described above, and that enable the training of machine-learning models to predict the risk. In other words, each sample represents a skier in a ski-resort for a specific day and time-slot, and it includes a number of relevant features and a label indicating whether the skier was involved in an accident during the considered period of time.

4.2.1 Classification of snow condition

While most of the features are obtained either by directly including raw data in the dataset or by performing simple transformations of available data (e.g., the affluence of skiers), information regarding the condition of the snow in ski-resorts is not available as easily. Therefore, we estimate this information by developing a model to classify the condition of the snow from past weather data.

The task of classifying the condition of the snow can be approached as a multi-class classification problem. Given a set of possible snow-condition classes \( S \), it is possible to train a classifier \( g : \mathbb{R}^n \rightarrow S \) that predicts the condition of the snow given a set of variables representing the recent history of temperature and precipitations in the considered ski-resort. If the resulting classifier achieves a good enough performance in predicting the condition of the snow, it can be used as an auxiliary model for the model to predict the risk of accidents, in order to account for the condition of the snow in the process of estimating the risk of accidents for a skier.

The data used for this task is collected by the SicurSkiWeb system, which provides an evaluation of the condition of the snow performed by the ski-patrol at each intervention. As introduced in Section 3.2, a number of classification models can be used for this purpose, such as Random Forests, Boosted Trees and Feedforward Neural Networks.
Details regarding the experiments done to build an estimator of the condition of the snow are provided in Section 5.3.

4.3 Risk models

Given the dataset obtained as described above, the problem of estimating the probability for a skier to sustain injuries during a period of time can be approached as a probabilistic binary classification problem. In other words, it is possible to apply probabilistic classification algorithms on the skiing-sessions dataset, in order to predict the probability that a sample (i.e., a skiing session) is labeled as 1 (i.e., an accident happened). For each sample, features contain information about the profile of the skier (e.g., age group, gender) along with other variables that may influence the risk (e.g., traffic, temperature, etc.). With this approach, given a set of samples it is possible to train a classifier $g : \mathbb{R}^n \rightarrow [0, 1]$.

As mentioned above, skiing accidents are rare events, therefore the dataset will contain mostly records of non-injured skiers, with a small minority of records representing injured skiers. As detailed in Section 3.4, standard classification models often perform poorly when trained on very unbalanced datasets, therefore a number of techniques to deal with this problem are compared when training the models. Specifically, both the Random Undersampling and the Balanced Bagging techniques are compared against the model trained with the standard approach.

In addition, as detailed in Section 3.3, the probabilistic output of classifiers is often skewed either towards the margins (i.e., close to 0 and 1) or towards the center, therefore a calibration phase may be useful to improve the quality of predicted probabilities. To address this problem, two well-known techniques are compared for this task: Platt scaling and Isotonic regression.

4.4 Analysis of behavior of skiers

According to literature, a number of factors that influence the risk of accidents for skiers are related to the behavior and skiing ability of skiers. For example, a higher rate of accidents was observed among beginners and people who have little fear of injuries. As mentioned in Section 2.3, past studies show that it is possible to infer some information related to the behavior of skiers by analyzing their travel times from their history of ski-lift runs. These studies were mostly focused on analyzing the congestion of slopes and identifying specific groups of skiers (e.g., courses, large groups of skiers skiing together, etc.), but their results suggest that it is possible to infer useful information about the behavior of skiers by analyzing their movements through ski-lifts.
In the context of this project, an analysis of the activities and behavior of skiers could potentially be useful in order to determine whether this information influences the risk of accidents.

However, there is a major challenge that makes this analysis difficult to perform: the dataset about injuries and the dataset of ski-lift runs use different keys to identify users, therefore they are not “linked”. As a consequence, once skiers are classified according to their behavior, it is not possible to know who of them were involved in an accident, therefore it is impossible to directly assess the relation between the behavior of skiers and their risk of injuries.

For this reason, this thesis addresses the task of analyzing the behavior of skiers separately from the rest of the experiments, as it is considered a proof-of-concept for potential improvements. First, an experimental approach to “link” the two datasets (i.e., to find the skipass-id of each injured skier) is proposed. Then, a number of potential approaches to analyze the behavior of skiers are proposed. The experiments done in this regard are detailed in Chapter 7.

### 4.5 Evaluation metrics

For this thesis work, a number of predictive models are developed, with different purposes. First, during the data-preprocessing phase, a simple model is developed to estimate the condition of the snow given past weather information. This task is a multi-class classification problem, as the objective is to assign a class label indicating the condition of the snow to each observation.

Furthermore, the focus of this thesis is on the risk models to predict the probability for skiers to sustain injuries. This task is a probabilistic classification problem, as the goal is to predict a probability $p \in [0,1]$.

This section details the metrics used to evaluate the predictive models developed in the context of this thesis work.

#### 4.5.1 Classification

The performance of models for classification tasks is measured with the Matthew’s Correlation Coefficient (MCC) [33]. The MCC is a measure of the quality of classification that returns a value in the interval $[-1,1]$ where 1 represents perfect classification, 0 random classification and -1 total disagreement between true labels and predicted labels. In its original form, the Matthew’s Correlation Coefficient works for binary classification only. However, a generalization of MCC for the multi-class case was proposed by J. Gorodnik [34].

The MCC is a popular metric to assess the performance of a classifier, as it provides more information compared to other simpler metrics (e.g., accuracy) and
it works reliably with unbalanced classes [35].

Consider a classification problem on \(S\) samples and \(N\) classes. Consider two functions \(tc, pc : S \rightarrow \{1, ..., N\}\) where \(tc(s)\) indicates the true class of a sample \(s\) and \(pc(s)\) indicates the predicted class of sample \(s\). The confusion matrix is a square matrix \(C \in \mathcal{M}(N \times N, \mathbb{N})\) where \(C_{ij}\) represents the number of elements of true class \(i\) that the classifier classified as class \(j\):

\[
C_{ij} = |\{s \in S : tc(s) = i \land pc(s) = j\}|
\]

Intuitively, elements on the diagonal of the confusion matrix represent samples classified correctly, as the predicted class is the same as the true class.

In the binary case, where the target variable can take only two values, the confusion matrix is \(C = \begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix}\) where \(TP\) indicates the true positives \((tc(s) = pc(s) = 1)\), \(TN\) the true negatives \((tc(s) = pc(s) = 0)\), \(FP\) the false positives \((tc(s) = 0 \land pc(s) = 1)\) and \(FN\) the false negatives \((tc(s) = 1 \land pc(s) = 0)\).

For the binary case, the MCC is defined as:

\[
\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\]

The multi-class generalization of MCC is defined as:

\[
\text{MCC} = \sqrt{\sum_{k,l,m=1}^{N} \frac{\left(C_{kl}C_{ml} - C_{lk}C_{km}\right)^2}{\sum_{k=1}^{N} \left(\sum_{l=1}^{N} C_{lk}\right) \left(\sum_{j=1}^{N} C_{gj}\right) \sum_{k=1}^{N} \left(\sum_{l=1}^{N} C_{kl}\right) \left(\sum_{f=1}^{N} C_{fg}\right)}}
\]

### 4.5.2 Probability estimation

Two performance measures are considered to evaluate the performance of binary probabilistic models: the ability of the model to discriminate between samples of positive and negative class, denoted as discrimination, and the quality of the predicted probabilities (as discussed above), denoted as calibration.

#### 4.5.2.1 Discrimination

The discrimination of models is evaluated using ROC curves and the Area Under ROC (auROC) metric. Consider a binary classification model with a continuous output \(g : \mathbb{R}^n \rightarrow [0, 1]\). A threshold can be set on the outputs of the model to
4.5. **Evaluation Metrics**

discriminate between positive and negative samples: outputs greater than the threshold are labeled as positive, and the others are labeled as negative.

Once a threshold is set, and all the samples are classified accordingly, four quantities can be obtained to represent the results of the classifier:

- True positives (TP): the number of positive samples correctly classified as positive;
- False positives (FP): the number of negative samples wrongly classified as positive;
- True negatives (TN): the number of negative samples correctly classified as negative; and
- False negatives (FN): the number of positive samples wrongly classified as negative.

The ROC and auROC metrics consider two metrics obtained from these quantities: the ratio of positive samples that are correctly classified as positive, the true positive rate ($TPR = \frac{TP}{TP+FN}$), and the ratio of negative samples that are wrongly classified as positive, the false positive rate ($FPR = \frac{FP}{FP+TN}$).

The ROC curve is obtained by considering a 2-dimensional space defined by the $FPR$ as $x$ and the $TPR$ as $y$, setting different threshold values $t$ (ranging from the minimum value emitted by the classifier to the maximum one) and, for each $t$, plotting the point $(FPR_t, TPR_t)$. A curve that gets closer to the upper-left corner represents a better classifier (as it indicates higher $TPR$ and lower $FPR$), while the diagonal represents a random classifier ($TPR = FPR$).

The auROC score is a metric that summarizes the ROC curve. It is defined as the integral of the ROC curve (i.e., it represents the area contained under the curve). The auROC score is a value in the interval $[0, 1]$, with 1 representing a perfect classifier and 0.5 representing a random one. An easy interpretation of the auROC score is that, given a randomly chosen positive sample $s^+$ and a randomly chosen negative sample $s^-$, the auROC score represents the probability that the model will emit a higher score for $s^+$ than for $s^-$, i.e., $\text{auROC} = P(g(s^+) > g(s^-))$.

Contrarily to other metrics such as the “accuracy” (defined as the ratio of correctly classified samples over the total number of samples, i.e., $\frac{TP+TN}{TP+FP+TN+FN}$), the ROC curve and auROC score are reliable even in case of imbalanced classes. However, while they are good metrics to evaluate the ability of a classifier to discriminate between positive and negative samples, they do not provide information about the reliability (i.e., calibration) of the predicted probabilities [36].
4.5.2.2 Calibration

The quality of fit of probability estimates to the observed data is evaluated using the Brier Score [37] and Reliability Diagrams [38].

The Brier score is a popular measure of quality of probability estimations, and it takes into account both the discriminative power and the calibration of models. Specifically, it is defined as the average squared difference between the observed labels and the estimated probabilities. More formally, let $y \in \{0, 1\}$ be the label associated with observation $x$, let $N$ be the number of observations and let $\hat{P}(y_i | x_i)$ be the estimated probability of $x_i$ to have label $y_i$. The Brier Score is defined as

$$BS = \frac{1}{N} \sum_{i=0}^{N} (y_i - \hat{P}(y_i | x_i))^2$$

A lower Brier score indicates a better estimation of probabilities, as the differences between the estimated probabilities and the observed labels are smaller.

However, the Brier score is not easy to interpret, especially when working with unbalanced datasets. The reason is that in imbalanced scenarios the Brier score is dominated by the majority class, therefore it is difficult to assess the calibration of estimated probabilities for samples of minority class.

To address this problem, a modification of the Brier score has been proposed, the Stratified Brier Score [39]. The Stratified Brier Score produces different scores for each class, thus giving a more reliable estimate of the quality of probability estimation in case of unbalanced data. The Stratified Brier Score is defined as

$$BS^+ = \frac{\sum_{y=1}^{N} (y_i - \hat{P}(y_i | x_i))^2}{N_{pos}}$$
$$BS^- = \frac{\sum_{y=0}^{N} (y_i - \hat{P}(y_i | x_i))^2}{N_{neg}}$$

where $BS^+$ represents the score relative to the positive class and $BS^-$ represents the score relative to the negative class.

The Brier Score and the Stratified Brier Score provide a quantitative evaluation of the performance of probabilistic models. On the other hand, reliability diagrams provide a qualitative representation regarding only the calibration of the model. Reliability diagrams are produced by first discretizing the prediction space into a number of bins, and for each bin plotting the mean predicted probability against the observed frequency of positive cases in the bin. Intuitively, the diagonal of the plot represents a perfect calibration, as the predicted probability perfectly matches the observed distribution.
An advantage of reliability diagrams over the Brier score is that, similarly to ROC curves and the auROC score, they can easily be interpreted even in highly class-imbalanced scenarios.

4.6 Tools and frameworks

4.6.1 Sklearn

Sklearn [40] is a popular open source machine learning library for the Python language. It provides efficient implementations for a variety of classification, regression and clustering techniques. In this project, the Sklearn implementation of a number of models is used, including Random Forests, Gradient Boosted Trees and Logistic Regression. Moreover, the Sklearn library is used to perform Hyperparameters tuning and probability calibration.

Some techniques not implemented in Sklearn (e.g., Balanced Bagging and adjustment of probabilities with undersampling) are implemented in this project over the Sklearn API.

4.6.2 Amelia

Amelia [41] is a library for the R programming language that provides utilities to perform imputation of missing data. More specifically, it provides an efficient implementation of the EM imputation algorithm. Furthermore, it adopts a number of additional techniques to improve the quality of results in some specific cases, such as to deal with ordinal or categorical variables (as the “standard” EM imputation algorithm assumes that variables are multivariate normal).

4.6.3 Keras

Keras [42] is a high-level API to develop neural networks. It can run on top of libraries such as Tensorflow and Theano. In this project, Keras is used to implement the Feedforward neural network models.
Chapter 5

Feature Engineering

In order to train the models, it is first necessary to transform and combine the available data into a consistent, suitable dataset. This chapter details the data available, along with the approaches taken to obtain the final dataset, enhance it with additional features and to deal with the problem of missing values.

5.1 Available data

In this thesis work, all the experiments were conducted on data that is commonly available for ski-resorts. More specifically, the data was collected from three sources:

- Ski-lift runs were collected from the ski-lift infrastructure of two ski-resorts, covering a time period of three ski-seasons in total;
- Accident reports were collected from the SicurSkiWeb system; and
- Weather information was retrieved from a set of publicly accessible weather station.

This section details the data used for this thesis work and its properties.

5.1.1 Ski-lift runs

In most of the ski-areas, the access of users at ski-lifts is regulated by turnstiles, that check the validity of the ski-pass (i.e., the access card) of skiers. Thanks to this system, in most of the ski-resorts it is possible to track the movements of each skier across different ski-lift gates during the day.

Each record in this dataset corresponds to a person accessing (and presumably riding) a ski-lift. More specifically, the following information is provided:
Table 5.1: Data available for each skiing season. The number of skiers is calculated on a daily basis, i.e., subjects who skied for more than one day are counted more than once.

<table>
<thead>
<tr>
<th>Ski-resort</th>
<th>Season</th>
<th>Start</th>
<th>End</th>
<th>Skilift trips</th>
<th>Skiers</th>
<th>Accidents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cermis</td>
<td>2012-2013</td>
<td>2012-12-06</td>
<td>2013-04-14</td>
<td>2 884 310</td>
<td>245 186</td>
<td>251</td>
</tr>
<tr>
<td>Cermis</td>
<td>2013-2014</td>
<td>2012-12-03</td>
<td>2014-04-21</td>
<td>2 993 852</td>
<td>248 469</td>
<td>271</td>
</tr>
<tr>
<td>Pinzolo</td>
<td>2012-2013</td>
<td>2012-12-06</td>
<td>2013-03-07</td>
<td>2 539 025</td>
<td>269 147</td>
<td>248</td>
</tr>
</tbody>
</table>

- Time of access at ski lift;
- ID of skipass;
- ID of ski lift and direction (uphill or downhill);
- Age group (child, junior, adult, senior) of the skier. This information is not always available; and
- Gender of the skier. Similarly to the age group, this information is not always available.

For this project, this data is available for a total of three ski-seasons, in two different ski-areas. Both the ski-resorts are located in the Trentino region, in the north Italian Alps. More specifically, this data is provided for the Cermis ski area for seasons 2012-2013 and 2013-2014, and for the Pinzolo ski area for season 2012-2013 only. A summary of the data available for each season is provided in Table 5.1. In total, this dataset contains about 8 million records.

Information about gender and age of skiers is often provided by their skipass. The age of skiers is aggregated into four classes:

- Child: maximum age of 7 years;
- Junior: between 8 and 15 years old;
- Adult: between 16 and 65 years old; and
- Senior: more than 65 years old.

Additionally, some ski-passes provide information about the category of skier who is using them. For example, it is occasionally possible to identify ski instructors, athletes, ski-resort employees and other categories of skiers.
5.1. AVAILABLE DATA

5.1.2 Ski accidents

Data regarding ski accidents is collected from the SicurSkiWeb platform. As mentioned in Section 1.1, SicurSkiWeb is a service developed by the Bruno Kessler Foundation (FBK) to provide ski resorts with technological tools to improve the safety for skiers. Among other functionalities that are not relevant for this project, this system provides ski-resorts with GeoICT tools to collect and analyze data regarding interventions of ski-patrols and accidents of skiers.

SicurSkiWeb debuted in year 2009, and today it is used by 19 distinct ski-areas in the Italian Alps. Currently, it contains more than 20,000 reports of accidents. Information about accidents are filled by the ski patrols (usually members of the State Police) after each intervention.

Specifically, the following information is recorded for each accident:

- Date and time of accident;
- Personal information about the skier(s) involved, including gender, age and country of origin;
- Additional information about the skier, such as equipment used (e.g., ski, snowboard, etc.), property of the equipment (e.g., rented or private), insurance coverage and use of helmet;
- Geographical information (coordinates of accident, slope, difficulty of slope);
- Descriptive information about weather, visibility, wind and snow condition;
- Descriptive information about causes of the accident (e.g., sudden illness, impact with object, etc.);
- Information about the injury (e.g., diagnosis, body location, etc.); and
- Means of transportation used to transfer the injured skier (e.g., toboga, helicopter, lift, etc.).

It is important to notice that, of the more than 20,000 records of accidents contained in the dataset, only 770 happened in a period and location for which data about usage of ski-lifts (as mentioned in the previous section) is available for this project as well. This means that only that subset of this dataset can be used in conjunction with the dataset of ski-lift runs.
5.1.3 Weather

Both the ski-areas considered in this project (Pinzolo and Cermis) are located in the Trentino region in northern Italy. Most of the ski-areas in this region are either provided with a weather station inside the ski-resort, or they are close to a village with a weather station. Those weather stations are managed by MeteoTrentino, and their data is publicly accessible, both for historical records and real-time data. The information available from the weather stations always includes hourly reports of temperature and precipitations, with occasionally additional variables as well.

It is important to notice that the temperature varies depending on altitude and atmospheric pressure, therefore raw data from weather stations cannot be considered completely accurate for an entire ski area, since the altitude can largely change between different points in the same ski area.

5.2 Obtaining dataset of skiing-sessions

The objective is to obtain a dataset of skiing-sessions, with labels indicating whether an accident happened during the considered interval of time. As mentioned above, each sample represents a skier in a ski-resort for a period of one time-slot, and it is labeled positively if the skier sustained injuries during the considered time.

5.2.1 Length of time slots

First, it is necessary to determine the optimal interval of time to consider as unit of time (i.e., time-slot) for the estimation of risk. As mentioned in Section 2.1, previous research indicates that the risk of accidents is higher at certain times of the day. Therefore, it is sensible to consider relatively short units of time, in order to account for the change of risk at different times of the day.

However, the presence of skiers in a ski-resort during each time-slot is estimated from their history of ski-lift runs, and it is therefore an approximated measure. Similarly, the reported times of accidents are collected by the ski-patrols when they aid injured skiers, therefore they may not be completely accurate. For these reasons, using excessively small periods of time as time units for this analysis may result in a noisy dataset. In other words, the unit of time must be a compromise between granularity and practicality.

In this project, time intervals of 1 hour are considered as units of time for the prediction of risk, in order to achieve a compromise on the issues described above. An additional advantage of using time-slots of 1 hour is that they allow a natural
5.2. OBTAINING DATASET OF SKIING-SESSIONS

and easy interpretation of the results (i.e., the estimated risk will represent the hourly risk of accidents).

5.2.2 Combining ski-lift runs and accident reports

Information about the presence of skiers in a ski-resort during each slot of time is inferred from the history of ski-lift runs of each skier. Specifically, the dataset relative to ski-lift runs is aggregated by time-slots of one hour of length. A skier is considered as present in a ski-resort during a slot of time if they did one or more ski-lift runs during the interval of time. Records obtained with this approach are labeled negatively (i.e. no accidents).

Similarly, each record of injured skier from the dataset regarding skiing accidents is considered as a skier present in the ski resort during the time slot when the accident happened, and those samples are labeled as 1 (i.e., an accident happened).

It can be noticed that, with this approach, skiers that sustained injuries are included twice in the combined dataset: first when aggregating data about the usage of ski-lifts (as they probably used a ski-lift shortly before the accident), and then when accident reports are included in the dataset. This problem introduces some noise among the negative samples, since those samples are wrongly classified as “without accidents”, and should thus be removed. The noise introduced by this issue can be considered as negligible, as these “false negative” samples correspond approximately to only 1/5000 of the total dataset. However, removing them from the dataset is fairly easy. Specifically it is sufficient, for each skiing session labeled positively, to identify a session with the same values (except for the label), and to delete it. Due to the redundancy of the dataset, there may be more than one matching samples, as during a time-slot there may be a large number of people with the same properties skiing in a ski-resort. Therefore, it is enough to remove one of the (potentially many) matching samples.

Once this process is completed, the obtained dataset contains a total of 3 394 736 records (one for each skiing session), 770 of positive class and 3 393 966 of negative class.

For each sample, a number of relevant features can be obtained. Date, time and location relative to each skiing-session can be easily obtained when combining the two datasets. Personal information about the skier can be retrieved either from the information contained in their skipass, or from data collected by ski patrols in case of injured skiers. Additionally, weather-related information can be retrieved from weather stations, as discussed in Section 5.1.3. Finally, information about the affluence of skiers in a ski-resort can be estimated by analyzing the number of ski-lift runs during each day.
5.3 Estimating the condition of the snow

According to literature (as detailed in Section 2.1), the condition of the snow on the slopes is a significant factor that influences the risk of accidents for skiers. This variable is not available for the ski-resorts analyzed in this project. However, the SicurSkiWeb system contains a large number of descriptive observations of the condition of the snow at different times, for all the locations where the system is in use. Presuming that the snow condition is affected by the time of the day and by the weather situation in the days preceding the observation, it could be possible to use this dataset, in conjunction with historical weather data, to build a classifier to estimate the condition of the snow at a particular time in a ski-resort.

This section describes the experiments performed and the results obtained to build a predictive model for the condition of the snow.

5.3.1 Data properties

Many reports of intervention in the SicurSkiWeb system are provided with a classification for the condition of the snow. The classification is performed on site by members of ski-patrol teams, and it can take 5 classes: compact, hard, mealy, fresh and wet.

A total of 12,770 observations are available. The distribution of classes is shown in Table 5.2. As can be seen, the different classes are fairly unbalanced, as the majority of records are labeled as “compact” snow.

Weather data is retrieved from publicly accessible weather stations, as mentioned in Section 5.1.3, which provide information regarding temperatures and precipitations.

<table>
<thead>
<tr>
<th>Snow</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compact</td>
<td>8159</td>
</tr>
<tr>
<td>Humid</td>
<td>1465</td>
</tr>
<tr>
<td>Mealy</td>
<td>1202</td>
</tr>
<tr>
<td>Fresh</td>
<td>1010</td>
</tr>
<tr>
<td>Hard</td>
<td>934</td>
</tr>
</tbody>
</table>

Table 5.2: Distribution of observations of the condition of the snow.

5.3.2 Experiments setup

A number of models are trained to estimate the condition of the snow from the history of temperature and precipitations before each observation. More specifically, in order to get a representation of the recent temperature and precipitations
5.3. Estimating the Condition of the Snow

history, and to avoid to deal with an excessively high number of features, the following features are considered (total of 15 features)

- Time of observation (approximated by hour);
- Ski area;
- Minimum, average and maximum temperature for 7 days, 3 days and 1 day preceding the observation;
- Instantaneous temperature; and
- Sum of precipitations for 7 days, 3 days and 1 day before observation;

The dataset is split into training (70%) and testing (30%) datasets. Since observations from the same day and ski area have very similar features (and mostly similar labels as well), the split is performed on the date of the observation. In other words, two observations from the same day must be in the same set (either test or training). This prevents overlapping training and test sets, which could lead to optimistic results in the evaluation of the model.

Three different classification models are trained and compared for this task: Random Forests, Gradient boosted Trees and Feedforward Neural Networks. Due to the class imbalance in the datasets, the classification models are trained and evaluated both using the original dataset, and the dataset balanced using the SMOTE oversampling technique.

Since feedforward neural networks are sensitive to features scaling, the data is normalized before it is used to train and evaluate the neural network. Specifically, Z-score normalization is performed in order to obtain, for each variable, values with a mean of 0 and standard deviation of 1 (i.e., the properties of a standard normal distribution). With this technique, the z-scores of the samples are obtained as:

$$z = \frac{x - \mu}{\sigma}$$

where $\mu$ is the mean of the values of a variable, and $\sigma$ is the standard deviation from the mean.

The process to obtain the optimal configuration for each model, and the obtained sets of parameters, is detailed in Appendix A.1.

5.3.3 Results

A summary of the results obtained with different classification models, both on the original dataset and the dataset balanced using the SMOTE technique, is provided in Table 5.3. As can be noticed, the use of SMOTE oversampling consistently
CHAPTER 5. FEATURE ENGINEERING

Table 5.3: Results obtained on the test dataset by three different classification models, trained both on the original dataset and on the dataset rebalanced using SMOTE. Results are evaluated with the Matthew’s Correlation Coefficient (MCC).

<table>
<thead>
<tr>
<th>Model</th>
<th>MCC</th>
<th>MCC (SMOTE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>0.374</td>
<td>0.418</td>
</tr>
<tr>
<td>Boosted Trees</td>
<td>0.376</td>
<td>0.392</td>
</tr>
<tr>
<td>Feedforward NN</td>
<td>0.345</td>
<td>0.378</td>
</tr>
</tbody>
</table>

Figure 5.1: Normalized confusion matrix for the predicted condition of the snow on test data.

resulted in a better performance of the model. The best results, in terms of MCC score, on the test dataset are obtained by the Random Forest classifier trained on the dataset balanced using the SMOTE technique, that achieves a MCC of 0.418. As detailed in Section 4.5.1, a naïve classifier that always predicts the most frequent class (in this case, that would be "compact snow") would achieve an MCC score of exactly 0, and similarly a random classifier (i.e., emitting a random class label at each prediction) would tend to have an MCC score of approximately 0. It is worth to highlight that there may be some similarity between some classes of snow condition, and this could have impacted on the frequency of misclassifications.

Fig. 5.1 shows the normalized confusion matrix for the predicted condition of the snow. As can be seen, the majority of errors consist either of non-compact snow incorrectly classified as compact snow, or misclassifications between mealy and fresh snow. An hypothesis for the reason why non-compact snow is often wrongly classified as compact snow could be that the process of grooming of
slopes may be effective at keeping the snow compact even in unfavourable conditions, that would otherwise lead to other snow conditions. Moreover, the frequent misclassifications between mealy and fresh snow suggest that there may be similarities between these two snow conditions and between the weather conditions that tend to generate them.

Even though the model is not perfectly accurate at predicting the condition of the snow, its probabilistic output (i.e., the probability of the snow to belong to each of the classes, as estimated by the model) can be used by the risk model as an indication of the properties of the snow at a particular time.

5.4 Imputation of missing data

As mentioned above, personal data about skiers is not always known. When analyzing the ski-lift runs of skiers, personal information is retrieved from their ski-pass, and the amount of information available on a ski-pass may depend on a number of factors.

In the datasets available to this project, the age-class of skiers is known in 85% of cases, while their gender is known only in 37% of cases. Little’s MCAR test rejected the null hypothesis with a p-value < 0.01, which indicates that data is not Missing Completely At Random (MCAR). Further analysis shows that the probability that gender information is missing is largely influenced by the age of the subject. Adults show the highest availability of gender information (51%), while other age groups have a much lower rate of available data, i.e., 30% for children, 1% for juniors and 1% for seniors. Regarding the “age” variable, no significant patterns of missingness are found.

As mentioned, the availability of information about age and gender of skiers mostly depends on the type of skipass used by the subject. In particular, “special” types of skipass often do not provide personal information about the skiers who used them. Special ski-passes include free tickets, event-specific skipasses and various types of promotional tickets.

Assuming that the distribution of age and gender does not significantly change between regular ski-passes and “special” ones, the data can be considered as Missing At Random (MAR), therefore the problem of missing values can be addressed using imputation techniques for MAR data.

Intuitively, it is not sensible to impute missing gender information for classes where only 30% (or less) of records are complete, as most of the data would then be artificial, and therefore the resulting dataset could potentially be biased. Considering gender information about adults, however, the rate at which data is missing is still relatively high (49%), but given the large size of the dataset and the redundancy of samples, imputation of missing data can still be considered a
valid option, under the assumption that data is MAR. For these reasons, after the
imputation process is complete the gender variable is considered for adults only,
therefore obtaining a single variable representing the demographic profile of users
that can take five values: child, junior, adult-male, adult-female and senior.

Regarding data about the age of skiers, since the rate of missingness is
relatively low (15%), imputation of missing data can be considered a good option
under the MAR assumption.

Imputation is performed with the Expectation Maximization (EM) technique.
Since the dimensionality of the dataset is relatively low (less than 50 features),
and the size of the dataset is large (more than 3 million records), all the available
features are used for imputation, in order to account for every possible relation
between features and the missing values.

It is worth to highlight that, while it is sensible to assume data to be MAR in
this case, this cannot be verified reliably. If this assumption is false, then some
bias could be introduced in the data. The only reliable approach to solve this
problem, in order to be completely sure that no bias is introduced during this step,
would be to retrieve better (i.e., more complete) data.
Chapter 6

Risk prediction

This chapter details the experiments done to train and evaluate the models to predict the probability of accidents. First, it introduces the features used to train the models. Secondly, it discusses and compares the different approaches used to obtain the risk model. Additional experiments to assess the relevance of different types of features, and to evaluate the suitability of the models for particular use-cases, are also performed.

6.1 Features

The dataset used to train the risk model is obtained with the process described in Section 5. A total of 3,394,736 records are available, 770 of positive class and 3,393,966 of negative class. Each record represents a skiing session, i.e., a skier doing activities in a ski resort during a time slot of one hour. For each sample, the following features are considered:

- Demographic profile of the skier, defined as a single variable that can take 5 values: child, junior, adult-male, adult-female and senior.
- Location (i.e., ski-resort).
- Time slot. Since data is aggregated by time slots of 1 hour, it is simply defined as the hour, i.e., ‘9’ represents the time slot from 9am to 10am.
- Day of the week, e.g., Monday, Tuesday, etc.
- Season level (e.g., high season, mid-season, low season), as defined by the ski-resort administration.
• Temperature, both the instantaneous one and the historic trends. In this project, the average temperature for 1 day, 3 days and 7 days prior to the record are used, along with the instantaneous temperature.

• Precipitation history. In this project, we consider the sum of precipitations for time windows of 1 day, 3 days and 7 days prior the observation.

• Estimated condition of the snow, obtained as the probability scores emitted by the classifier detailed in Sec. 5.3.

• Daily affluence, inferred from the usage of ski-lifts. In order to normalize the affluence information in relation to the size of the ski-resort, this feature takes a value between 1 and 4, representing the quartile of the historical daily affluence (for the same ski-resort) where the daily affluence falls. For example, if the affluence for one day is higher than the historical median affluence for the ski-resort and lower than the 3rd quartile (i.e., the 75th percentile), this feature takes a value of 3.

Categorical features are encoded using one-hot encoding (i.e., as an array of binary bits, where the bit corresponding to the category is set to 1 and the others are set to 0).

6.2 Risk model

As detailed in Section 4.1, the goal of the model is to estimate the probability for a skier to be involved in an accident during a time-slot of one hour. This problem is approached as a binary probabilistic classification task. A number of classification models are trained and compared for this task. The ability of models to discriminate between samples more at risk of accidents and samples less at risk is evaluated with the ROC and auROC metrics.

6.2.1 Experiments setup

Four classification models are compared: Logistic Regression, Random Forests, Gradient Boosted Trees and Feedforward Neural Networks. The Feedforward Neural network is implemented using the Keras framework. For the other models, the implementation provided by the Sklearn library is used.

The dataset is split into training (70%) and test (30%) datasets, performing the split on the day of the skiing session (i.e., ensuring that there cannot be samples from the same day in both the training and testing dataset, for the reasons mentioned in Section 5.3.2). Since the dataset is heavily unbalanced between
positive and negative classes, a number of techniques to deal with this problem are evaluated. Specifically, the following approaches are evaluated:

- No balancing, therefore training a regular model on the full dataset;
- Random Undersampling with a final positive-to-negative ratio of 1, i.e., resulting in a dataset where the number of positive samples is the same as the number of negative samples;
- Random Undersampling with a ratio of $\frac{1}{10}$, resulting in a dataset with a reduced imbalance, (in this case, ten negative samples for each positive one); and
- Balanced Bagging, training an ensemble of probabilistic classifiers over datasets balanced (with ratio of 1) with Random Undersampling.

Even though the dimensionality of the used data is not particularly high, feature selection is performed in order to exclude the less relevant variables from the model. Feature selection is performed using $L_1$ regularization in the case of Logistic Regression, while for Random Forest and Gradient Boosted Trees it is performed via Recursive Feature Elimination with a 5-fold cross-validation, using a Random Forest model to assess the importance of features.

As mentioned in the previous chapter, feedforward neural networks are sensitive to feature scaling, therefore the data used to train and evaluate the neural network model is first normalized using the Z-score standardization technique for each feature.

### 6.2.2 Results

The discriminative power of models is evaluated with the auROC score. A summary of the performance obtained using different models and techniques to deal with class imbalance is provided in Table 6.1. It is worth to remember that a naïve model predicting always the same probability (regardless of the value chosen) would achieve an auROC score of 0.5, and similarly a model predicting always a random probability would achieve an auROC score of approximately 0.5.

It can be noticed that performing Random Undersampling does not improve the performance over the model trained on the full dataset. Specifically, using Random Undersampling to bring the imbalance ratio to $\frac{1}{10}$ (i.e., ten negative samples for each positive sample) leads to similar performance as the regular model trained on the full dataset. However, using Random Undersampling to bring the imbalance ratio to 1. (i.e., obtaining a fully balanced dataset) results
### 6.3 Probability calibration

Up to this point, models were evaluated in terms of discriminative power, i.e., their ability to distinguish between more risky and less risky scenarios. As detailed in Section 3.3, the scores emitted by those models do not always represent good probability estimates, and they may be skewed. Therefore, a successive calibration phase may be necessary in order to produce well-calibrated probabilities.

This section compares the quality of probability estimations of the models considered, along with the results obtained by performing a calibration process on them, either using Platt scaling or Isotonic regression.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Feedforward NN</td>
<td>0.580</td>
<td>0.586</td>
<td>0.619</td>
<td>0.627</td>
</tr>
<tr>
<td>Gradient Boosted Trees</td>
<td>0.677</td>
<td>0.605</td>
<td>0.681</td>
<td>0.689</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.622</td>
<td>0.580</td>
<td>0.615</td>
<td>0.621</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.667</td>
<td>0.605</td>
<td>0.668</td>
<td>0.671</td>
</tr>
</tbody>
</table>

Table 6.1: Comparison of auROC scores obtained by the four classification models, using different techniques to deal with the class imbalance.

In generally bad performances, probably due to the relatively small number of samples of positive class, which results in an excessive loss of information when performing undersampling. Conversely, the Balanced Bagging technique generally leads to a slight improvement of the performance of models.

Regarding the classification algorithms, overall the tree-based models (i.e., Gradient Boosted Trees and Random Forest) achieve the best performances, while Logistic Regression performs poorly in comparison. This indicates that there may be non-linear relations that the Logistic Regression model, being a linear model, cannot detect. An interesting result is the poor performance of Feedforward Neural Networks, which reported results similar to the Logistic Regression models. This may be caused by the relatively low dimensionality and volume of the dataset used, as neural-networks are known to be well-suited especially for large-volume and high-dimensionality data.

A visualization of the ROC curves obtained by the models trained using the Balanced Bagging technique is shown in Figure 6.1.
Figure 6.1: ROC curves obtained by the four classification models using the Balanced Bagging technique to deal with class imbalance.
Table 6.2: Comparison of the performance of models relative to the estimation of probabilities. \( BS^+ \) denotes the Stratified Brier score for the positive class, while \( BS^- \) denotes the Stratified Brier score for the negative class.

### 6.3.1 Experiments setup

For this experiment, the four classification models are configured with the same hyperparameters used in the previous experiment. To improve the performance under class imbalance, the Balanced Bagging technique is used, as it is the method that achieved the best discriminative performance in previous experiments. Since the Balanced Bagging technique performs Random Undersampling on the training dataset, the predicted probabilities are adjusted using Equation 3.3, in order to account for the change in class distribution between the original dataset and the randomly undersampled one.

The probabilities estimated by the “raw” models are compared with the probabilities estimated by the same models after probability calibration is performed on them. The calibration techniques considered are Platt scaling and Isotonic regression, and they are both applied with a 5-fold cross validation.

The quality of probability estimations is evaluated with the Brier score. Due to the high class imbalance, both the standard Brier score and the Stratified Brier score are considered.

Additionally, a visualization of the calibration of probability estimations is provided with reliability diagrams. Since the test dataset is relatively small, the size of probability bins in the diagrams is not fixed, but it varies to ensure that each bin contains a sufficient number of values to make a reliable estimation of the empirical probability. To achieve this, the prediction space is divided in \( k \) bins using the \((k - 1)\)-quantiles of estimated probabilities as delimiters, instead of dividing it into \( k \) bins of equal size.
6.3. Probability Calibration

Figure 6.2: Reliability diagram for the classification models trained using the Balanced Bagging technique. The diagonal line represents the best possible calibration.

6.3.2 Results

Table 6.2 shows the comparison of the Brier score and the Stratified Brier score obtained by the four classification models, along with the results obtained by applying probability calibration techniques (i.e., Platt scaling and Isotonic regression).

As can be seen, calibration techniques slightly improve the quality of probabilistic outputs (as measured by the Brier score) in many cases. However, the best result in this regard (i.e., lowest Brier score) is surprisingly obtained when no calibration technique is used, by the Gradient Boosted Trees model.

The Stratified Brier scores \(BS^+\) and \(BS^-\) suggest that the models calibrated using Isotonic regression tend to predict better probabilities for the positive class than models calibrated using Platt scaling, and conversely using Platt scaling generally results in better probabilities for the negative class, compared to Isotonic regression. The discrimination power of models (i.e., auROC) is generally not
Figure 6.3: Reliability diagram for the classification models trained using the Balanced Bagging technique, with probabilities calibrated with Platt scaling.

Figure 6.4: Reliability diagram for the classification models trained using the Balanced Bagging technique, with probabilities calibrated with Isotonic regression.
6.3. Probability calibration

significantly affected by the use of calibration techniques.

The interpretation of Brier scores can be unintuitive, since most of the scores are either very close to 0 (for Brier score and \( BS^- \)) or to 1 (for \( BS^+ \)). The reason is that skiing accidents are very rare events, therefore the predicted probabilities are always fairly low. As a consequence, the error for the negative class tends to be close to 0, while the error for the positive class tends to be close to 1. It can be useful to consider that a naïve model always predicting a probability of 0 would obtain a Brier Score of 0.000231049, while a model always predicting the total ratio of accidents (i.e., the mean probability of accidents) would obtain a Brier Score of 0.000230995137.

A more intuitive approach to evaluate the quality of probabilistic outputs is to observe the reliability plots. As mentioned above, reliability plots only measure the calibration of models and therefore, contrarily to the Brier score, they are not affected by the auROC score of a model.

Figure 6.2 shows the reliability diagrams relative to the models when no calibration technique is used. As can be seen, some models tend to concentrate their probability estimations around the median point, thus overestimating the probability of the less frequent events and underestimating the probability of more frequent ones. This behavior is particularly noticeable for the Random Forest model, and to a lesser extent for the Feedforward Neural Network. On the other hand, the Gradient Boosted Trees model achieves an apparently good calibration compared to other models.

Figure 6.3 and Figure 6.4 show the reliability diagrams obtained by the same models, after a probability calibration process is performed on them (Platt scaling and Isotonic regression, respectively). Platt scaling appears to be overall fairly ineffective at improving the calibration of models. In fact, some models (e.g., Gradient Boosted Trees) show a worse calibration when this technique is used. Conversely, Isotonic regression seems to be more effective, as the resulting models seem to be better calibrated, in particular for lower probabilities (i.e., the left-half of the diagram).

Surprisingly, the Gradient Boosted Trees model appears to have a better calibration when no calibration technique is used. Since Gradient Boosted Trees (with no calibration) is also the model that achieved the best performance in terms of discrimination (auROC = 0.689), this model can be considered the most suitable for this task among the analyzed models.

The low effectiveness of calibration techniques may be imputable to the relatively small volume of records of positive class, which (in the cross-validation process) results in small validation datasets used for calibration of the models, hence producing unstable results. Another important consideration is that reliability diagrams need a sufficient number of positive samples in the test dataset in order to provide stable estimates of the empirical probability. Therefore, a
larger dataset would enable both a better calibration of models and a more reliable evaluation of the quality of predicted probabilities.

6.4 Features relevance

In order to understand what factors influence the risk of accidents for skiers the most, we perform an evaluation of the importance of different types of features. To obtain meaningful and interpretable results, this evaluation is done by grouping the features in a number of clusters according to their nature, in order to minimize problems caused by correlation between features.

Specifically, the features detailed in Section 6.1 can be grouped into four clusters, as follows:

- Environment: features relative to temperature and precipitations, along with the estimated condition of the snow;
- Skier info: demographic profile of skier, i.e., gender and age group;
- Location and affluence: ski-resort and affluence of skiers;
- Date and time: day of week, season level, time of the day (i.e., time slot).

One of the most popular approaches to estimate the relevance of features for binary probabilistic classification problems is to fit a logistic regression and to evaluate the magnitude of the normalized coefficients relative to each variable (or the average among a group of features). However, previous results suggest that some of the variables are correlated with the risk of accidents by a non-linear relationship. For this reason, we use a Random Forest model to assess the importance of features in this context, as it is able to detect non-linear relations. The model is configured with 100 trees as base learners, using 6 features for each tree split.

The importance of single features in the Random Forest model is assessed with the Gini Importance (a.k.a. Mean Decrease Impurity) [43], which is defined as the total decrease in node impurity weighted by the proportion of samples reaching that node, averaged over all trees in the forest. The importance of each cluster of features is determined as the sum of the normalized importance of single features.

A visualization of the resulting importance scores is shown in Fig. 6.5. As can be seen, the types of feature that seem to influence the risk of accidents the most are related to the environment and to the skiers themselves, while variables related to date, time, location and affluence appear to be slightly less important, but still relevant.
6.5. Applicability to New Ski-resorts

Ideally, the model should be trained on past data collected in the same ski-resort in which the model will later be used. However, the data necessary to train the model may be unavailable for a ski-resort, therefore it may be desirable to use a model trained on data from another ski-resort in that case. To understand how a model trained on a ski-area would perform when applied on another ski-area, it is possible to simulate this scenario by splitting the full dataset into training and test dataset by keeping samples from each ski-area in different datasets (either the

Figure 6.5: Relevance of the different clusters of features for the prediction of risk, obtained from a Random Forest model.

However, it is important to notice that these results may be partially skewed by the presence of correlations among features belonging to different clusters. If two variables are correlated, part of the information they hold is common among them. The Random Forest model considers this information as relevant only the first time it analyzes it, hence only the first of the variables to be considered during the tree-building process is considered as “important” due to the information that is common among them.

For example, the date and time variables may intuitively correlated with both the environmental variables (i.e., the temperature vastly changes between different days and times) and the affluence.
CHAPTER 6. RISK PREDICTION

<table>
<thead>
<tr>
<th>Classification model</th>
<th>auROC</th>
<th>Brier Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient Boosted Trees</td>
<td>0.639</td>
<td>0.000219932</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.623</td>
<td>0.000219923</td>
</tr>
</tbody>
</table>

Table 6.3: Performance of the models trained on a ski-area (Cermis) and evaluated on a different one (Pinzolo).

As detailed in Chapter 5.1, the data available for this project covers a total of three ski seasons in two ski-resorts:

- Pinzolo 2012-2013;
- Cermis 2012-2013; and

It is therefore possible to simulate the described scenario by training the model on data from the Cermis ski-resort and evaluating it on data from the Pinzolo ski-resort. The two models that achieved the best results in previous experiments (i.e., Random Forest with Isotonic Regression and Gradient Boosted Trees, both applied using the Balanced Bagging technique) are evaluated using this approach.

Results are reported in Table 6.3. It is important to notice that the Brier Score can only be used to compare models evaluated on the same test dataset, as its interpretation changes depending on the prevalence of positive samples in the test dataset. As can be seen, the discriminative performance of models is significantly reduced when applied to “new” ski-resorts. An interesting outcome is that, while the Random Forest model performs worse than the Gradient Boosted Trees model in terms of discrimination, it achieves a slightly better Brier score.

The results obtained may be considered pessimistic, as the data available for this project are relatively limited, and therefore it is possible that the model would be able to better generalize to other ski-resorts if more training data was available. However, they suggest that the risk patterns are different among different ski-resorts, therefore in order apply the model to a new ski-area with acceptable results it is first necessary to collect the needed data from the target ski-area.
Chapter 7

Analysis of skiers’ activities

The risk models presented above rely only on demographic information of skiers (i.e., age and gender) and on external environmental variables in order to determine the risk of accidents. As discussed in Section 2.1, research literature indicates that the risk of accidents for skiers is also influenced by factors related to their activities (e.g., level of tiredness) and behavior (e.g., attitude towards risk, experience, etc.). Therefore, an analysis of the activities and the behavior of skiers could potentially be useful to obtain a more accurate estimation of their risk of accidents.

The most effective approach to study the activities of skiers would be to track their movements with a GPS device and potentially other sensors. However, this would require users to activate a GPS tracker and to use it during all their skiing experience, which could be impractical, especially for the less technically proficient users. A more practical, albeit less accurate, approach to study the activities of skiers is to analyze their history of ski-lift runs. As mentioned in Section 2.3, this approach was used in past studies to estimate the distribution of traffic on the slopes and to identify different groups of skiers in a ski-resort, and it could potentially be applied in the context of risk estimation as well.

In order to study the relation between the activities of skiers and their risk of accidents, it is necessary to associate this data with the information about accidents. This would be trivial if reports of accidents included the skipass-ID of injured skiers, but unfortunately this does not currently happen. Therefore, given a skier (identified by their skipass-ID) and their history of ski-lift runs, it is currently not possible to know whether they sustained injuries.

This section details the experiments done to study the possibility to analyze the activities and the behavior of skiers with the purpose of predicting their risk of accidents. First, it discusses an experiment performed to address the aforementioned problem, with the objective of identifying the skipass-ID of injured skiers by searching for possible candidates among people who, according
their history of ski-lift runs, could have been near to the place of an accident at the time when the accident happened. Secondly, it introduces a possible approach to study the activities and behavior of skiers in the context of estimation of risk.

7.1 Retrieving skipass-id of injured skiers

The ability to identify the activities of injured skiers is a needed precondition in order to effectively study the relation between behavior of skiers and risk of accidents.

To address this problem, an attempt to retrieve the history of ski-lift runs of injured skiers (i.e., identifying their skipass-ID) is performed. The general idea is to analyze the history of ski-lift runs in order to identify the skiers that could possibly correspond to the skiers who were involved in accidents. More specifically, a skier is considered a candidate for an accident if they used a ski-lift that leads to the slope of the accident shortly before the time of the accident (during a defined window of time), and if their profile (i.e., age and gender) matches the profile of the injured skier according to the accident report.

When accidents happen, there is often a large number of skiers on the slope of the accident, and many of those skiers could match the profile of the injured person. Therefore, finding a single matching profile among them would be impossible most of the times.

To restrict the number of candidates relative to each accident, it is possible to assume that injured skiers do not generally access any other lift after the accident happens. In this case, it would be possible to consider only skiers that used a ski-lift that leads to the slope of the accident and did not do any other ski-lift run afterwards. If correct, this assumption would be particularly convenient, as in general the last ski-lift runs of skiers are concentrated in a small subset of ski-lifts (intuitively, the lifts that leave the ski resort and lead to parking lots). As a consequence, the number of skiers having their last access to a “regular” (i.e. not part of the mentioned subset) ski-lift would be relatively small, thus possibly making it easier to identify a single candidate.

The following paragraphs describe the experiments done, and the results obtained, trying to retrieve the ski-pass ID of injured skiers using the approach introduced above.

7.1.1 Technical approach

Let \( a \in A \) be a tuple \((loc, t, p)\) representing a skiing accident that happened at time \( t \) on slope \( loc \), involving a skier with profile \( p \). Let \( s \in S \) be a tuple \((p, l_{last}, t_{last})\) representing the last ski-lift run for the day of a skier, where \( p \) is the profile of the
7.1. RETRIEVING SKIPASS-ID OF INJURED SKIERS

skier, and \( l_{last} \) and \( t_{last} \) are respectively the ski-lift and the time of arrival relative to the ski-lift run. Finally, let \( is\_connected(l, loc) \rightarrow \{0, 1\} \) be a function that, given a ski-lift \( l \) and a slope \( loc \), outputs 1 if the ski-lift can lead to that slope and 0 otherwise.

The function to find the set of candidates for an accident \( a \), considering a time window \( TW \), is defined as:

\[
\text{candidates}(i, TW) = \{ s \in S : is\_connected(s, l_{last}, a, loc) \\
\quad \land s.t_{last} < a.t < (s.t_{last} + TW) \\
\quad \land s.p = a.p \} 
\] (7.1)

Conceptually, \( TW \) represents an estimation of the maximum time that can pass between the moment a skier leaves the ski-lift and the moment the accident happens. A smaller value of \( TW \) intuitively leads to less skiers being considered as potential candidates, with the risk of missing the target skier. Conversely, a larger value leads to more skiers being considered as potential candidates, potentially resulting in more than one matching profile and therefore the impossibility to uniquely identify the target skier. For these reasons, the \( TW \) parameter should be tuned in order to obtain a compromise between these two issues.

7.1.2 Experiments and results

As mentioned above, the network of ski-lifts and slopes is often fairly complex, and it is usually impossible to determine exactly which slope a skier skied solely from the ski-lifts they used before. As a consequence, extracting candidates can be challenging for accidents that happened on slopes that can be reached using multiple ski-lifts, as all the skiers that used one of these ski-lifts represent potential candidates (i.e., they could have been on the slope of the accident). For this reason, this methodology is first evaluated in optimal conditions, by considering only accidents that happened on slopes where this problem is negligible. Specifically, we consider only slopes that can be reached using a minimum number of ski-lifts, in order to minimize the pool of candidates for each accident. Among the two ski-resorts analyzed in this project (Pinzolo and Cermis), only two slopes meet this requirement: the “Brenta” slope in Pinzolo, which is reachable by 2 ski-lifts, and the first part of the “Salera” slope in Cermis, which is reachable by a single ski-lift.

During the three ski-seasons considered in this project, 21 accidents happened on the Brenta slope and 14 accidents happened on the Salera slope.

This experiment is performed considering all the ski accidents that happened on the two slopes during the seasons for which the necessary data is available (i.e.,
Table 7.1: Results obtained trying to identify the skipass-ID of skiers who were involved in accidents on the Brenta and Salera slopes. *Matching candidates* indicates the number of accidents for which a single candidate with a matching profile is identified.

<table>
<thead>
<tr>
<th>Slope</th>
<th>Accidents</th>
<th>Time window</th>
<th>Tot. candidates</th>
<th>Matching candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brenta</td>
<td>21</td>
<td>20</td>
<td>22</td>
<td>6 (29%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40</td>
<td>33</td>
<td>10 (47%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>60</td>
<td>36</td>
<td>9 (42%)</td>
</tr>
<tr>
<td>Salera</td>
<td>14</td>
<td>20</td>
<td>56</td>
<td>3 (21%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40</td>
<td>70</td>
<td>3 (21%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>60</td>
<td>84</td>
<td>3 (21%)</td>
</tr>
</tbody>
</table>

2012/2013 for the Brenta slope, and both 2012/2013 and 2013/2014 for the Salera slope). Time windows of 20, 40 and 60 minutes are considered.

A summary of the results obtained is provided in Table 7.1. As can be seen, the best results are obtained on the Brenta slope considering a time window of 40 minutes, which resulted in the identification of single candidates for 47% of accidents. For the Salera slope, the results are much worse, as single candidates are identified for 3 accidents only, over a total of 14 accidents.

While the accuracy achieved on the Brenta slope may be considered acceptable, the fact that this approach achieved a fairly low accuracy on the two “optimal// slopes indicates that this methodology cannot be considered reliable.

The reason for the poor accuracy could be that the assumption at the base of this approach may often be invalid, as injured skiers could be transported downhill via ski-lift when the injury is not serious, thus “blending them in the crowd”.

The most effective approach to fix this problem for future data would be to collect the ski-pass ID of injured skiers when filling the reports of accidents. This would enable a consistent and reliable identification of the activities (i.e., ski-lift runs) of injured skiers, therefore allowing a detailed analysis of the relation between the activities of skiers and their risk of sustaining injuries.

### 7.2 Behavior and risk

Since the previous experiment did not achieve an acceptable accuracy at identifying injured skiers, it is not possible to effectively study the relation between the activities of skiers (as estimated from their ski-lift runs) and their risk of accidents. The reason is that it is necessary to know exactly which skiers sustained injuries in order to compare their activities with those of non-injured skiers and determine the potential risk-factors. Therefore, this problem can currently be addressed only
7.2. Behavior and Risk

as a proof-of-concept, in order to assess the feasibility and relevance of a more
complete study on this matter, that could be performed once the necessary data
becomes available.

For this purpose, we propose and evaluate a small number of simple metrics
that can be obtained by analyzing the activities of skiers (from their history of
ski-lift runs) that could potentially be related to their risk of accidents.

7.2.1 Experiments and results

As detailed in Section 5.1.1, the data collected by the ski-lift infrastructure
provides information on all the ski-lift runs performed by each skier. A total
of 8,559,935 records of ski-lift runs are available, with an average of 10.8 ski-lift
runs per skier/day.

The activities of skiers can be represented as a series of transitions between
successive pairs of ski-lifts, where each transition represents a skier using the first
ski-lift and then skiing from the arrival point of that lift to the starting point of
the second one. Due to the complexity of slopes networks, the exact path that the
skier used to transit from the first ski-lift to the second one is unknown.

For this analysis, skiers that did less than 3 ski-lift runs in a day are ignored,
as the data would not be sufficient to obtain meaningful information about them.
Similarly, transitions between pairs of ski-lifts that account for less than 1/1000 of
the total transitions in the same ski-area are not considered, as they would mostly
represent pairs of ski-lifts that are not connected by regular slopes (for which the
transitions were generated by skiers taking unusual paths, or possibly ski-patrols
and ski-area maintainers occasionally traveling on snow mobiles).

An important information for this analysis is the travel time for each transition
between pairs of ski-lifts. It is worth to highlight that the time that a skier spends to
travel from the arrival point of a ski-lift to the entrance of the successive one does
not consist only of skiing time, but it also includes the time that a skier spends
resting, chatting or otherwise doing other activities.

Formally, let \( r \in R \) be a tuple \((s, r_1, r_2)\) representing two successive ski-lift
runs \((r_1 \text{ and } r_2 \text{ respectively})\) by a skier \( s \). The travel time for the transition of skier
\( s \) from the first lift to the second one is defined as:

\[
t_{\text{time}}(r) = r_2.\text{timestamp} - (r.r_1.\text{timestamp} + r.r_1.\text{lift.time})
\]

where \( r.r_1.\text{lift.time} \) indicates the travel time of the ski-lift used in \( r_1 \).

Intuitively, the scale of travel times can vastly change between different pairs
of ski-lifts, as the time required to transit between the two depends on distance,
difference of altitude and potentially other variables. To normalize the information
about travel-times, it is necessary to obtain an approximation of the time generally
required to transition between each pair of ski-lifts. A possible approach to estimate the expected travel time for a pair of ski-lifts is to consider the median travel time among all the transitions between the same ski-lifts.

More formally, given a pair \((l_1, l_2)\) of ski-lifts, the expected travel time can be obtained as:

\[
est_{\text{time}}(l_1, l_2) = \text{median}_{r \in R} \{ t_{\text{time}}(r) \mid r.r_1.lift = l_1 \land r.r_2.lift = l_2 \}\]

At this point, the travel time relative to a transition between two ski-lifts can be normalized by dividing it by the expected travel time for the same pair of ski-lifts, as follows:

\[
norm_{\text{time}}(r) = \frac{t_{\text{time}}(r)}{\est_{\text{time}}(r.r_1.lift, r.r_2.lift)}\]

Conceptually, this value indicates how faster (or slower) a skier transitioned between two ski-lifts compared to the rest of the population.

With the information obtained in previous steps, the activities of each skier can be represented as a series of transitions between successive pairs of ski-lifts, with the associated travel times.

### 7.2.1.1 Tiredness

A first metric that could intuitively have an impact on the risk of accidents for a skier is their level of tiredness. While the actual tiredness of a skier depends on many factors that are not known (e.g., the age and strength of a skier), it is possible to obtain a relevant metric by estimating the intensity of the activities they did up to a certain time.

For this purpose, we evaluate a simple metric consisting of the sum of the expected travel times \(\est_{\text{time}}\) of the transitions done by the skier. The idea is that the expected travel time between two ski-lifts should approximate the level of activity needed to transit between the two ski-lifts. Formally, the tiredness metric of a skier \(s\) at time \(\text{time}\) is obtained as:

\[
tiredness(s, \text{time}) = \sum_{r \in R \atop r.s = s} \{ \est_{\text{time}}(r.r_1.lift, r.r_2.lift) \mid r.r_1.timestamp < \text{time} \}\]

To evaluate the relevance of this metric for the estimation of the risk of accidents, it would be necessary to test the correlation between the tiredness of each skier and the incidence of injuries. As mentioned above, this cannot be done directly since the necessary information (i.e., labels indicating which skiers got injured) is missing. It is however possible to verify whether there is a correlation between the average level of tiredness in the population (i.e., among all the people skiing in a ski-resort at a particular time) and the ratio of accidents.
For this purpose, the average level of tiredness (according to the metric defined above) in each ski-resort is calculated for each day for time-slots of 1 hour, and it is applied to the dataset of skiing-sessions described in Section 5.2.

The estimated level of tiredness is discretized into 10 bins, and for each bin the ratio of accidents is calculated as the number of accidents over the total number of skiers in the ski-resort. A confidence interval of 95% for the ratio of accidents is obtained via bootstrapping (with 5000 bootstrap resamples). In addition, a logistic regression is estimated on the data, in order to examine the relation between the average tiredness and the ratio of accidents. A confidence interval of 95% is obtained for the logistic regression model as well, using the bootstrapping technique with 5000 bootstrap resamples.

The resulting plot is shown in Figure 7.1. The plot suggests that there is a non-linear relationship between the average estimated tiredness of skiers in a ski-resort and the frequency of accidents. Specifically, the lowest risk is associated to the lowest level of estimated tiredness, while the frequency of accidents appears to stabilize after a certain point. An interesting observation is that a peak in the ratio of accidents is reported before the frequency of accidents stabilize to a lower value.

It is worth to highlight that the average level of tiredness will generally be correlated with the time of the day, therefore the ratio of accidents in this experiment may be influenced by other unknown factors (e.g., related to lunch time) in addition to the tiredness of skiers.
7.2.1.2 Behavior

In addition to the level of tiredness at a particular moment, it could be interesting to evaluate the relation between the behavior of skiers (i.e., their skiing ability and habits) and their risk of accidents.

For this experiment, we consider a number of fairly simple metrics that can be obtained by analyzing the ski-lift runs of each skier. Specifically, the following metrics are evaluated:

- **Permanence**: the average time (in seconds) that a skier spends in a ski-resort during the day, determined as the average time difference between their last ski-lift run in a day and their first one.

- **Median normalized travel time**: the median value of \( \text{norm}_{\text{time}} \) for a skier. It represents how “fast” a skier generally moves through pairs of ski-lifts compared to other skiers.

- **Skied ratio**: the ratio between the sum of the expected travel times \( \text{est}_{\text{time}} \) for all the transitions done by a skier and the permanence of the skier in the ski-resort. It approximates the intensity of activities by a skier: a higher ratio represents a skier who, on average, skies more in a unit of time compared to skiers with a lower ratio.

Similarly to the previous experiment, the evaluation is performed by averaging each metric among the population of skiers in a ski-resort for each slot of time, and studying the impact of this factor on the frequency of accidents.

In addition, we compare the distribution of each metric among categories of skiers that can be assumed to have significantly different behaviors. Specifically, we consider the following classes of skiers:

- **Children** with a maximum age of 7 years;
- **Members of FISI** (Italian Federation for Winter Sports); and
- **Athletes**.

The motivation for this choice of categories is that they can often be identified from their skipass (as mentioned in Section 5.1.1) and they should intuitively represent skiers with substantially different skiing experience and habits. Specifically, it is assumed that children mostly consist of beginners, while members of FISI mostly consist of passionate skiers.

The distribution of each metric among the mentioned categories of skiers is visualized using violin plots, showing both the quartiles (i.e., 25th, 50th and 75th percentiles) and a kernel density estimate. In addition, the relation with the ratio
7.2. BEHAVIOR AND RISK

Figure 7.2: Violin plots representing the distribution of behavior-related metrics among different categories of skiers.

of accidents is visualized using the same approach used for the tiredness plots, showing a 95% confidence interval for both the ratio of accidents in each bin and the obtained logistic regression.

The resulting plots are shown in Figure 7.2 for the distribution of each metric among different categories of skiers, and in Figure 7.3 for the relation of each metric (averaged over all the skiers in the ski-resort) with the hourly ratio of accidents. As can be seen, the distribution of the obtained metrics tend to be noticeably different among the considered categories of skiers. In particular, children report a generally lower skied-ratio and a higher median normalized travel time compared to athletes and FISI member. Conversely, the distribution of permanence does not change significantly among different categories of skiers, even though it is interesting to notice how athletes tend to rarely report a long permanence in comparison to other classes.

Regarding the relation of the metrics (averaged over all the skiers in the ski-resort) and the frequency of accidents, no significant correlation is observed. The reason could be that the average value of a metric does not generally change significantly between different days and times, hence making it difficult to draw conclusions on this matter. In order to obtain meaningful results it would be necessary to perform this analysis on a personal level, by analyzing the relation between the metrics for a single skier and their personal risk of accidents.

7.2.1.3 Discussion of results

This analysis was substantially limited by the impossibility to evaluate the relation of each metric with the risk of accidents for skiers on a personal (i.e., per-skier)
level. For this reason, the experiments were performed by considering the relation between the metrics, averaged over all the skiers in the ski-resort, and the hourly ratio of accidents.

The experiments related to the tiredness of skiers suggest that there may be a non-linear relation between the estimated tiredness of skiers and their risk of accidents. Furthermore, the experiments related to the behavior of skiers show that the distribution of the considered metrics appears to be substantially different among different categories of skiers, such as children, athletes etc. However, no substantial relationship is observed between the average of behavior-related metrics among all the skiers in a ski-resort and the hourly ratio of accidents.

To conclude, the experiments performed in this chapter suggest that it is possible to retrieve meaningful information related to the behavior of skiers and their level of tiredness by analyzing the history of ski-lift runs of each skier. However, to perform a complete and meaningful evaluation of the relevance of this information for the estimation of risk it is necessary to know which activities are associated to injured skiers. The necessary information could be easily collected for future ski-seasons, simply by collecting the skipass-ID of injured skiers when the accidents reports are filled. Therefore, it could be possible to perform a more in-depth analysis on the behavior and tiredness of skiers in future iterations of this project, which could potentially lead to more accurate risk-models.
Chapter 8

Discussion and conclusion

This chapter discusses the methodology used, the results obtained and a number of other aspects related to this thesis. Also, it provides a conclusion, a discussion of the future work that could be done to improve this project, and a discussion of the ethical implications of this thesis.

8.1 Discussion

8.1.1 Methodology

As detailed in Section 2.2, a limited number of past studies already addressed the task of predicting the risk of accidents for skiers. The main differentiation of this thesis is the fact that it aims to estimate the probability of skiing accidents on a personal basis, while other studies were limited to estimating the total number of accidents during a day in a ski-resort.

In order to obtain a fairly generic model, which can potentially be applied to many different ski-resorts, this study relies on data that is commonly available to ski-resorts, specifically the reports of accidents, the history of usage of ski-lifts, and weather reports. This project cleans and combines this data in order to obtain a representation of the presence of skiers in a ski-resort at different times of the day, along with information about skiing accidents.

Once this dataset is obtained, the problem is addressed by training probabilistic classification models to predict the probability of accidents for skiers. A number of techniques to address the problems of class-imbalance and probability calibration are evaluated and compared.

The resulting models are evaluated in terms of discrimination ability and calibration of probabilities. Due to the lack of previous state-of-the-art studies on this specific problem, it is not possible to compare the obtained results with
the results achieved by other studies. For this reason, the evaluation is performed by considering the natural interpretation of the metrics used, and the baseline performance (i.e., the results that would be obtained by a naïve or random model) for these metrics.

8.1.2 Data preparation

The dataset is obtained by combining the history of ski-lift runs in the ski-resort with reports of accidents collected by ski-patrols. Additional information related to the weather situation is also retrieved from publicly accessible weather stations.

In addition, the condition of the snow is estimated by an auxiliary predictive model, trained on historical information about temperature and precipitations. The best model obtained for this task achieves a MCC score of 0.418, and its probabilistic prediction of the snow condition is used as an additional feature for the risk model. An MCC of 0.418 indicates that the model does not have a perfect accuracy, but it still provides useful information regarding the condition of the snow. It is worth to remember that the MCC score can take a value between −1 and 1, where 0 indicates a random or static (i.e., predicting always the most frequent label) classifier, and 1 represents a perfect classifier.

Since personal data about skiers is occasionally unavailable, an imputation process is performed to fill the missing values, using the Expectation Maximization technique. This process relies on the assumption that the data is Missing At Random (MAR), which means that the probability that a value is missing does not mostly depend on the value itself. If the MAR assumption is correct, the bias introduced by the imputation process should be minimal and would not represent a problem. However, it is important to notice that this cannot be verified, and the only way to make sure that this problem does not happen would be to collect better (i.e., more complete) data.

8.1.3 Risk model

A number of learning algorithms are trained on the obtained dataset and evaluated. Since the dataset is heavily class-imbalanced due to the rarity of skiing accidents, a number of techniques to deal with class-imbalance are compared. The best results in terms of discrimination ability are achieved by a Gradient Boosted Trees model using the Balanced Bagging technique, which achieves an auROC score of 0.689. This means that, considering a random positive sample and a random negative sample, the model will output a higher risk score for the positive sample in 68.9% of the cases. Considering the randomness associated with skiing accidents, this can be considered a fairly good result. It is worth to remember that random and static classifiers would obtain an auROC score of 0.5.
8.1. Discussion

To estimate reliable probabilities for the risk of accidents, two well known probability calibration techniques are applied on the models and compared. Surprisingly, the best calibration among all models is achieved by a Gradient Boosted Trees model when no calibration technique is used. Reliability diagrams suggest that, while this model achieves a relatively good calibration of probabilities, there is room for improvements. We can suppose that a better calibration of probabilities could be obtained if a larger amount of data was available.

An evaluation of the relevance of features shows that the features that appear to have a higher predictive power are related to the environment, and to a lesser extent to the personal profile of skiers. However, it is worth to highlight that the estimated relevance of features may be skewed due to correlations among different types of features.

8.1.4 Analysis of skiers’ activities

Finally, we evaluated the possibility to study the activities of skiers in order to perform a more accurate estimation of their risk of accidents. This experiment was limited by the impossibility to know exactly which skiers sustained injuries, which made it impossible to effectively study the relation between the activities of skiers and their risk of accidents. For this reason, an experiment was performed with the goal of determining the identity of injured skiers by searching for candidates among the skiers who may have been near to the place of an accident at the time when the accident happened. However, this experiment was not successful, and the aforementioned problem was not solved. Despite this issue, we performed a simple experiment by defining a small number of metrics related to the level of tiredness of skiers and their behavior, obtained from their history of ski-lift runs. We then studied the relation between the average of each metric among the population of skiers and the reported ratio of accidents. Results suggest that there is a non-linear correlation between the estimated tiredness of skiers and their risk of accidents. Furthermore, a number of fairly simple behavior-related metrics showed noticeably different distributions among different categories of skiers, thus suggesting that similar metrics may be useful for the estimation of risk as well.

8.1.5 Challenges

As it often happens in machine-learning projects, the biggest limitation encountered by this work is related to the available data. First, some data (specifically, personal information about skiers) had a relatively high ratio of missing information, therefore the imputation process could have introduced some bias in the dataset. As mentioned above, if the MAR assumption is correct, the introduced
bias should not represent a problem. However, this cannot be verified, therefore it would be necessary to collect better data in order to remove the need for this assumption.

Furthermore, some of the risk factors identified by past literature (detailed in Section 2.1) could not be considered in this project, as information about them was not available. In particular, the personal information about skiers was limited to their gender and age group, and no information was provided regarding their skiing experience, winter sport practiced (e.g., skiing or snowboarding) or other potentially useful variables.

Also, while the volume of data was not excessively low, a larger dataset (i.e., spanning a higher number of ski-seasons) would almost certainly lead to a better precision and generalization ability of models, in particular from the perspective of calibration of probabilities, which is fairly sensitive to the quantity of available data.

Finally, the impossibility to link the dataset regarding ski-lift runs with the dataset of skiing accidents prevented us from accurately studying the relation between the activities of skiers and their risk of accidents.

8.2 Conclusion

The objective of this thesis is to propose a methodology to perform a personal estimation of the probability for skiers to sustain injuries, relying on data that is commonly available to ski-resorts. To our knowledge, this is the first study that addresses the problem of estimating the risk of accidents for a skier on a personal level.

This problem is addressed by combining data from a number of different sources in order to obtain a dataset representing the presence of skiers in a ski-area at different times, along with the incidence of skiing accidents. The risk of accidents for a skier is then computed by probabilistic binary classification models. The obtained models achieve a reasonably good performance from the perspectives of both discrimination ability and calibration of probabilities.

Finally, additional experiments suggest that it may be possible to improve the accuracy of the risk estimation by performing an analysis of the behavior of skiers. However, in order to enable this kind of analysis it would be necessary to collect additional information, specifically the skipass-id of injured skiers.
8.3 Future work

As mentioned above, this project could largely benefit from the availability of additional data. Collecting data of ski-lift runs relative to more skiing seasons could be useful to improve the performance of models, as the data available to this project was relatively limited.

In addition, it would be possible to start collecting the skipass-ID of skiers who sustain injuries for future skiing accidents. This would only require a small update to the SicurSkiWeb platform, and it would enable to study the relation between the usage of ski-lifts by skiers and their risk of accidents for future data.

Furthermore, in this project the task of estimating the condition of the snow is addressed by training a relatively simple model, that achieves a satisfactory, but not excellent, performance. Further work can be performed to improve the performance of this model, for example by considering more variables (e.g., wind, sunlight, etc.) or by employing more sophisticated classification techniques.

Another area where this project could be expanded is the estimation of the type and severity of injuries. While this project focused on studying the probability of a person to get involved in an accident in a ski resort, it could be useful to estimate the most probable types of injuries, or their severity. The accident reports collected by the SicurSkiWeb platform contain detailed information about causes of accidents, profile of the injured skiers and diagnosis of the injuries. From this data, it is possible to study the patterns in injuries related to alpine skiing. For example, it would be possible to study whether some variables (e.g., age of subject, sport practiced, etc.) is correlated with the place of the injury (e.g., head, arms, etc.) or with the diagnosis. Estimating the type and entity of injuries could be useful for a number of use cases, for example to estimate the expected losses for insurance companies, or simply to provide skiers with a better analysis of their risks, thus encouraging them to practice winter sports safely.

Finally, it could be possible to address this problem from a different perspective. An example could be to perform the analysis of risk on a smaller scale, for instance by considering the risk of sustaining injuries on single slopes. This analysis would require the availability of additional data (specifically, the exact set of slopes where each skier skied), but it could potentially achieve a better performance at estimating the risk of accidents, by considering features which are specific to each slope (e.g., difficulty, width, etc.).

8.4 Ethical considerations

A common ethical issue faced by projects that apply machine learning techniques to personal data of users regards the sensitivity of the data involved. In this project,
the only personal information currently considered consists of the age group and gender of skiers. The sensitivity of this information would depend on the use-case. For example, if this project is used to inform skiers about their risk to sustain injuries, it would not represent a problem from an ethical perspective. However, the insurance use-case may be more problematic, as providing different offers to people of different genders or ages could be considered as unfair or unethical.

In addition, in case future iterations of this project will consider more features, it will be necessary to assess the related ethical implications as well. For example, health conditions of a person may intuitively affect their risk of being involved in an accident, but using such a sensitive information could be considered as a violation of privacy and an unethical practice.

Another important ethical consideration is that the estimation of the risk for skiers should be transparent, in order to allow users to know the factors that influenced their score. As mentioned above, some models provide a straightforward way to interpret the predictions. All the predictive models considered in this project, except for neural networks, can be considered white-boxes, as their results can easily be interpreted and justified.
Bibliography


Appendix A

Details on configuration of models

This appendix provides information on the process of optimization of hyperparameters performed in order to determine the configuration of the predictive models used in this project.

The configuration of models is mostly optimized using either the grid-search or random-search techniques. It is important to note that not all the parameters can be determined with these methods. For example, the number of trees for a Random Forest cannot be optimized with this approach, since a higher number of trees tends to always achieve a more accurate model at the cost of a higher computational complexity, hence this parameter needs to be determined as a compromise between accuracy and computational cost of the model. Therefore, the tuning of models is based both on the automatic search of optimal parameters and a manual tuning process.

A.1 Snow condition model

For each classifier, most of the hyperparameters are optimized either via grid-search or random-search, with a 5-fold cross validation, while some parameters are manually tuned. Cross validation is performed ensuring that all the samples from the same day are in the same fold, in order to reduce the bias introduced by similar samples.

Table A.1 contains the distribution of hyperparameters considered during the process of optimization of hyperparameters for the models used to predict the condition of the snow. For the FeedForward Neural Network, a grid-search approach was used, on a fixed set of hyper-parameters. For the other models, random-search was used, testing 250 sets of hyperparameters for each model. The multi-class Matthew’s Correlation Coefficient (MCC) was used as metric to select the best models.
### Random Forest

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Distribution</th>
<th>Best value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max features per split</td>
<td>[2, 12]</td>
<td>3</td>
</tr>
<tr>
<td>Criterion</td>
<td>{gini, entropy}</td>
<td>gini</td>
</tr>
</tbody>
</table>

### Gradient Boosted Trees

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Distribution</th>
<th>Best value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning rate</td>
<td>{0.001, 0.01, 0.1, 0.2}</td>
<td>0.1</td>
</tr>
<tr>
<td>Maximum tree depth</td>
<td>[2, 15]</td>
<td>4</td>
</tr>
<tr>
<td>Number of estimators</td>
<td>[50, 200]</td>
<td>193</td>
</tr>
<tr>
<td>Subsample ratio</td>
<td>[0.3, 1.0]</td>
<td>0.9</td>
</tr>
</tbody>
</table>

### Feedforward Neural Network

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Distribution</th>
<th>Best value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of neurons</td>
<td>{10, 15, 30, 50, 70, 100}</td>
<td>100</td>
</tr>
<tr>
<td>Activation function</td>
<td>{tanh, relu, sigmoid}</td>
<td>tanh</td>
</tr>
<tr>
<td>Dropout rate</td>
<td>{0.25, 0.5, 0.75}</td>
<td>0.25</td>
</tr>
<tr>
<td>Optimizer</td>
<td>{SGD, Adam}</td>
<td>Adam</td>
</tr>
<tr>
<td>Epochs</td>
<td>{30, 50, 70, 100}</td>
<td>30</td>
</tr>
</tbody>
</table>

Table A.1: Hyperparameters space considered for tuning of models to predict the condition of the snow.

The Random Forest model is configured with 100 trees, a maximum number of features per tree of 3, using “gini” as the criterion to determine the best split. The Gradient Boosted Trees model is configured with a learning rate of 0.1, a maximum depth per tree of 4, 193 estimators and a subsample ratio of 0.9. Finally, the Neural Network is configured with a single hidden layer of 100 neurons with the hyperbolic tangent activation function (a.k.a. “tanh”), a dropout ratio of 0.25, and “Adam” as optimizer. To train the neural network, 30 epochs of training are performed.

### A.2 Risk models

For each model, the optimal hyperparameters are determined by using either a grid-search or a random-search approach, performed with a 5-fold cross-validation.

Table A.2 shows the hyperparameters-space explored to optimize the models used for the prediction of the probability of accidents. The search for the optimal configuration is performed with a grid-search approach for the Logistic
A.2. RISK MODELS

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Distribution</th>
<th>Best value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Penalty type</td>
<td>${L_1, L_2}$</td>
<td>$L_1$</td>
</tr>
<tr>
<td>Inverse regularization strength</td>
<td>${0.001, 0.01, 0.1, 1.0, 10.0, 100.0}$</td>
<td>0.1</td>
</tr>
<tr>
<td>Random Forest</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum features per split</td>
<td>$[2, 16]$</td>
<td>6</td>
</tr>
<tr>
<td>Criterion</td>
<td>${\text{gini, entropy}}$</td>
<td>gini</td>
</tr>
<tr>
<td>Gradient Boosted Trees</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learning rate</td>
<td>${0.001, 0.01, 0.1, 0.2}$</td>
<td>0.01</td>
</tr>
<tr>
<td>Maximum tree depth</td>
<td>$[2, 15]$</td>
<td>6</td>
</tr>
<tr>
<td>Number of estimators</td>
<td>$[50, 200]$</td>
<td>112</td>
</tr>
<tr>
<td>Subsample ratio</td>
<td>$[0.3, 1.0]$</td>
<td>0.8</td>
</tr>
<tr>
<td>Feedforward Neural Network</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of neurons in hidden layer</td>
<td>${10, 15, 30, 50, 70, 100}$</td>
<td>50</td>
</tr>
<tr>
<td>Activation function</td>
<td>${\text{tanh, relu, sigmoid}}$</td>
<td>tanh</td>
</tr>
<tr>
<td>Dropout rate</td>
<td>${0.25, 0.5, 0.75}$</td>
<td>0.5</td>
</tr>
<tr>
<td>Optimizer</td>
<td>${\text{SGD, Adam}}$</td>
<td>Adam</td>
</tr>
<tr>
<td>Epochs</td>
<td>${30, 50, 70, 100}$</td>
<td>50</td>
</tr>
</tbody>
</table>

Table A.2: Hyperparameters space considered for tuning of models to predict the probability of accidents.

Regression and for the Feedforward Neural Network models, while it is performed with a random-search for the other models.

The optimization of hyperparameters is performed on the dataset obtained using the Random Undersampling technique, using the auROC score as metric to compare the different configurations. The models trained on the full dataset are manually tuned starting from the configuration of the same model obtained on the undersampled dataset. The reason is that training a model on the full dataset takes a considerably longer time (generally tens of minutes for each model), therefore performing grid-search or random-search is not practical.

The final configuration for each model is now described. The Random Forest classifier is configured with 200 trees when trained on the undersampled dataset, while 100 trees when trained on the full dataset, using “gini impurity” as criterion to measure the quality of a split and with a maximum number of features per split of 6.

The Gradient Boosted Trees model is configured with a learning rate of 0.01,
A number of estimators of 112, a maximum depth of trees of 6, and a subsample ratio of 0.8. When trained on the full dataset, a subsample ratio of 0.6 is used. The loss function to optimize is set to “deviance”, since it is considered the optimal one when the aim is to obtain probabilistic outputs.

The Logistic Regression is configured to use L1 penalty with the C parameter set to 0.1 (note that C represents the inverse strength of regularization, therefore a smaller value represents a stronger regularization).

Finally, the Feed-forward neural network is configured with a single hidden layer of 50 neurons and a dropout rate of 0.5, using the hyperbolic tangent (a.k.a. “tanh”) as activation function and “Adam” as optimizer. 30 epochs of training are performed to train the network.

The Balanced Bagging technique is performed with a number of estimators of 50 (i.e., on 50 balanced datasets).