POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Survival analysis techniques applied to car insurance for claims and frauds risks prediction

Tesi di Laurea Magistrale in
Computer Science Engineering - Ingegneria Informatica

Author: **Omar Abdrabou**

Student ID: 963500
Advisor: Trovò Francesco
Academic Year: 2021-22

Abstract

In today's insurance industry, the ingredients for success are not a secret. With the increase in the digitalization of business processes, swiftly handling policy applications and claims is key in providing excellent customer service. In this complex and fast-paced environment, a deep understanding of the customer is no longer optional. This thesis presents the results of applying Survival Analysis techniques and Machine Learning algorithms to the car insurance sector. First, in order to define a starting point, the trends and structure of the Italian Insurtech sector were investigated to comprehend the environment and the underlying principles governing it. Then, the users' data was analyzed with the data owner's cooperation to discover what elements and which characteristics were relevant to the task at hand, in order to define a work perimeter. The prepared dataset was then used as input for various models. Before implementing the survival model, first we started with a classification model, to understand how a more classical approach would fare, and check if it could have been sufficient. The results analysis process revealed that classification is unable to achieve the set goals, especially as censored data, which are common in this scenario, cannot be handled properly by the model. The comparison between the survival baseline and the final model highlighted the need to exploit all features to properly estimate an acceptable survival function, in particular in this sector, where the indicators to be used are many. Lastly, possible points of improvement and considerations regarding future developments of the taken approach are listed at the end of the document.

Keywords: Survival analysis, Risk prediction, Machine learning, Insurtech, Fintech

Abstract in lingua italiana

I fattori che contribuiscono al successo all'interno del settore assicurativo del giorno d'oggi non sono un segreto. Con l'incremento del livello di digitalizzazione dei processi di business, una rapida e accurata gestione dei sinistri è la chiave nel fornire un servizio clienti di qualità. In questo settore così dinamico e complesso, una profonda consapevolezza delle caratteristiche dei propri clienti non rappresenta un fattore ormai ignorabile. Questa tesi espone i risultati ottenuti dall'applicazione di tecniche di Analisi di Sopravvivenza e algoritmi di Machine Learning al settore assicurativo automobilistico. Il lavoro svolto si è basato su un'analisi dei trend e della struttura del settore Insurtech, in particolare italiano, con l'obiettivo di comprenderne le caratteristiche e i principi che ne regolano le dinamiche. Successivamente, i dati condivisi sono stati analizzati scrupolosamente in collaborazione con il data owner, in modo da identificare quali elementi e aspetti fossero rilevanti per il raggiungimento dell'obiettivo prefissato, definendo quindi un perimetro di lavoro ben preciso. Il dataset così definito è stato poi usato come input per diversi modelli. Prima di implementare i modelli di sopravvivenza, è stato testato un modello di classificazione con l'obiettivo di comprendere come un approccio più classico avrebbe performato nel caso in esame, e se fosse stato eventualmente sufficiente. Il processo di analisi dei risultati ha rivelato come un modello di classificazione non sia in grado di raggiungere gli obiettivi prefissati, in particolar modo dovuto alla prevalente presenza di dati censurati, che il modello non riesce ad utilizzare correttamente. Il confronto finale tra i modelli di sopravvivenza, baseline e ultimo modello implementato, ha evidenziato la necessità di sfruttare a pieno tutte le feature per stimare una funzione di sopravvivenza sufficientemente precisa, in particolar modo in questo contesto, dove gli indicatori da utilizzare sono verosimilmente molti. Infine, possibili aspetti migliorabili e considerazioni su sviluppi futuri riguardo al lavoro svolto sono esposti nell'ultima sezione dell'elaborato.

Parole chiave: Analisi di sopravvivenza, Previsione del rischio, Machine learning, Insurtech, Fintech

Contents

1 | Introduction

In recent years, the insurance industry has gradually emerged as an area of opportunity for entrepreneurs seeking to address the inefficiencies and lack of customer-centricity in the current insurance ecosystem. The technological advancements brought by the development of AI, coupled with the collection and analysis of vast quantities of data, have paved the way for tech-driven insurance platforms and technology providers to address significant challenges and shortcomings in this industry. In this already complex ecosystem, the paradigm shift brought by the introduction of new, tech-driven techniques, generated new roles for these new businesses, aside from the traditional agents, brokers, and plain insurers. These new entities can be classified as technology solution providers, dealing with more than simple policies, offering solutions related to new platforms, or the definition and implementation of data processes.

A tentative definition of the term Insurtech is the following: *an insurance company, intermediary, or insurance value chain segment specialist that utilizes technology to either compete or provide valuable benefits to the insurance industry [15].*

One of the factors that have been plaguing this sector and remained constant, possibly even increased with its evolution and its transition towards a more tech-driven one, is the concept of fraud. In this case, it is possible to define it as an action committed by an individual or a group by unjustifiably claiming or being credited with accomplishments or qualities for illicit financial gains.

A glaring example of the impact fraud have is the American Insurtech sector, one of the most advanced. A 2022 study by The Coalition Against Insurance Fraud (CAIF) indicates that insurance fraud can cost U.S. consumers $308.6 billion yearly. That amount includes estimates of annual fraud costs across several liability areas, including Life Insurance ($74.7 billion), Property and Casualty ($45 billion), Workers Compensation ($34 billion), and Auto Theft ($7.4 billion) [2].

The objective of this work is to define an AI solution that has the capability of estimating the fraud risk for each customer, effectively supporting businesses in their fight against fraud by giving them an additional tool to provide insight into their customers.

The problem was approached from a high-level perspective, modeling fraud and claims

as events and applying a statistical technique called survival analysis to calculate their likelihood of occurring over an observation window. This method was combined with Machine Learning algorithms to define what is commonly referred to as Survival Function and plot its structure.
Chapter 1 presents the companies that have been involved in this work and goes into detail about the process undertaken to arrive at the definition of the goals. Chapter 2 introduces the empirical settings by employing a business analysis of the Italian Insurtech sector, before looking into the ethical guidelines of AI solutions and then analyzing the literature to understand what the State of the Art for similar issues is. Chapter 4 describes the dataset and explains all steps taken to arrive at a final dataset and the rationale behind them. Chapter 5 presents the model implementation process and discusses the results obtained. The last Chapter discusses possible points of improvement of the work done.

2 | Process Introduction: An Applied Research

As a student of the EIT Digital double master program, the process that originated this thesis entailed the involvement and direct collaboration of companies to define a use-case that encompassed real, Fintech-related scenarios that required the usage of the knowledge acquired during the program to carry out a hands-on project for a customer and present them with its findings.
The result is a deliverable that describes the undertaken process, from the definition of the goals, and the background research to the more technical parts.

2.1. Home Company

The first company that was involved with the thesis' project as the solution supplier was Red Reply, a company belonging to the Reply Group.
The Reply group is made up of a network of highly specialized companies which support leading industrial groups in defining and developing business models to optimize and integrate processes, applications, and devices by using new technology and communication paradigms, such as Artificial Intelligence, Big Data, Cloud Computing, Digital Communication, IoT, Sustainability.
Red Reply specializes in Oracle Public Cloud with a focus on Oracle Cloud IaaS and PaaS and is the only Italian Partner certified by Oracle as Managed Service Provider.
Functionally, the company is divided into three Business Units: Infrastructure, Software Development, and Data Integration. The BU that was involved for the majority of the project was the Data unit, with frequent exchanges also occurring with members of the

Infrastructure BU to configure the data links to prepare the data transfer from the client on-premise databases to a landing zone on the company's systems.
Recently, it had been evaluating the addition of Machine Learning services to its current offer, implementing Oracle's Machine Learning tools to its projects. For this reason, this thesis also served as a testing ground to gauge the tools' effectiveness and potential, and Red's prospects in this new part of the market.

2.2. Customer Company

The second company represents the 'client' interested in exploring the application of Machine Learning algorithms to its business and seeing their effectiveness directly with a 'test' before possibly approving official projects if the results were deemed satisfying. **Verti Assicurazioni** is an Italian insurance company belonging to MAPFRE, the most prominent Spanish insurance multinational, offering insurance solutions for vehicles, houses, and people. Red Reply had already established a relationship with them as a customer through other big projects, like the definition and management of their data warehouse and daily data processes, meaning that they had knowledge of their structure as a company and of their business processes, which allowed for an easier definition of potential solutions, as their business processes were familiar.

2.3. Proposal Definition Process

2.3.1. Working Methodology Overview

The only projects currently ongoing at that time with the client were related to business processes that had no affinity with the area of interest of the thesis, as they represented core tasks that had no room for any change. As such, it was not possible to take advantage of an already-defined project to base the research and future solution, adding an ulterior layer of complexity to the overall process.

To delineate an area of their business that could be improved by researching applied AI methods and applying them to the client's business, the first step was to investigate and analyze the structure of the Italian Insurtech market from both a legal and economic perspective. The objective of this stage was obtaining enough information to gain insight into technological trends that might have been overlooked and might prove beneficial to the business if taken advantage of. Then, the client's business structure and currently offered services were analyzed with the help of the Data Business Unit manager's knowledge and compared with results obtained from the previously performed research to uncover pain points and possible improvement areas.
The resulting offer had to possess all the characteristics to qualify as an interesting thesis project in the ML and FinTech field of research, while also clearly displaying its potential and the improvement it could bring to the client, convincing them to pursue the project after the thesis' end, and advance it further to fully integrate it into its business.
At the time, the only active project was the management and monitoring of an autonomous data warehouse and all its scheduled processes, ensuring that all data processes terminate correctly and fixing any error or missing data instances that might occur, along with developing new data streams when required.
The result of this analysis highlighted that the majority of the processes were not only extremely complex and already fully automatized but were also tightly bound to the legal aspects of an insurance company, meaning that there was very little space for improvement or alteration. As such, the focus shifted to processes that were not carried out inside the data warehouse's system.

2.3.2. Research Motivation

The research carried out, along with the internal brainstorming sessions with the data team, highlighted that two main routes could be taken to define a proposal. The first option would have been for the business to implement the solution as a completely new service, enabling the company to tap into a new section of the market, reaching more customers and expanding its business. This would have made it harder to define a proposal, as a successful integration with the already existing processes would have required more attention, but offered a high degree of freedom when choosing the technical details of the proposal, due to it being an addition. The second option was more in line with the initial approach, comprising of identifying business processes of relevance, but not core to the company, that could potentially be improved by the usage of ML and AI techniques. The reason for the additional restriction on the importance of the targets was due to the allowed scope of work: as core processes represent the most vital tasks, an extreme

amount of attention and planning is required to attempt altering them, which was incompatible with the 'testing' nature of the overall project. This route allowed for the easiest integration with the existing business, though at the cost of being more restricted with the selection of approaches that could be taken, which needed to be compatible with the hypothetical chosen process' goal. Taking into consideration all these characteristics and the time restrictions imposed by the client, the improvement of already existing tasks was chosen as the approach.
The business process qualified as being relevant enough and that was compatible with AI approaches was the claims and fraud detection and handling. The detection made use of information passed from external, legal entities that define the regulation of the entire insurance sector, and an archive of past identified frauds. As such, the process in place was of the reactive type, and investigations were carried out manually, with no automation or forecast tools available.
In addition, according to IVASS, the authority with the duty of managing and overseeing the work of insurance companies in Italy, and also the one publishing extensive reports of the insurance sector, 250 million euros were saved in 2019 with the current frauds detection mechanism implemented by insurance companies, **equaling 1,9% of that year insurance premium.**
Overall, these factors provided justification from both a business and economic standpoint to continue to believe claims and fraud detection and handling to be the proposal target.

2.3.3. Objectives Definition

The initial actions were carried out in close contact with the client to make use of its expertise and domain knowledge and help define in advance the work parameter of the hypothetical solution.

- First, as stated earlier, based on the part of the business we had visibility on, and by consulting the shared documentation, pain points and chances of improvements were discovered through a detailed analysis.
- The second step was to carry out research on the structure of the Italian insurance market and on the global InsurTech sector to spot trends and assess the current state of the market. The current State-of-the-Art was also investigated to understand the technologies already adopted by other players who offered similar services as Verti.
- Then, the findings were presented to the client, who collaborated to define the final goals that the solutions should reach, together with selecting the data of interest contained in his database, effectively determining the working perimeter.

The confrontation with the client was fundamental: considering the procedures already in place for fraud detection and management, it was highlighted that the business already made use of a proprietary tool for fraud detection, even if without using ML algorithms for it. As such, the previously defined target, claims and fraud detection, was not valid anymore, as not only it would not bring any relevant improvement, but it would also conflict with the internal tools already used.
For this reason, the focus shifted to **defining indicators on the clients that opened claims to determine their behavior**. More specifically, the new goal was to give an estimation of the likelihood that a client would open a claim or commit fraud, at different points in time, **effectively framing the case as claim and fraud prediction**, rather than detection.
This new addition to the existing process would have improved not only the claim management process but also brought benefits to transversal business processes by making information about fraud internally available for decision-making processes.

3 | Business Analysis and Related Works Overview

3.1. Analysis Methodology

The main objective of the market analysis was to identify the structure and driving forces that characterize the overall InsurTech market, with particular attention to the Italian section, to present the client with powerful and convincing reasons to justify the implementation of the proposed project.
The first phase of the research surveyed the juridical structure of the Italian insurance market to understand the regulations in place and the official entities that enforce them, to understand the limitations that must be respected when defining the proposal. The available statistical data about claims and fraudulent claims made public by IVASS was also examined and organized in an easy-to-understand manner to convey the importance of improving the related processes. The next stage concentrated on the study of the technological trends of the InsurTech market to acquire the necessary information to rank the client's position against other competitors, both startups and big incumbents, and detail a high-level prospect for future growth.
Then, the outlined context and requirements to satisfy were used as a starting point to analyze the current state of the art for claims and fraud detection, looking at the existing techniques, their characteristics, and use cases, to define the starting baseline from a theoretical and practical standpoint. Lastly, the focus shifted to the current literature and legislation about ethical and moral aspects of implementing ML algorithms with processes that are particularly 'close' to people and directly influence business decisions, like claims and fraud management. The definition of a solution that is efficient and performant but lacks the characteristics necessary to carefully and properly handle such delicate data, and the ethical repercussions that it could cause, would be unable to be realistically implemented.

3.2. The Italian Insurance Sector

Insurance companies of all branches in the Italian market are managed by **IVASS**, an entity with legal personality under public law which ensures adequate protection of policyholders by pursuing the sound and prudent management of insurance and reinsurance companies and their transparency and fairness towards customers.
In order to guarantee the fulfillment of the institutional objectives, IVASS oversees the supervised entities, the so-called micro-prudential supervision), by carrying out corporate governance, capital, financial and technical controls and carries out supervisory functions on the stability of the system (so-called macro-prudential supervision). IVASS also has the duty of analyzing data about individual insurance companies and the market state, compiling the results into reports that are published periodically.
It also contributes to the fight against fraud in the third-party liability sector auto, carrying out analyzes and evaluations of the information obtained from the management of a unified claims database, interacting with companies regarding the reports that have emerged, collaborating with the police forces, and the judicial authorities.

3.3. InsurTech in Italy

Concerning the development of the InsurTech phenomenon, there currently are 564 Italian companies active in the FinTech and InsurTech sectors, most of which have their offices in Northern Italy, with a total of 564 billion euros gathered by the end of December 2021, which is a small sum when compared to America or other countries belonging to the EMEA.[19]
According to the Italian InsurTech Association (IIA), the Italian InsurTech sector is currently relatively small but is characterized by a sustained growth rate, as it can be seen in figures 3.1 and 3.2. It is a phenomenon emerging in the international and European panorama as a whole and which, in its broadest meaning, identifies the entire process of digitalization of the insurance sector, from filling insurance policies to the management of claims.[8]
To date, digital consumers already correspond to around 32% of insurances' customer base, a value expected to exceed 80% by 2030 and reach 100% in 2040. But if the digital consumer is ready, the same cannot be said about the digital insurance offer, standing at only 1.5% at the end of 2020. With growth as the goal, IIA, in collaboration with the major insurance players, proposed the establishment of an Insurtech ACT, a system of incentives and dedicated funds in favor of investments in the Insurtech sector. In synthesis, this document details a plan to promote research initiatives aimed at framing

customers' needs and behaviors, integrating home insurance and customer services with new technologies, with a focus on the collection of users feedback, designing new products, and offering models starting from the needs and lifestyles of consumers and above all and offering the possibility of distributing products and services in an open insurance paradigm. The last addressed point is about the creation of digital culture by planning to define courses for insurance employers about these new technologies and technological trends.

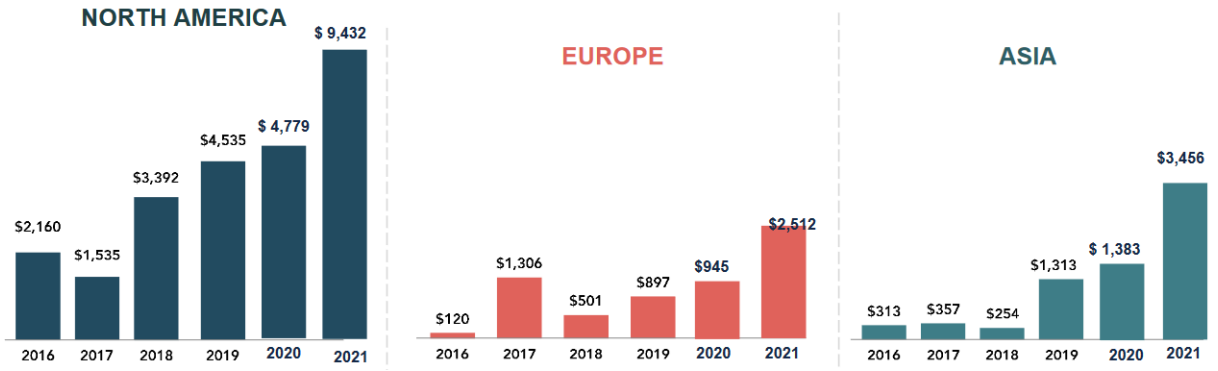

Figure 3.1: Geographical breakdown of investments in InsurTech startups.
Source: Cb Insights.

Analyzing more in detail the InsurTech sector in Italy, there are 130 innovative companies, and they are divided into two categories: 64% are proper InsurTech companies, while 36% are Tech Insurance companies, therefore companies that offer technologies for the players in the insurance sector.
Since 2021, IVASS, Bank of Italy, and CONSOB brought forth an initiative in Italy called **Regulatory Sandbox**. It is a controlled environment in which traditional and FinTech/InsurTech operators test technologically innovative products and services under a simplified transitional regime, in constant dialogue and confrontation with the supervisory authorities, thus reducing the spreading of potential risks associated with these activities. Where necessary, operators can also be granted regulatory exemptions to ease the testing phase, facilitating the overall process and encouraging the development of advanced solutions.

3.4. The Power of AI in InsurTech

The passage of time has proven that the traditional insurance business model is remarkably resilient. However, the advent of new technologies has increased the importance of a

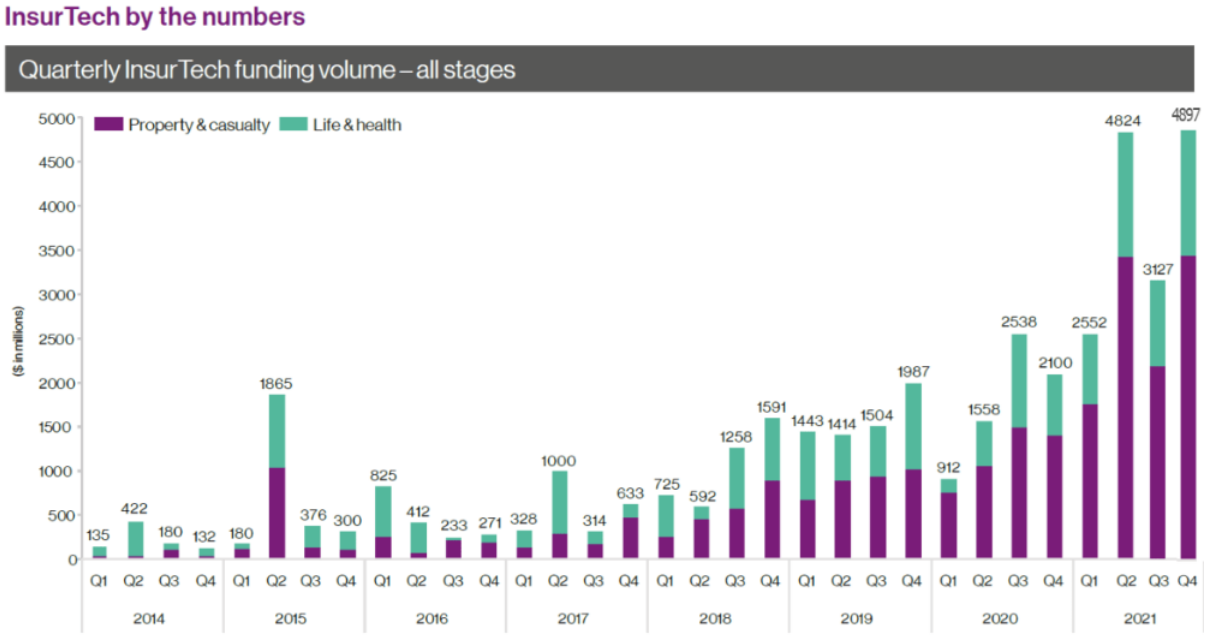

Figure 3.2: Time-wise breakdown of investments in InsurTech startups from a global perspective.
Source: Cb Insights.

transition towards more advanced digital processes, influencing all fields related to finance and indirectly changing customers' expectations, creating an increase in the number of new players in the sector. InsurTechs have attracted consumers with selective discounting based on the intersection of intelligent devices and risk-minimizing behaviors, offering, for example, meters for car mileage or in-home flood and fire detectors that autonomously signal emergency services. In our use case of car insurances, forward collision avoidance, blind-spot assist, and adaptive cruise control are already fitted in many new cars, increasing vehicle safety.
Figure 3.3 shows that large incumbents digitizing existing businesses could more than double profits over the course of five years. In the longer term, however, earnings from traditional businesses will face headwinds as driving becomes less risky due to the use of sensors and telematics or, as in the case of autonomous cars, due to liability being transferred to manufacturers. Fifteen years on, profits for traditional personal lines auto might fall by 40% or more from their peak.

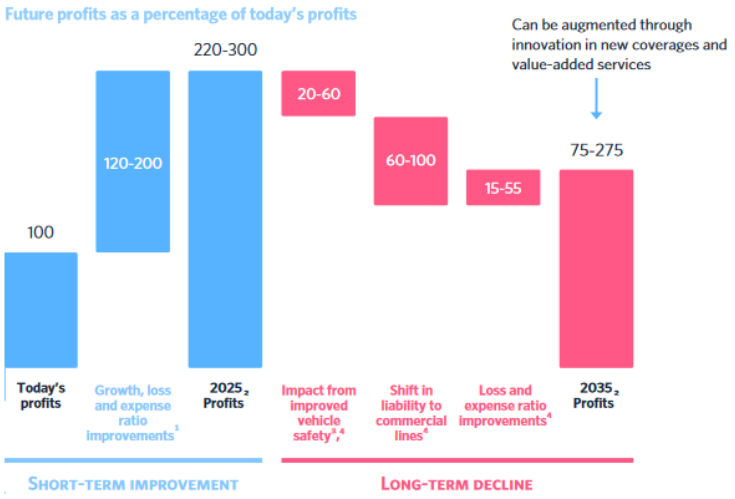

Figure 3.3: Profit prospects for traditional insurance companies.
Source: McKinsey Panorama InsurTech Database.

By 2020, 20% of vehicles globally are expected to come with safety systems already installed, reducing the number of accidents and the value of personal auto insurance policies, thus creating a future threat for insurance companies. Leading companies use data and analytics not only to improve their core operations but to launch entirely new business models: as an example, imagine a situation where companies with the necessary data and tools to analyze and understand it began to locate and catch low-risk customers in significant numbers. In this case, the insurers' traditional business model that uses the premiums collected from low-risk policy holders to balance the claims opened by high-risk profiles could easily fall apart.[4]

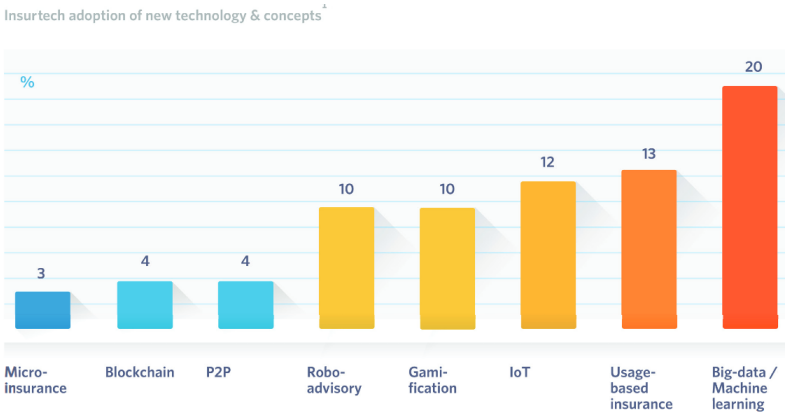

Figure 3.4: New technologies in InsurTech based on the 500 most well-known examples in the database.
Source: McKinsey Panorama InsurTech Database.

All parts of the insurance value chain can become more efficient by understanding the potential of new technology and exploiting it: according to McKinsey, automation can reduce the cost of a claims journey by as much as 30%, also increasing customers' satisfaction and the company's growth [4]. The boundary with BigTech firms and partners are blurring as there is a reorientation of internal priorities and value systems, and InsurTechs startups are already integrating eight critical new technologies and paradigms that incumbents have not widely adopted to solve real business problems, as displayed in Figure 3.4.
Insurtech innovations can enable incumbents to expand insurance coverage to market segments that were previously underinsured, introduce new products and services, simplify claims management, reduce transaction costs, and deliver enhanced risk identification and measurement.
The number one expense item for insurers is loss and with its value estimated in the billions, they must actively identify and minimize losses, and the implementation of AI into the insurance context has brought benefits in the area of fraud detection, pricing, underwriting, claim management and the customer journey. Among the various success stories recorded, the following emerge:

- Image Verification: The fraudulent ways in which third parties seek misleading gain are nowadays constantly evolving. To guarantee solidity in the recognition of such frauds, it is required the support of technologies like Computer Vision and Deep learning for image analysis and to identify counterfeits.
- Risk Engineering: As an example, with Natural Language Processing technology it is possible to read and analyze the risk reports that describe the structures to be insured, managing to identify the most suitable insurance premium, extracting the relevant information from documents of over 100 pages in a few seconds. It is estimated that the minimum reduction of the total process time is 50%, with an accuracy level of the risk analysis equal to 85%. [8]

AI is also effective during the policy review process. Not only it can potentially identify and detect overexposures, misalignments, and other warning signs, but it can also compare policies with others to identify potential loss areas. Another area of intervention of artificial intelligence is that of the preliminary analysis of the types of damage and the preparation of a forecast of possible compensation according to the damage in a manner to be put in reserve for consistent amounts, improving efficiency and reducing the risk of incurring sanctions.

3.5. Current Frauds Management System

This section contains the analysis carried out over the reports published by IVASS regarding the years 2019 and 2020 anti-fraud activities, with the objective of showing a complete overview of the processes already in place while also taking into consideration the effect that the 2020 lockdown had on overall fraudulent behaviors.
As briefly mentioned in the earlier sections, one of IVASS's duties is the contribution to the fight against fraud in the third-party liability car sector. It does so by analyzing and evaluating the information obtained from the management of the central claims database, interacting with companies regarding the recorded reports, and collaborating with the police forces and the judicial authorities to perform investigations.
It also manages a unified database of all registered claims to data, called **BDS**, along with an archive of past recorded frauds, characterized by specially developed KPIs that identify how potentially dangerous a claim could be. Companies or privates can consult both of these internal systems after the necessary authorization has been given by IVASS, maintaining a central distribution of authority, and all information is concentrated in a single entity. The BDS is powered directly by the information gathered from insurance companies, which are subjected to strict regulations regarding deadlines, and the cleanliness of data. It contains detailed, historical data about every claim closed in the last five years and can be accessed by ordinary citizens to consult their data after sending IVASS a request. After this period, the data is transferred to a different system, which can be accessed exclusively at the direct request of the direct data owner or for criminal justice needs.
The archive is called **AIA** and is strictly related to the fight against fraud. It is directly connected to the BDS, and other data sources external to IVASS, and can be interrogated by authorized users to generate values for various indicators to generate a final **Fraud Score** by comparing the new claims to the overall claims history recorded. An example of this functionality can be observed in Figure 3.5. The AIA authorization process is stricter than the BDS's, as the information contained within is much more sensible and is used directly by insurance companies for their activities.

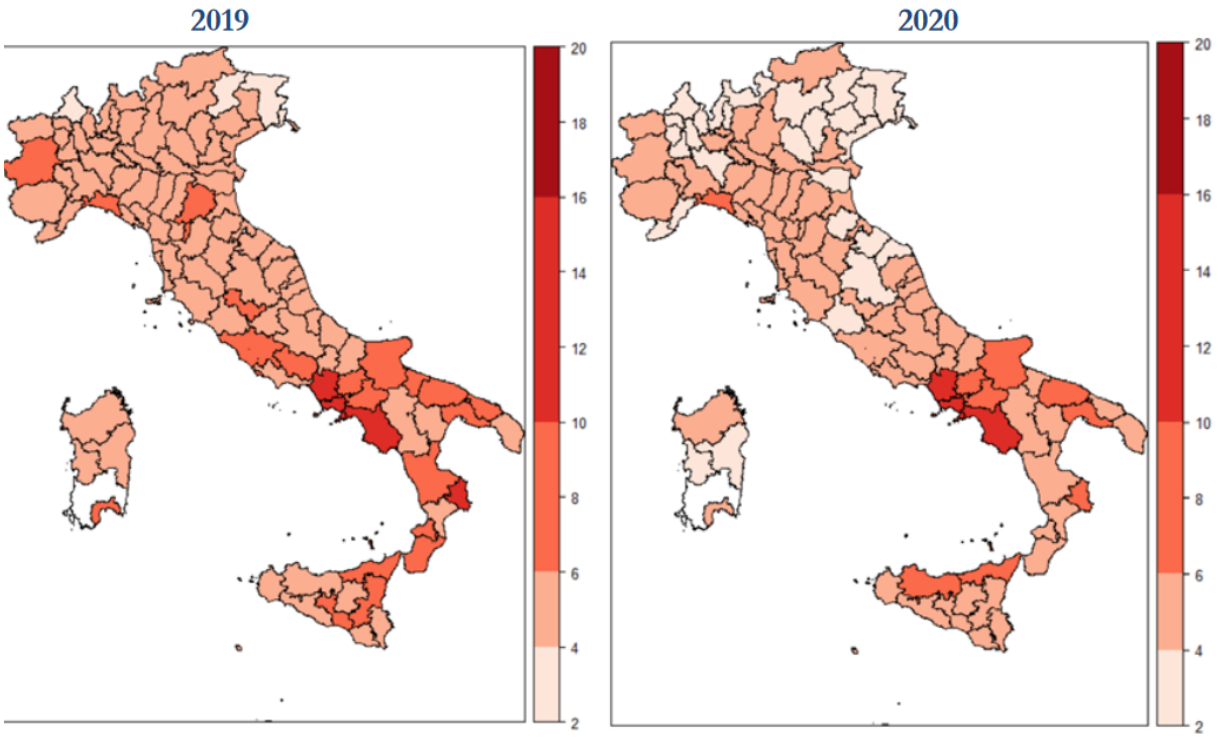

Figure 3.5: AIA's fraud scores geographical distribution.
Source: IVASS 2020 fraud report.

Figure 3.6 is the result of IVASS's yearly analysis: it displays the geographical distribution of the number of opened claims and claims exposed to fraud risk. It highlights that while the northern regions have the most managed claims, they also present the lowest percentage of possible frauds, in opposition to the statistics related to the southern regions [10, 11].

Regions	Claims Opened		Claims Opened Exposed to Fraud Risk		Potential Frauds Investigated		Potential Frauds Not Investigated	
	2020	2019	2020	2019	2020	2019	2020	2019
EMILIA ROMAGNA	146.708	209.804	20,6%	20,3%	10,1%	11,5%	13,9%	10,8%
FRIULI-VENEZIA GIULIA	30.899	43.336	18,7%	17,9%	6,9%	6,9%	15,0%	13,2%
LIGURIA	64.238	86.973	21,4%	23,3%	10,3%	11,5%	16,7%	12,8%
LOMBARDIA	336.117	481.558	18,2%	17,9%	8,0%	8,2%	14,8%	11,5%
PIEMONTE	149.759	221.134	21,8%	21,2%	9,5%	9,4%	15,4%	13,8%
TRENTINO ALTO ADIGE	41.775	75.907	25,7%	23,8%	6,3%	4,1%	18,7%	15,3%
VALLE D'AOSTA	3.858	5.801	18,3%	15,8%	8,1%	7,4%	22,9%	16,7%
VENETO	139.071	196.085	15,4%	16,1%	6,5%	7,0%	13,9%	11,0%
North Total	*912.425*	*1.320.598*	*19,3%*	*19,3%*	*8,4%*	*8,7%*	*14,9%*	*12,0%*
LAZIO	262.510	374.615	24,7%	23,5%	14,7%	13,4%	14,5%	12,0%
MARCHE	46.053	64.949	20,3%	18,7%	10,1%	8,9%	12,4%	11,5%
TOSCANA	138.258	198.895	20,4%	20,5%	10,1%	10,1%	13,6%	10,6%
UMBRIA	28.098	39.749	21,9%	22,0%	11,4%	10,2%	16,1%	12,7%
Center Total	*474.919*	*678.208*	*22,8%*	*22,1%*	*12,7%*	*11,8%*	*14,2%*	*11,7%*
ABRUZZO	37.059	52.230	22,7%	22,1%	10,2%	9,5%	17,8%	14,6%
BASILICATA	13.502	18.437	28,1%	26,7%	15,8%	14,8%	15,9%	15,0%
CALABRIA	42.394	57.155	33,6%	32,0%	21,8%	19,8%	16,2%	15,2%
CAMPANIA	194.353	259.743	53,1%	48,4%	37,9%	33,6%	17,7%	14,9%
MOLISE	8.549	11.647	39,0%	35,4%	25,4%	22,6%	18,6%	15,5%
PUGLIA	108.069	140.013	29,4%	27,8%	17,7%	16,6%	12,1%	12,7%
South Total	*403.926*	*539.225*	*40,8%*	*37,8%*	*27,3%*	*24,5%*	*16,6%*	*14,5%*
SARDEGNA	50.666	69.452	17,2%	16,6%	8,9%	8,4%	14,7%	12,2%
SICILIA	151.397	202.820	25,5%	25,5%	14,9%	14,8%	15,2%	12,6%
Islands Total	*202.063*	*272.272*	*23,4%*	*23,3%*	*13,4%*	*13,2%*	*15,1%*	*12,6%*
	1.993.333	2.810.303	24,9%	23,9%	13,7%	12,9%	15,4%	12,9%
Past Years	2018	2017	2018	2017	2018	2017	2018	2017
	2.813.191	2.857.883	22,3%	22,4%	13,3%	12,4%	14,9%	14,2%

Figure 3.6: Numerical analysis of claims and reported frauds per region.
Source: IVASS 2020 fraud report.

The latest available analysis shows that after considering all the claims managed in the two preceding years, 2019 and 2020, the total amount of capital the current system saved thanks to the fight against fraud was 254 and 248 million euros, respectively. Figure 3.7 highlights these results in more detail, also dividing the claims into different classes of risk, which are dependent on various indicators.

Claims risk classes and estimation of money saved due to antifraud activities

(Million of euros and Percentages)

Risk Classes	2019		2020	
	Amount	Market Share	Amount	Market Share
I	205,1	80,8%	161,5	75,2%
II	33,7	13,2%	38,2	6,9%
III	6,9	2,8%	41,6	10,2%
IV	4,3	1,7%	3,4	2,3%
V	3,7	1,5%	3,1	5,4%
Total	**253,7**	**100%**	**247,8**	**100%**

Figure 3.7: *Source: IVASS 2020 fraud report.*

3.6. State-of-the-Art Analysis

The **technical questions** that needed to be answered to obtain insight into the next steps to progress were mainly two:

- *Which AI techniques are used in fraud detection related to the financial sector?*
- *More generally, what approaches are used to estimate the likelihood of events occurring, like in our case for claims and frauds?*

We still want to study fraud detection methods as they could provide assistance with our case, even though the goal is different. The first result that can be obtained by analyzing the current literature is that AI is vastly more researched and implemented with use cases related to credit card fraud, rather than claim fraud. In the case of credit card fraud, the general approach is to treat this as an anomaly detection problem, using mainly data mining techniques to classify transactions and alert if any risky payment is detected. An example of this is [16] and more advanced versions that treat the incoming data as streaming and use a sliding window approach to also take into consideration the concept drift related to customers' behavior, like [20]. In contrast, systems that focus more on ML algorithms like [17] are less popular.
For the more specific case of insurance fraud, the typical approach is to treat it as a classification problem over the opened claims like [9] and [6], assigning labels to each claim instance. Obtaining labels for claims is a very costly and challenging process, especially in a field as complicated as insurance. To address the inefficiencies of supervised learning, in the last years more advanced methods employing neural networks have arisen as strategies [18].

By abstracting the research from the specific case of frauds and claims, seeing them only as 'events,' the focus shifted toward a particular type of research called **survival analysis**, also called **time-to-event analysis**. Survival analysis is a sub-field of statistics where the goal is to analyze and model data where the outcome is the time until an event of interest occurs.
Analytically, the function that represents the probability that the time to the event of interest is not earlier than a specified time t is called **survival function**, while the opposite function is called **cumulative death distribution function** and its derivative is the **death density function**. The last fundamental function is the **hazard functions**, also called instantaneous death rate: it represents the likelihood of the event occurring at time t given that no event has occurred before time t. The meaning behind these functions can be observed in Figure 3.8, and their mathematical expressions are:

$$S(T) = P(t \geq T) \textbf{ Survival Function.}$$

$$D(T) = P(t < T) = 1 - S(T) \quad \textbf{Cumulative Death Function.}$$

$$d(T) = \frac{d}{dt} D(T) \textbf{ Death Density Function.}$$

$$h(T) = \frac{\lim_{\Delta t \to 0} P(t \leq T < t + \Delta t)}{\Delta T} \textbf{ Hazard Function.}$$

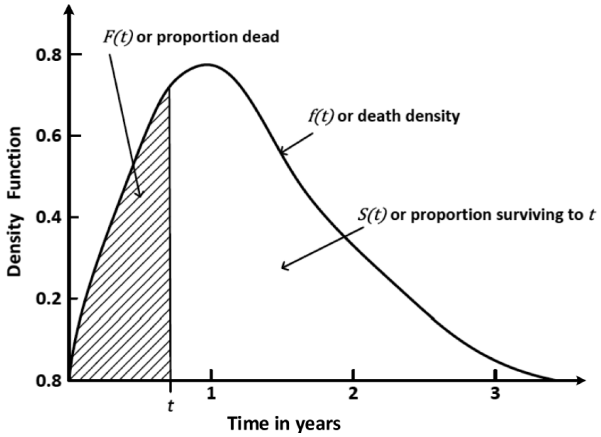

Figure 3.8: Graphical representation of the relation between a survival analysis relevant functions.

While most used in the medical sector, it can be applied to any type of event, making it useful in various fields: an example of this is [1], which describes its application in criminology for predicting the time until recidivism. An example of the fields where survival analysis can be implemented and its application can be seen in Table 3.1

Application	Event of interest	Estimation	Features
Healthcare (Miller Jr 2011) (Reddy and Li 2015)	Repeated hospitalizations Disease Cancer Survival	Likelihood of hospitalization within t days of discharge.	**Demographics**: age, gender, race. **Measurements**: height, weight, disease history, disease type, treatment, laboratory, procedures, medications.
Student Retention (Murtaugh et al. 1999) (Ameri et al. 2016)	Student dropout	Likelihood of a student dropping out within t days.	**Demographics**: age, gender, race. **Financial**: cash amount, income, scholarships. **Pre-enrollment**: high-school GPA, ACT scores, graduation age. **Enrollment**: transfer credits, college, major.
Customer Lifetime Value (Zeithaml et al. 2001) (Berger and Nast 1998)	Purchase behavior	Likelihood of a purchasing within t days.	**Customer**: age, gender, occupation, income, education, interests, purchase history. **Store/Online store**: location, customer reviews, customer service, price, quality, shipping fees and time, discounts.
Click Through Rate (Yin et al. 2013) (Barbieri et al. 2016)	User clicks	Likelihood of opening an advertisement within time t.	**User**: gender, age, occupation, interests, users click history. **Advertisement**: time of the ad, ads' location on a website, topic, format, and average clicks on it.

Table 3.1: Examples of real-world application domains for survival analysis.

A rather common occurrence in survival analysis problems is the event's time of occurrence not being observed for every instance. In the current case, it would correspond to clients not opening claims or committing fraud due to either the limited observation time window or missing traces caused by other uninterested events. This situation is called **censoring**: an illustration of this concept is shown in Figure 3.9.

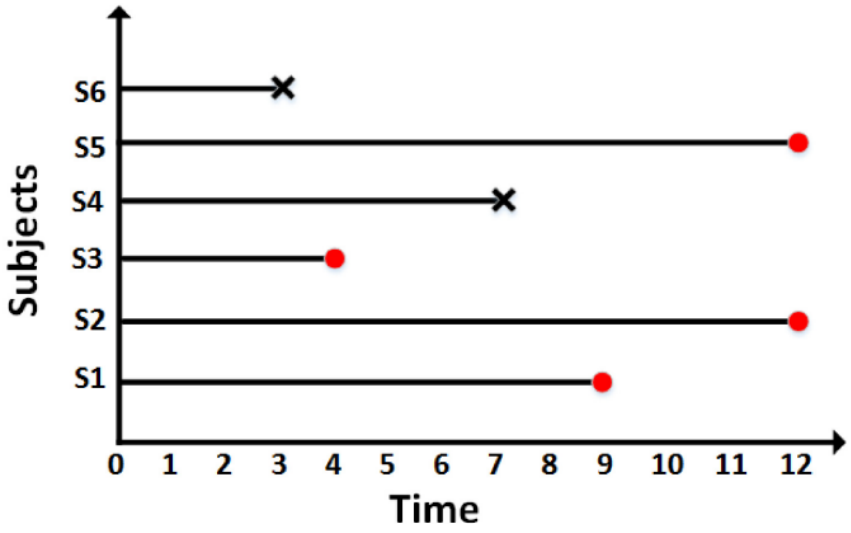

Figure 3.9: Graphical representation of the censoring problem in survival analysis.

In this case, only subjects S4 and S6 experienced the event during the observation windows, so it is possible to precisely calculate their survival. Everyone else either arrived at the end of the observation without encountering the event or their status became invalid at some point due to a variety of reasons, like a patient withdrawing from the clinical test. This particular case of censoring is the most common and is called **right-censoring**. In the case of insurance fraud, each instance could be modeled as a policyholder, their observation period is the duration of the policy, and the events that are used as reference are the claims and frauds. In addition, because of the nature of the events, it is not uncommon for a client to not open a claim or never commit fraud, arriving at the end of their policy without incurring the events. This situation fits perfectly with the concept of right-censoring.
Over the years, ML-based techniques have been developed to perform survival analyses, able to take into account the presence of censored data and model nonlinear relationships. Some of the most common algorithms are:

- **Survival Trees and Random Survival Forest**
  - Survival trees are classification and regression trees specifically tailored to han-

dle censored data, the primary difference with traditional decision trees being the splitting criteria. Decision trees perform recursive partitioning on the data by setting a threshold for each feature, while survival trees aim to minimize within-node heterogeneity or maximize between-nodes homogeneity. The RSF works very similarly to its traditional counterpart, the significant difference being how the final tree is constructed by taking the average Cumulative Hazard Function of each tree.

- **Naive Bayes and Bayesian Networks**
  - The experimental results of applying Bayesian methods to survival data show that these methods have excellent interpretability and uncertainty reasoning [21]. The Naive Bayes method is able to estimate various probabilities from the data, providing a link between the posterior probability and the prior probability, but implicitly assumes zero collinearity between the used features. A Bayesian Network allows the features to be related to each other at several different levels, and can graphically represent a theoretical distribution over a set of variables. Recently, [12] proposed a novel framework that combines its representations with an AFT model by extrapolating the prior probabilities to future time points.
- **Support Vector Machines and Support Vector Regression**
  - A first, naive approach is to consider only those instances that actually have events in the Support Vector Regression, ignoring the censored instances, and thus losing potentially useful information. To avoid this, [22] studied a learning machine designed for the predictive modeling of independently right-censored survival data by introducing a health index, which serves as a proxy between the instance's covariates and the outcome.
- **Gradient Boosting**
  - The boosting algorithm is one of the most widely used ensemble methods and is designed to combine base learners into a weighted sum that represents the final output of a strong learner. This algorithm iteratively fits a set of appropriately defined residuals based on the gradient descent algorithm.

There are also other techniques for survival analysis, statistics like Kaplan-Maier functions, Life Tables, and Cox Regression.
Due to the presence of censoring in survival data, standard evaluation metrics for regres-

sion like RMSE are not suitable for measuring performance.
A common way to evaluate a model is to consider the relative risk of an event occurring for different instances rather than the absolute survival times for each of them: this can be done by calculating the Concordance Index (C-Index) [7], which is a measure of the rank correlation between predicted risk scores and observed time points. Consider both the observations and prediction values for two instances, (x1, y1) and (x2, y2), where xi and yi represent the actual observation time and the predicted value, respectively. The concordance probability between them can be computed as:

$$C = P(y_1 \geq y_2 | x_1 \geq x_2).$$

While Harrell's concordance index is easy to interpret and compute, it is known to be overly optimistic with survival rates when the amount of censoring in the data is high. As a solution to this, it exists a variant of the C-Index that does not depend on the distribution of censoring times in the test data. Therefore, the estimate is unbiased and consistent for a population concordance measure that is free of censoring. It is based on the inverse probability of censoring weights, and thus requires access to survival times from the training data to estimate the censoring distribution.
Another valuable metric is the Brier Score [7], which can only be used to evaluate models with a probabilistic output, so a value is confined in the rage [0,1]. Given a time point t, it is defined as:

$$BS(t) = \frac{1}{N} \sum_{i=1}^{N} I(y_i \leq t | \delta = 1) \frac{(0 - \pi(t|\mathbf{x}_i)^2}{G(y_i)} + I(y_i > t) \frac{(1 - \pi(t|x_i)^2}{G(t)}.$$

Where the first function is the model's predicted probability of remaining event-free up to time point t for feature vector $\mathbf{x}$, and 1/G(t) is an inverse probability of censoring weight, which can be calculated via an estimator.

3.7. Ethical and Moral Guidelines

3.7.1. Global Landscape

These concerns sparked a heated debate over the definition of an 'ethical AI' and which requirements, technical standards, and best practices are needed for its realization. This concept revealed to be complex to define, as it is strictly related to the chosen definition of AI itself and the tasks that can be carried out, forcing the community to look at each case individually first and then try to formulate 'global' guidelines. Following this

methodology, [13] compiled the available information and defined eleven main principles that an ethical AI should conform to, each with its own set of constituent ethical issues or guidance related to it.
The four principles that were the most prevalent and highlighted when analyzing the literature were:

- **Transparency**
  - The concept of transparency encompasses all actions performed to increase explainability, interpretability, or other means of communication and disclosure, with the goal of fostering trust with third parties and legal entities. To achieve greater transparency, many sources suggest increased disclosure of information by those developing or deploying AI systems , although specifications regarding what should be communicated vary greatly: source code, what data is used and the reason, processes impacted, etc.
- **Justice, Equity, and Fairness**
  - Justice and fairness are expressed mainly by the prevention, monitoring, or mitigation of unwanted bias and discrimination due to the model structure or training data used. Whereas some sources define justice as the respect for diversity, inclusion, and equality, others focus on human accountability and on the actions necessary to ensure that the data being used does not contain errors and inaccuracies that will corrupt the response and decisions taken by the AI, and consequently by the humans use it.
- **Non-maleficence**
  - This concept refers to general calls for safety and security, asserting that AI should never cause foreseeable or unintentional harm. While more evident than other qualities, non-maleficence acquires particular importance in contexts like autonomous cars and drone technology, where physical harm could be involved. Another, more detailed explanation entails the explicit definition and avoidance of specific risks or potential harms, like intentional misuse via cyberwarfare and malicious hacking. Technical solutions include in-built data quality evaluations or security and privacy by design, while governance strategies include active cooperation across disciplines and stakeholders.
- **Responsibility and Accountability**
  - Responsibility and accountability aim to clarify the attribution of responsibility and legal liability to all parties involved. Not only are the developers

responsible for the architecture of a solution, but organizations, in general, have to be aware of the potential dangers of poorly thought solutions or biased data. To reach a satisfactory level of responsibility and accountability, there is a need for documentation, built-in record-keeping, or other tools to enable auditing.

While [13] presents an already detailed set of guidelines for the concept of ethical AI, others took it as a starting point and adapted it to more specific cases to propose a new set of guidelines, like [5] and [14], applying them to sectors like healthcare.

3.7.2. European Community Legislation

After getting an idea of the overall characteristics that an AI should possess to be classified as 'ethical', even if only from a theoretical standpoint, the next step consists in effectively looking into the current legislation that manages the usage of AI. In 2021, the European Union released a decree proposing a legal framework for laying down harmonized rules for trustworthy AIs(Artificial Intelligence Act). [3] The Commission aims to address the risks generated by specific uses of AI through a set of complementary, proportionate, and flexible rules, which will also provide Europe with a leading role in setting the global gold standard. To reach this objective, the Commission set four specific goals:

- Ensure that AI systems are fully compatible with existing laws and respect the values set by the Union.
- Ensure legal certainty to promote AI growth in the market.
- Enhance governance and effective enforcement of existing law on fundamental rights and safety requirements applicable to AI systems.
- Facilitate the development of a single market for lawful, safe, and trustworthy AI applications and prevent market fragmentation.

Before the publication of this document, an online public consultation was launched on February 2020 to gain insight into the point of view of the stakeholders involved with AI in the Union market. While the majority agreed on the existence of a legislative gap regarding AI, they also warned the Commission on the possibility and dangers of overregulation, effectively obstructing the market's growth potential. Many also underlined the importance of a technology-neutral and proportionate regulatory framework that defined fuzzy concepts like 'risks', 'biometric identification', 'real-time', and 'harm'.
The result was a framework that presented clear requirements and obligations regarding specific uses of AI with an easy-to-understand approach based on four different levels of

risk, depending on the field of application: unacceptable, high, limited, and minimal.

- Unacceptable risk
  - It includes all AI systems considered a clear threat to the safety, livelihoods, and rights of people, from social scoring by governments to toys using voice assistance that could encourage dangerous behavior.
- High risk
  - This classification level encompasses systems dealing with critical infrastructures, educational or vocational training, employment, management of workers, law enforcement, asylum and border control, administration of justice and democratic processes, etc.
- Limited risk
  - Limited risk refers to AI systems with specific transparency obligations such as chatbots, as users should be aware that they are interacting with a machine.
- Minimal risk
  - This level includes the vast majority of applications in today's market, such as AI-enabled video games or spam filters, where the impact that AI could have on users is negligible.

Out of the four risk classification levels, high-risk AIs require particular attention, as they are involved with complex sectors of society that require particular attention but could also provide the highest benefits if implemented properly. As such, they are subjected to a structured approval process to carry out all the necessary steps to check the solution's compliance, as shown in Figure 3.10

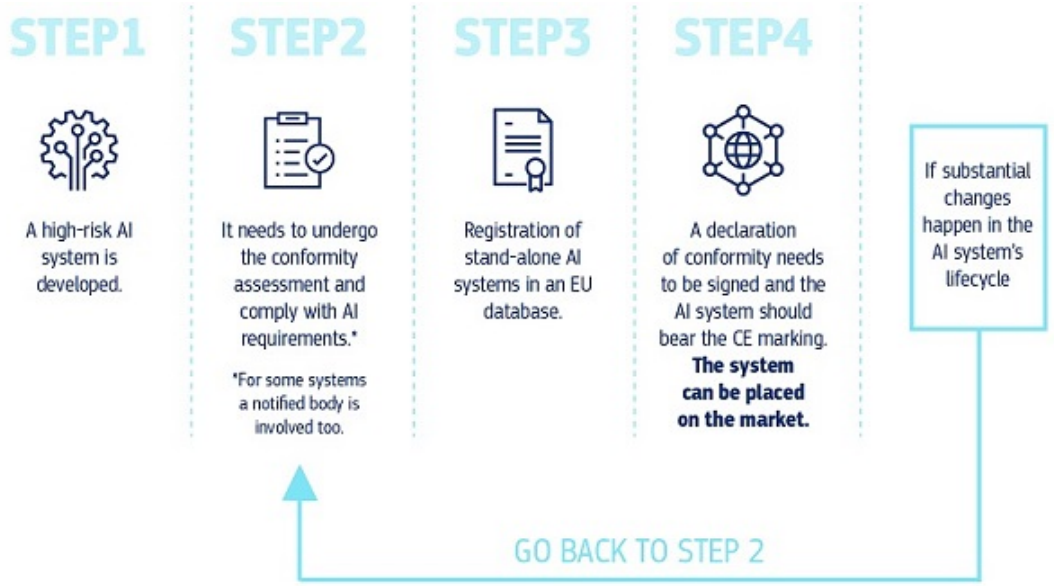

Figure 3.10: Review and approval process for high-risk AIs.
Source: European Commission proposal for a legal framework on AI.

3.8. Overall Findings

To summarize the results of the theoretical research done, the Italian insurance sector is characterized by strict regulations, all enforced by a central authority, IVASS, which coordinates the anti-fraud activities across all individual insurance companies. Even though Italian insurance companies have not widely implemented new technologies, and lag behind other countries, in 2021, IVASS implemented a system designed for the incubation and nurturing of technologically advanced proposals for the FinTech and InsurTech sectors, which represents a clear signal that there is a room for the implementation of new solutions.
Classical fraud detection is more involved with the classification of claims by employing various algorithms, from traditional to more advanced techniques, like deep neural networks.
By abstracting ourselves from the nature of the event, claims, and fraudulent claims, in this case, the research discovered that a survival analysis would be able to satisfy the business needs imposed by the customer, returning as output both a prediction on the time of the event occurrence and the survival rate at a specified moment.
Many Survival models are available, and almost all of them can manage censored data

without manipulating the source data, an important intrinsic characteristic of data related to claims and fraudulent claims.
Considering the scope of the project, the restraints on the time and resources available, and how well-documented these ML models are, the traditional ML approach was chosen as the go-to solution. In particular, the neural networks were set aside for two specific reasons:

- **Resources Available**: The degree of complexity, amount of input data needed for an effective training, and hardware resources necessary to sustain the usage of neural networks across the entire business, would have surpassed the resources available at the time this project was carried out.
- **Interpretability** and **Explainability**: The data at hand contains sensible information, and the model's results could influence the business actions towards their clients. The research performed over the requirements showed that the AIs developed for the defined goal have a high likelihood of being classified as high-risk. As the field of interest of this work requests for personal information to be processed, the degree of attention required to topics like BIAS and discrimination is very high. This means that there is a strong need for a model whose reasoning behind its predictions can be understood and explained, if requested, to prove that no discriminatory decision-making is carried out by the model when the information is used.

4 | Data Analysis and Feature Processing

In this chapter, the goal is to define the scope of the analysis by presenting the dataset and processing it to gain insight into its content. First, a selection of the most relevant features by using the analysis results was performed. Then, these base features were used to craft new ones, defining the final dataset to give as input to the model.

4.1. Dataset Definition

As the database at hand contains data related to the entire claim center managed by the insurance company, the information stored inside goes beyond what is necessary for The amount of data collected in the available database: a quick scan of the number of tables returns a value of over **two thousand tables**.
The first step was to reduce the number of tables to handle by removing the ones unrelated to claims or fraudulent claims, backup tables, and empty tables. This process was carried out by simply running SQL queries on the system tables and consulting the data dictionary shared by the client: it contains a high-level explanation of the fields of each table included in the claim center database. This purging process allowed us to arrive at around **one thousand tables** potentially of interest.
As the number of tables was still too high to define a clear dataset, a reverse-engineering approach was instead adopted.

- First, an approximate list of potentially useful characteristics and predictors for each of the two tasks was created.
  - Examples would be the age at claim/fraud time, number of vehicles insured, number of past claims/frauds, number of past insurances, time between accident and report, etc.
- Then, the database's system tables were queried to find the tables with the most number of records.

- Tables containing information about clients and claims must have a considerable amount of information stored inside, and should also be the central point of the schema. From them, additional information could be retrieved using these links, which are represented in the form of foreign keys.

The table list was used as input to search for these desired proprieties. This allowed the determination of which tables were more relevant, without getting distracted by the amount of data available.

The final dataset was determined by joining the discovered tables to obtain a single, central entity and performing feature selection and creation to define the desired qualities.

4.2. Exploratory Data Analysis

4.2.1. Individual Tables

The process mentioned above concluded with identifying three main tables that contained the majority of the previously identified information, which can be seen in Table 4.1. A first examination of the three tables revealed that most of their features were keys used to connect them with other tables or fields obfuscated for privacy reasons. Their initial exclusion reduced the total number of relevant features of each table to less than twenty for each table.

Table Name	Row Count	Description	Features Count
PDW_CLAIM	1.831.918	Contains information about all registered claims and indications of their potential fraudulence.	13
PDW_POLICY	2.075.698	Contains information about all registered policies, both active and expired.	8
PDW_PERSONAL_INFO	7.542.973	Contains information about all registered clients.	15

Table 4.1: Target tables descriptions and basic information.

The reason behind the big difference in the stored records between the claim and policies tables and the table containing personal information was the source. The first two tables came from the claim database shared by the client, while the other came from the data

warehouse, which contains information about all clients, even those not present in the shared data. The first step was to look closely at the actual data inside these tables to understand not only what values are inside, but also the informative content of each field. The objective was to make use of this information to select the features that the final dataset would be composed of and also to create new features to enrich the dataset.
The section below contains a high-level description of each table's fields and the number of null and distinct values.

Feature name	Description
CD_CLAIM	Incremental identifier of an individual claim. It changes if the claim is reopened.
CD_CLAIM_NUMBER	Identification number of an individual claim. It does not change if the claim is reopened.
CD_POLICY	Incremental identifier of an individual policy. It changes if the policy is renewed.
DT_CLAIM_LOSS	Date the claim was reported to have occurred.
DT_CLAIM_REPORTED	Date the claim was reported to the insurance company.
DT_CLAIM_CREATION	Date the claim record was created in the database.
DT_CLAIM_CLOSED	Date the claim was closed.
DAYS_DIFF_LOSS	Difference in days between the claim occurrence date and the claim report date.
CD_FAULTRATING	Categorical field defining the amount of fault attributed to the insured entity.
FL_TESTIMONI	Field expressing the presence or not of witnesses of the reported claim.
FL_RITROV_VEIC	Field expressing the discovery of the insured vehicle after a theft.
CD_LOSSCAUSE	Categorical field defining the type of loss.
CD_ACCIDENT_TYPE	Categorical field defining the type of accident.
FL_FRAUD	Field expressing if the claim was found out to have been a fraud.
AIA_SCORE	Numerical score assigned by AIA.

Table 4.2: Claim table overview.

Feature Name	Feature Type	Distinct Values	Null Values
CD_CLAIM	ID	1.831.918	0
CD_CLAIM_NUMBER	ID	1.831.918	0
CD_POLICY	ID	1.831.918	0
DT_CLAIM_LOSS	Date	6.783	0
DT_CLAIM_REPORTED	Date	6.361	0
DT_CLAIM_CREATION	Date	6.361	0
DT_CLAIM_CLOSED	Date	4.705	0
DAYS_DIFF_LOSS	Integer	1.622	0
CD_FAULTRATING	Categorical	11	628.881
FL_TESTIMONI	Flag(Integer)	2	1.530.288
FL_RITROV_VEIC	Flag(Integer)	2	0
CD_LOSSCAUSE	Flag(Integer)	23	0
CD_ACCIDENT_TYPE	Categorical	8	1.740.948
FL_FRAUD	Flag	2	0
AIA_SCORE	Integer	22	0

Table 4.3: Claim table analysis.

Feature Name	Feature type	Distinct Values	Null Values
BIRTH_DT	Date	29.119	105.888
PARTY_TYPE_CD	Categorical	3	0
GENDER_CD	Categorical	3	0
MARITAL_STATUS_CD	Categorical	8	0
OCCPTN_CD	Categorical	24	0
Z_PROFESSION_CD	Categorical	24	0
CITY_NAME	Categorical	10.491	0
CNTRY_CD	Categorical	3	0
COUNTY_NAME	Categorical	115	0
POSTAL_CD	Categorical	8.652	0
Z_ADDR_TYPE_CD	Categorical	4	0
Z_PROVINCIA	Categorical	218	0
Z_TOPONYM_CD	Categorical	69	0
POL_NO	Integer	1.394.321	0
Z_VRT_TERM_NO	Integer	22	0

Table 4.7: Personal information table analysis.

Feature Name	Description
CD_POLICY	Incremental identifier of an individual policy. It changes if the policy is renewed.
CD_POLICY_NUMBER	Incremental identifier of an individual policy. It does not change if the policy is renewed.
CD_POLICY_TYPE	Categorical field describing the type of policy.
DT_POLICY_EFFECT	Date of the start of the policy.
DT_POLICY_EXPIRATION	Date of the expiration of the policy.
CD_RISKUNIT	Categorical field describing the category of risk unit.
CD_RISKUNIT_TYPE	Categorical field describing the sub-category of risk unit.
VEHICLE_MANUFACTURER	Field containing the name of the insured vehicle manufacturer.
NM_INSURED_VALUE	Field containing the amount the related vehicle is insured for.

Table 4.4: Policy table overview.

Feature Name	Feature Type	Distinct Values	Null Values
CD_POLICY	ID	2.075.698	0
CD_POLICY_NUMBER	ID	1.557.751	0
CD_POLICY_TYPE	Categorical	1	0
DT_POLICY_EFFECT	Date	6.939	0
DT_POLICY_EXPIRATION	Date	8.734	0
CD_RISKUNIT	Categorical	2.075.698	0
CD_RISKUNIT_TYPE	Categorical	1	0
VEHICLE_MANUFACTURER	String	923	4.951
NM_INSURED_VALUE	Integer	17.832	0

Table 4.5: Policy table analysis.

This process immediately highlighted how the Policy table possesses some features of cardinality equal to one or that do not bring any additional information. In particular:

1. CD_POLICY_TYPE is a categorical feature of cardinality equal to one. As the value is static, not only its informative content is zero, but it could also impact negatively the model to be implemented. *As such, this field was discarded.*
2. CD_RISKUNIT is a code that, by itself, does not communicate anything about the

Feature Name	Description
BIRTH_DT	Birth Date of the related entity.
PARTY_TYPE_CD	Categorical field describing the type of entity.
GENDER_CD	Categorical field describing the gender of the related entity.
MARITAL_STATUS_CD	Categorical field describing the gender of the related entity.
OCCPTN_CD	Categorical field describing the work occupation of the related entity.
Z_PROFESSION_CD	Categorical field describing the profession of the related entity.
CITY_NAME	Categorical field describing the city where the related entity is located.
CNTRY_CD	Categorical field describing the code of the related entity's country.
COUNTY_NAME	Categorical field describing the county of the related entity.
POSTAL_CD	Categorical field describing the postal code of the related entity.
Z_ADDR_TYPE_CD	Categorical field describing the type of address of the related entity.
Z_PROVINCIA	Categorical field describing the district of the related entity.
Z_TOPONYM_CD	Categorical field describing the code of toponym of the related entity. It augments its location.
POL_NO Z_VRT_TERM_NO	Fields containing information about the policy number of the related entity.

Table 4.6: Personal information table overview.

policy or associated vehicle. *As no additional data was available on the risk units, this field was discarded.*

3. CD_RISKUNIT_TYPE is a categorical feature of cardinality equal to one. As the value is static, not only its informative content is zero, but it could also impact negatively the model to be implemented. *As such, this field was also discarded.*

4. VEHICLE_MANUFACTURER can be considered as a categorical field. While its null count is not particularly high, its distinct count reveals issues: as it should contain data about a vehicle brand, not its model, the maximum amount of brands should not be as high as the analysis uncovered.
   As such, this field required additional investigation.

In addition, the data available in the Claim and Personal-Information tables was rarely fully complete, as some features were characterized by a high null count. For each of these cases, a specific strategy was adopted in order to handle this situation.

1. For the Claim table
    - CD_FAULTRATING was characterized by 34,3% of null values of the table total row count. The values that this field expressed are related to the percentage of fault attributed to the insured party, from 0% to 100%: the easiest way to deal with its missing values was simply to add another category identifying an unknown fault percentage.
    - FL_TESTIMONI was characterized by 83.53% of null values of the table total row count. As the only allowed values for this field were 0 or 1, *the solution chosen was to simply map all the nulls with a '2', indicating that the presence of witnesses was unknown.*
    - CD_ACCIDENT_TYPE was characterized by 95.03% of null values of the table total row count, nearly all the table's records. This field could provide additional utility when read together with the loss cause, which instead happened to always not have any null values. *As such, this field is discarded.*
2. For the Personal-Information table
    - BIRTH_DT was characterized by 1.4% of null values of the table total row count, meaning that the number of rows with a missing value was extremely small. *Considering that the birth date is not a field that can be deduced by other values and that few rows are impacted, the choice this time was to delete the impacted records.*
    - OCCPTN_CD and Z_PROFESSION_CD not only had very similar information content but were also characterized by the exact same number of categories. *An additional analysis was deemed necessary to understand if these fields were different from each other, and how often.*
    - POL_NO and Z_VRT_TERM_NO were used only to generate the corresponding entity policy code, *so they were ignored when defining the actual dataset.*

The analysis of the fields that raised suspicions was performed by investigating the data directly on the database through SQL queries, which proved to be the most powerful for the environment the data was in.

For the Policy table, it was fundamental to understand the reason behind the high cardinality of the 'brand' field. A quick select revealed that both brands and models were mixed in, generating many different combinations, as seen in Figure 4.1 explaining the messy situation that the first query highlighted.
As the data contained was too dirty to be used and no mapping was available to convert each model to its own brand, the feature was discarded.

```
SELECT DISTINCT(VEHICLE_MANIFACTURER)
FROM PDW_POLICY;
```

	VEHICLE_MANIFACTURER
43	SANTA FE 2.2
44	VAUXHALL
45	KA PLUS 1.2
46	C1 1.0 5P. C
47	TOURAN 1.9 T
48	PEGEOUT
50	RX 400H AMBA
51	ASV
52	ATALA
53	MONTESA
54	CH RACING
55	DODGE-CHRYSLER

Figure 4.1: Different manufacturers values.

For the Personal Information table, the check carried out was about seeing what was the difference between the two fields. The queries run revealed that, while they are never null, as the previous check showed, blank values are instead present, and that the same categories were expressed by the two fields. The only difference between them was that Z_PROFESSION_CD was more often not null, making it a more meaningful field: f*or this reason, the feature OCCPTN_CD was dropped.*

4.2.2. Final Dataset

With all the results obtained from this initial analysis, we could finally create the actual dataset by joining together the three tables, as described in Table 4.8

Table Name	Features Number	Censored Records	Not Censored Records
Final_Dataset	28	61.295	224.396

Table 4.8: Final dataset overview.

With all the information concentrated in one table, it was now possible to properly carry out an analysis of the distribution of certain properties of the available data.
The first course of action was to visualize the number of safe claims and fraudulent claims,

as they virtually represent different events and will need to be handled separately when later implementing the models. As the vast majority of the features are categorical, there was a need to visualize their distribution, assigning each category a meaning, to help make sense of the model predictions.

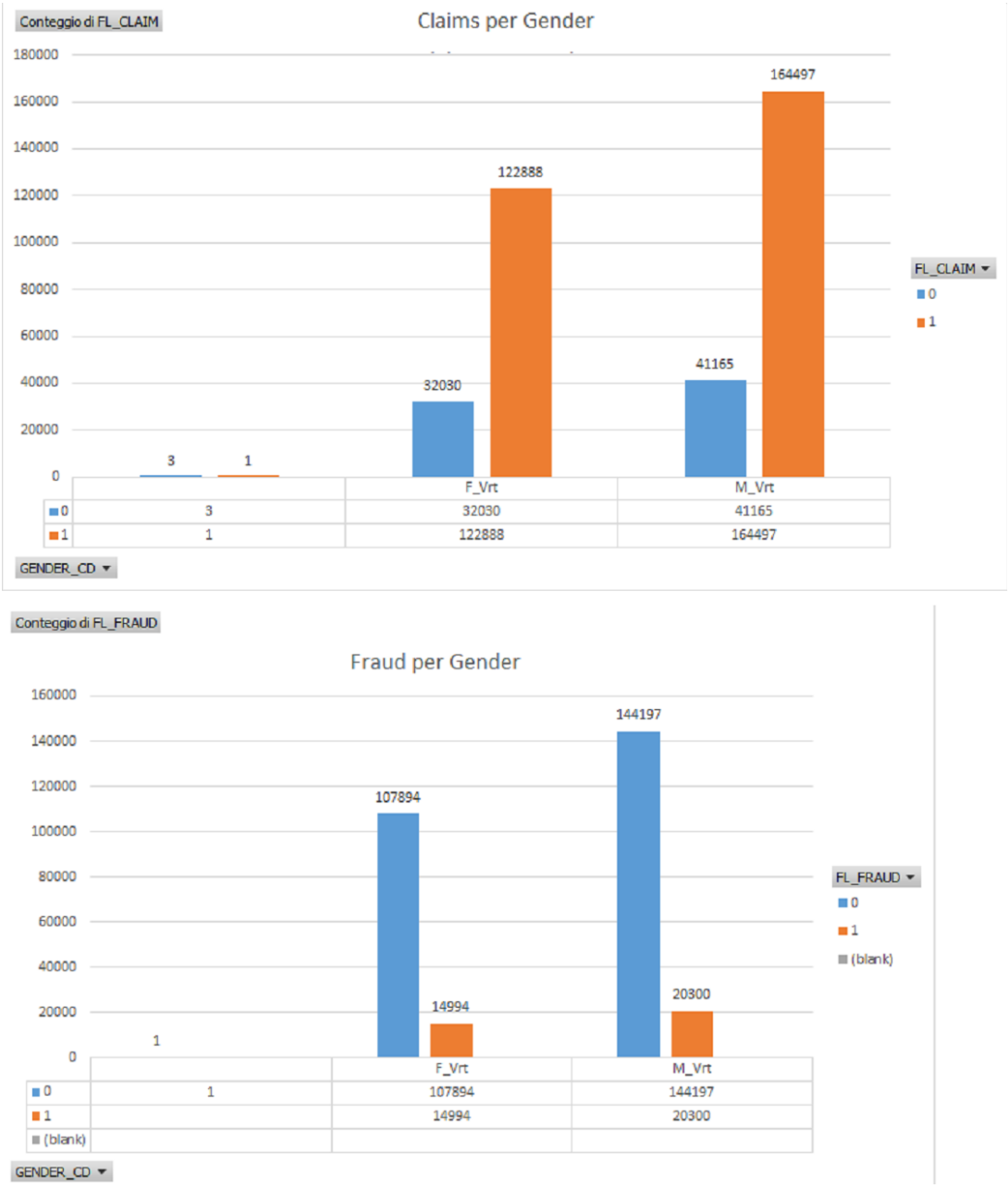

Figure 4.2: Graphical representation of the distribution of claims and frauds against customers' gender.

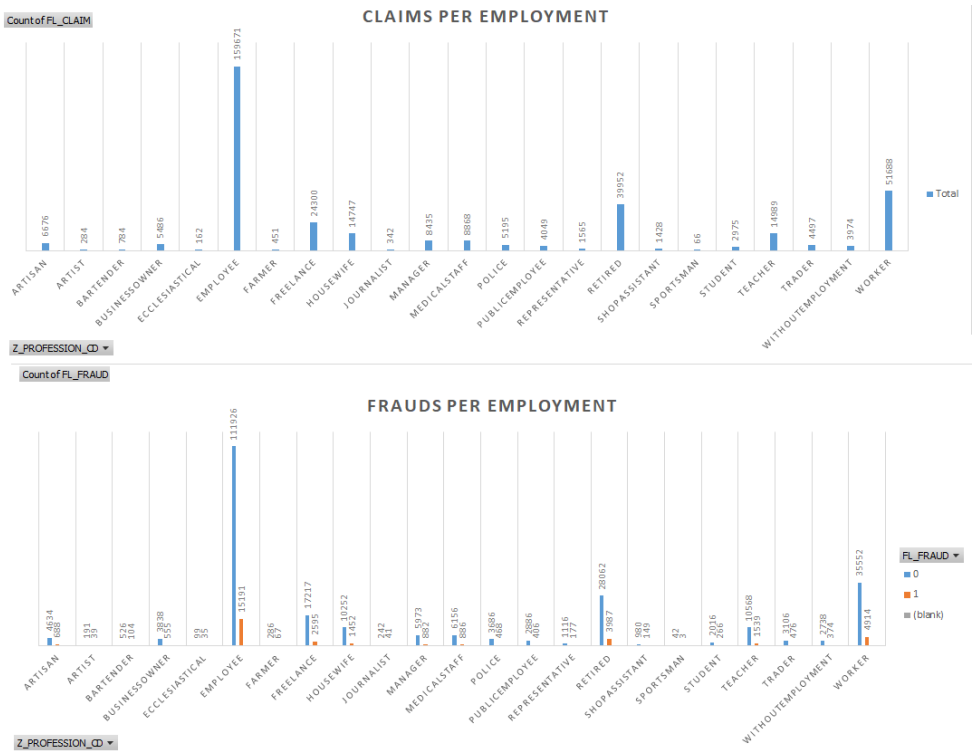

Figure 4.3: Graphical representation of the distribution of claims and frauds against customers' work occupation.

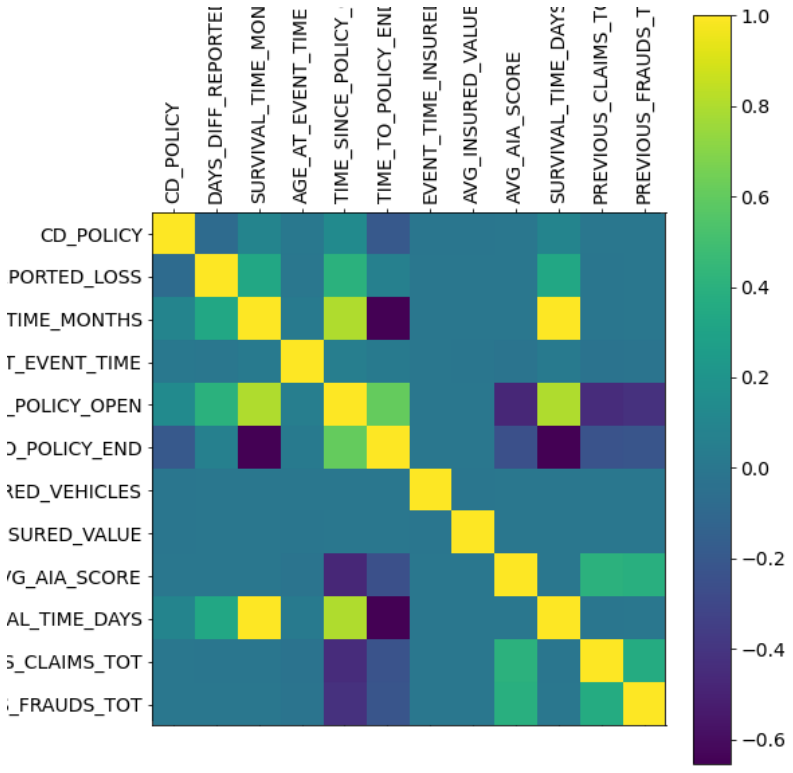

Figure 4.4: Graphical representation of the correlation between the dataset's features.

From the above plots, we can see that the occurrences of claims and frauds are even across the two genders, but more concentrated on people who work as employees, who show higher claims and frauds committed than any other occupation, while people with no job display among the lowest values, against a possible first intuition. The correlation matrix shows that there seems to be some correlation between the average AIA score and the previous number of claims/fraud, as could be expected.

4.3. Feature Engineering

The last step before going into defining and implementing the actual models was to check if any useful feature could be crafted starting from the available ones.

1. A proper survival time was not defined, meaning that the value necessary for any survival model to work was missing. As such, it was defined as 'time an individual has been a client for'.
    - For censored records, this value can be set to be the total length of the observation interval, so the difference between the moment an entity became a client and the policy expiration.
2. An interesting quantity would also be the number of vehicles insured by a single entity at claim/fraud time.
3. Rather than the birth date, the age when the claim/fraud occurs is a more interesting quantity.
    - For censored records, this quantity can be calculated as the entity's age in the current year.
4. If instead of considering the time the entity became a client we use as reference the policy start and end times, we can also define two new intervals.
    - For censored records, the quantities cannot be calculated, as the related claims' times do not exist.
5. For claims and frauds we can also calculate the average value of the insured vehicles.
6. For only the frauds, it would make sense to calculate the average AIA score of previous claims. For entities with no previous claims, this value is set to one, the lowest valid score.
7. At last, we can also count the total number of previous claims and fraud committed. If we do not have any previous claim or fraud, we set the counters to zero.

The detail of the queries defined to create the above-mentioned features are contained in the appendix.

4.3.1. Data Scarcity

After having finished crafting additional features, one characteristic of the dataset became obvious: the near absence of information related to fraud. This situation occurred due to limitations imposed by the business on the data that could be shared for a not 'official project'. As such, even before looking at any model, the expectation is that the predictions for frauds would not be significantly accurate, and their survival curve would resemble a lot the claims' curve.

5 | Models Implementation

This chapter will go over the implementation process of the models to reach the goals defined at the start. The starting point consisted in defining a baseline, a very simple model to determine the lower threshold for the performance in the given task. Then, survival models were implemented and compared to the baseline, showing the difference in performances.

5.1. Feature Selection

After the end of the engineering process, the features chosen to be used for the models were:

CD_FAULTRATING	Z_PROFESSION_CD
SURVIVAL_TIME_DAYS	COUNTY_NAME
GENDER_CD	AGE_AT_EVENT_TIME
MARITAL_STATUS_CD	TIME_SINCE_POLICY_OPEN
TIME_TO_POLICY_END	PREVIOUS_CLAIMS_TOT
EVENT_TIME_INSURED_VEHICLES	PREVIOUS_FRAUDS_TOT
AVG_INSURED_VALUE	AVG_AIA_SCORE

From this list, it can be observed that out of all the geographical information, only the county name was selected. The reason behind this choice was that topography, city name, postal code, and address type had a very high number of possible values, which could have led to the models being unable to make proper use of them, producing worse results.

5.2. First Implementation

The analysis performed of the problem and the SOA highlighted how survival models were the models most likely to be suitable to reach the set goals: *estimating the probability of opening a claim or committing fraud at a certain point in time.*

To confirm the goodness of this intuition, the chosen course of action was to look at the problem from another angle, framing it differently and checking if other techniques could produce acceptable results.
A quick scan of the entire dataset showed that each instance could be divided into one of three categories:

- User opened a claim.
- User committed fraud.
- User did nothing.

By defining the time as an input feature and translating these cases into labels, it was possible to frame this situation as a classification problem, where instead of the labels the focus is on the probability of belonging to the claim/fraud class.

5.2.1. AutoML

One of the business requirements was to employ one of the tools the client had available, instead of directly implementing the models directly by coding their structure.
In order to quickly design a classification model, one of the services of the OML4PY platform was employed. **AutoML** enables users to quickly implement ML regression or classification models by automatically selecting the best-performing algorithms based on the chosen dataset and metrics. It allowed the testing of multiple algorithms at the same time, while also performing its own selection features based on the input data.
This phase was run with two metrics: *weighted recall* and *f1 macro*. Weighted recall allowed the model to place more importance on correctly recognizing high probabilities of claims and fraud, even at the cost of false positives. On the other hand, the f1 macro metric tries to balance precision and recall without assigning any weight to the different classes.
In both cases, AutoML found the *Naive Bayes* algorithm to be the best performing one: this was not a surprise, as the correlation matrix revealed very little relation between the features, so the algorithm implicit assumptions worked well for the current use-case.
The models based on the two metrics both performed reasonably well, scoring **88%** on the test set, but a closer look at their confusion matrix revealed their limitations.

CLASSIFIER_LABEL	count_(Claim)	count_(Fraud)	count_(No Claim or Frauds)
Claim	154173	11	131
Fraud	25078	11	22
No Claim or Frauds	140	0	48852

Figure 5.1: Confusion matrix of the f1 macro classification model.

The model based on the f1 macro metric confused 'Claim' and 'Fraud' the vast majority of the time, while the one based on recall failed to completely recognize the 'Fraud' class, meaning it always assigned a low probability of fraud to every sample. This situation can be attributed to three main reasons:

1. From the start, classification techniques were not a good fit for the problem at hand, with its high complexity, dependency on time, and presence of censored data.
2. The information on frauds specifically was insufficient, as every available feature characterized both claims and frauds.
3. The number of labeled frauds is also very low, especially when compared to claims (190K vs 30K).

These combined issues likely caused the model to incorrectly recognize fraud events, the most important events for the business. Hence, this proved to be a significant limitation of the currently designed benchmark model.

5.3. Survival Model

Survival models are closely related to the concept of time, as their objective is to estimate the survival rate over the observation time. As such, instead of dividing the datasets into training and test subsets using the traditional splits, like 70-30 or 80-20, the approach chosen was to define them according to a division mainly based on time while maintaining a ratio close to 70-30 to keep balance.
The training sets gathered data for up to 2018, while the samples related to the remaining years were included in the test sets, creating a time separation. The objective was to develop a model that, after being trained on 'old data,' could perform well on data from recent years, learning an optimal survival function.

5.3.1. Baseline definition: Kaplan-Maier Estimator

The implementation of a survival model required a different approach in the division of the dataset, as the events that needed to be addressed were two, claims and frauds. For

this reason, the dataset was separated into two subsets to keep them distinct:

- Claim data + Censored data
- Fraud data + Censored data

From the research carried out on the SOA, the easiest model that could be implemented to get an idea of the structure of the survival functions of the claim and fraud events was the **Kaplan-Meier estimator**, which can calculate the survival probabilities and plot the respective survival function based on a target variable. The estimators obtained from this model were then immediately used to plot the survival functions of both frauds and claims based on a few chosen features, in this case, gender and some selected AIA scores, as depicted in Figure 5.2 and Figure 5.3.

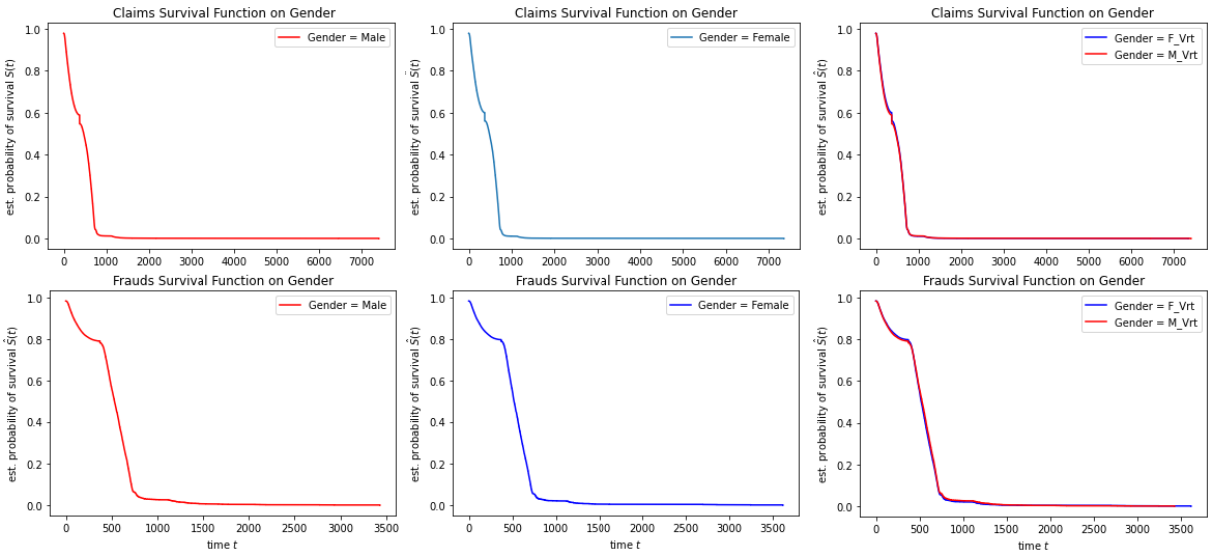

Figure 5.2: Survival function calculated by the KME on the customers' genders.

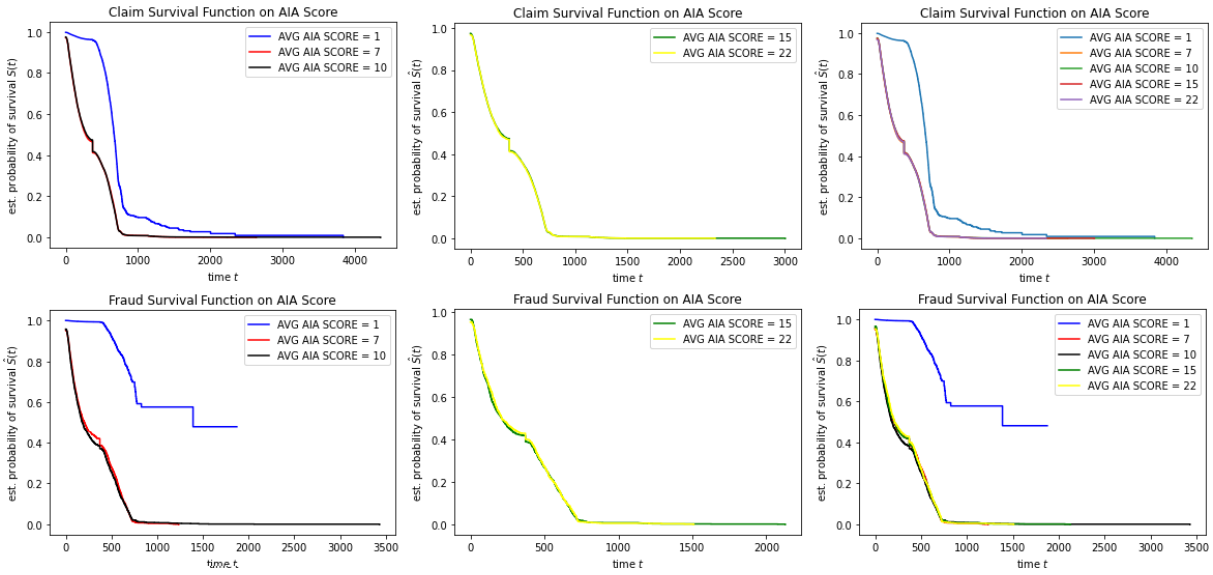

Figure 5.3: Survival function calculated by the KME on the customers' average AIA score.

As shown above, while the gender did not influence at any level the survival rate for both claims and frauds, the average AIA score instead had a notable impact.
While this model looked to provide the results that were needed, its simplicity came with a series of limitations that hindered its effectiveness: KME struggles to consider the effect of multiple features on the survival time at the same time, making it unsuitable for cases when many features are involved.
Before moving on to another, more appropriate model, due to the inability of a KME to consider multiple features simultaneously, each curve's performance had to be assessed with the C-Index calculated on the test set. This process was necessary to comprehend how well the simplest model performed through actual, concrete results, defining a lower threshold for the performance, and setting a baseline as a comparison with all future implementations.
The testing process revealed that no individual feature could successfully capture the underlying survival function with an acceptable degree of precision, as the C-Index computed in each case was particularly low. Consequently, the average of the individually calculated C-Index values could be considered representative of the model's overall predictive power. These results are displayed in Table 5.1 and Table 5.2.

Feature for KME	C-Index Claim Dataset	C-Index Fraud Dataset
CD_FAULTRATING	0,3297	0,2621
Z_PROFESSION_CD	0,2361	0,20,1
COUNTY_NAME	0,1711	0,1231
GENDER_CD	0,1886	0,1153
AGE_AT_EVENT_TIME	0,2511	0,2235
MARITAL_STATUS_CD	0,10,0,	0,0,18
TIME_SINCE_POLICY_OPEN	0,1735	0,1179
TIME_TO_POLICY_END	0,2120,	0,1260,
PREVIOUS_CLAIMS_TOT	0,3888	0,3331
EVENT_TIME_INSURED_VEHICLES	0,2768	0,2460,
PREVIOUS_FRAUDS_TOT	0,3413	0,3297
AVG_INSURED_VALUE	0,4335	0,40,5
AVG_AIA_SCORE	0,4868	0,4776

Table 5.1: KME C-Index score for each feature.

Average C-Index Claim Dataset	Average C-Index Fraud Dataset
0,2766	0,2343

Table 5.2: KME average C-Index score for the datasets.

5.3.2. Cox Model

Cox's proportional hazards model provides a way to estimate survival and cumulative hazard function in the presence of additional covariates. This is possible because it assumes the existence of baseline hazard function and that covariates change the "risk" (hazard) only proportionally. In other words, it assumes that the ratio of the hazard of experiencing an event of two patients remains constant over time.
The fitting process to obtain an estimator was performed exactly like any other model in the traditional scikit library, and it quickly produced the first version of the searched estimator. With the model ready, random samples from both datasets were taken as an example to plot the calculated survival functions, which are displayed in Figure 5.4.

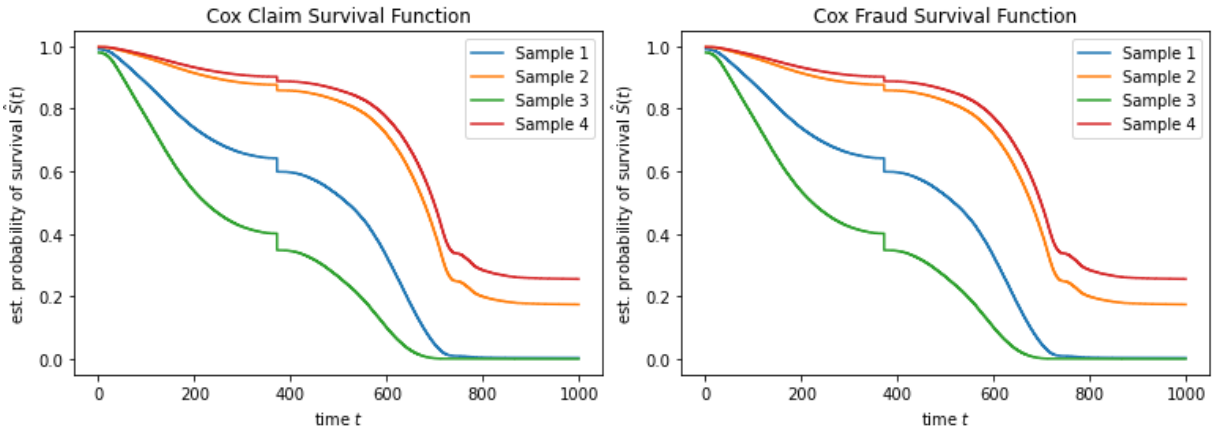

Figure 5.4: Survival functions calculated by the Cox model for both claims and frauds

As shown in the figure above, the survival functions calculated with the Cox models are smoother than the ones produced by the KME, which presented functions that were less defined and very quickly converged toward low values.
To assess the quality of the models, the next step was to calculate their performance on the testing dataset by employing the C-Index. The C-Index represents the global assessment of the model discrimination power, and its ability to correctly provide a reliable ranking of the survival times based on the individual risk scores.

C-Index Claim Dataset	C-Index Fraud Dataset
0,8263	0,9037

Table 5.3: Cox C-Index score for the datasets.

The results displayed in Table 5.3 show that the model performed reasonably well in both cases, especially when predicting fraud, which is the most important event to consider for the business.

5.4. Results Analysis: Performance Comparison

At the end of the entire implementation process, we can now properly compare all the models implemented and reach our conclusions.

5.4.1. Classification Model vs Survival Models

The first implemented model was a classification model: the objective was to see if a more traditional approach could be suitable for the problem at hand, instead of immediately

pursuing a survival analysis.
The results obtained highlighted that classification was not suitable as an approach. Two main reasons can be given to support this statement:

- Censored data handling
  - In the current use case, instances of clients that have never opened a claim or committed fraud correspond to censored data, meaning they do not possess a quantifiable survival time, as no event occurred. As previously stated, the nature of the problem imposes an expected high level of censoring, especially for frauds, since the number of people guilty of fraud is significantly lower than the opposite. While survival models can handle them natively and can also be tuned accordingly to the amount of censoring, a classification model has no way of doing so. In this last case, the survival time for these records would have to be left empty, making it unusable for training, essentially discarding the majority of the data.
- Risk function definition
  - The results obtained by the classification model express the likelihood of opening a claim or committing fraud at a specific point in time, defined by the time given as input. It fails to capture the function that determines those events and is incapable of displaying the variations of the risk values over time. On the other hand, survival models can output the survival rate for any given time and also plot the survival function. This is an advantage for a business, which would be able to better interpret the results and plan their actions accordingly.

5.4.2. Kaplan-Maier Model vs Cox Model

The KME was the chosen model to act as a baseline, providing the results needed to set a reference for the Cox model. While easy to implement and understand, the limitations shown by the KME, namely the inability of handling more features at once, represent a significant impediment. The dataset used was characterized by few features, but considering the amount of missing information, a more realistic and well-constructed dataset would have many more features. Consequently, this would lower even more the performance of the KME.
On the other hand, the Cox model was able to make use of all the features available simultaneously to define the survival function, achieving good performances for both events, which far surpassed the baseline results.

6 | Conclusions

6.1. Final Observations and Improvement Points

The research carried out revealed that the insurance sector has much potential for additional growth, but also its high complexity due to the high amount of regulations in place and the restrictions imposed by authorities on AI implementation.
The survival analysis proved to be an effective and interesting technique for the car insurance field, as proved by the results obtained from the models explored. The described solution can be improved much further if we take into consideration the factors that had to be ignored during the project duration, due to incompatibilities with the thesis' scope.

- First of all, the data that was shared to develop a model contained many dirty records. In addition, it was not possible to discuss in detail with the client the informative content of each table, forcing us to manually inspect them.
  - The addition of **domain knowledge expert** would improve drastically the effectiveness of the data selection process and the quality of the business-related goals.
- Almost all of the geographical features were ignored, mainly due to having no way of processing and making use of them properly.
  - On this side a lot of work can be done: the implementation of a **structured pipeline** and a **GIS** would enable the analysis of such information, generating new features that could be used by the models in the learning process.
- Due to business restrictions, we were also not able to integrate data coming from the internal fraud detection tool, which would have provided more information on fraudulent behavior.
  - This would enrich the historical data of the past claims, giving the survival model an additional valuable feature to make use of.

The only proper survival model that was explored was the Cox Proportional model. This was because all other models do not rely on a proportional hazards assumption, hence the score obtained by their predictions are not probabilities, but are of arbitrary scale. As such, without further research, models like the survival random forest were judged to be not adequate for the current use case. However, when prediction performance is the main objective, more sophisticated, non-linear, or ensemble models might lead to better results.

These approaches will be explored in more detail in the future, in order to craft a more precise and fully-functional solution that can satisfy the business needs and provide quantifiable help to internal processes.

List of Acronyms

IoT	Internet of Things
BU	Business Unit
ML	Machine Learning
IVASS	Istituto per la Vigilanza sulle Assicurazioni
AI	Artificial Intelligence
EMEA	Europe - Middle East - Africa
IIA	Italian Insurtech Association
CONSOB	Commissione nazionale per le società e la Borsa
BDS	Banca Dati Sinistri
AIA	Archivio Integrato Antifrode
SOTA	State Of The Art
EDA	Exploratory Data Analysis
RSF	Random Survival Forest
SVM	Support Vector Machine
AFT	Accelerated Failure Time
RMSE	Root Mean Squared Error
C-Index	Concordance Index
GIS	Geographic information system
KME	Kaplan-Meier Estimator

List of Figures

List of Tables

Bibliography

[1] C.-F. Chung, P. Schmidt, A. D. Witte, and A. D. Witte. Survival analysis: A survey. *Journal of Quantitative Criminology*, 7(1):59–98, 1991. ISSN 07484518, 15737799.

[2] Coalition Against Insurance Fraud. Fraud stats: The impact. 12 2021.

[3] E. Commission. Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. 04 2021.

[4] Digital McKinsey. Digital disruption in insurance: Cutting through the noise. pages 6–9, 3 2017.

[5] E. K. Frost, R. Bosward, Y. S. J. Aquino, A. Braunack-Mayer, and S. M. Carter. Public views on ethical issues in healthcare artificial intelligence: protocol for a scoping review. *Systematic Reviews*, 07 2022.

[6] M. R. Hanafy M. *Risks*, 2, 2021. doi: 10.3390/risks9020042.

[7] F. Harrell, K. Lee, D. Matchar, and T. Reichert. Harrell jr fe, lee kl, matchar db, reichert taregression models for prognostic prediction: advantages, problems, and suggested solutions. cancer treat rep 69: 1071-1077. *Cancer treatment reports*, 69: 1071–77, 11 1985.

[8] IIA. ll mercato insurtech e la penetrazione delle polizze digitali al 2030. 03 2021.

[9] B. Itri, Y. Mohamed, Q. Mohammed, and B. Omar. Performance comparative study of machine learning algorithms for automobile insurance fraud detection. In *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, pages 1–4, 2019. doi: 10.1109/ICDS47004.2019.8942277.

[10] IVASS. Relazione antifrode 2019. 05 2021.

[11] IVASS. Relazione antifrode 2020. 05 2021.

[12] M. Jahanbani Fard, P. Wang, S. Chawla, and C. Reddy. A bayesian perspective on early stage event prediction in longitudinal data. *IEEE Transactions on Knowledge and Data Engineering*, PP, 09 2016. doi: 10.1109/TKDE.2016.2608347.

[13] A. Jobin, M. Ienca, and E. Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, page 389–399, 09 2019.

[14] N. Köbis, C. Starke, and I. Rahwan. The promise and perils of using artificial intelligence to fight corruption. *Nature Machine Intelligence*, 05 2022.

[15] J. Mueller. Insurtech rising: A profile of the insurtech landscape. 12 2018.

[16] C. N, F. G, and C. M. A data mining based system for credit-card fraud detection in e-tail. pages 1–9, 2017. doi: 10.1109/ICCNI.2017.8123782.

[17] D. V. Nath and S. Geetha. 2019. *Procedia Computer Science*, 165:631–641, 1 2015.

[18] K. Nian, H. Zhang, A. Tayal, T. Coleman, and Y. Li. Auto insurance fraud detection using unsupervised spectral ranking for anomaly. *The Journal of Finance and Data Science*, 2(1):58–75, 2016. ISSN 2405-9188. doi: https://doi.org/10.1016/j.jfds.2016.03.001.

[19] Politecnico di Milano. Cresce il fintech in italia: quali sono le startup e i numeri e della rivoluzione. *FinTech & InsurTech Observatory of the School of Managemen*, 12 2021.

[20] A. D. Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi. Credit card fraud detection and concept-drift adaptation with delayed supervised information. pages 1–8, 4 2015.

[21] A. E. Raftery. Bayesian model selection in social research. *Sociological Methodology*, 25:111–163, 1995. ISSN 00811750, 14679531.

[22] B. V. Van, K. Pelckmans, J. A. K. Suykens, and S. V. Huffel. Support vector machines for survival analysis. *Proceedings of the 3rd International Conference on Computational Intelligence in Medicine and Healthcare (CIMED'07).*, pages 1–8, 2007.

Appendix

List of Queries

- Survival Time

  For uncensored records:

  –
```
UPDATE FINAL_DATASET T1
SET CLIENT_SINCE = (SELECT MIN(DT_POLICY_EFFECT)
FROM PDW_POLICY T2
WHERE T1.CD_POLICY_NUMBER = T2.CD_POLICY_NUMBER);

UPDATE FINAL_DATASET T1
SET SURVIVAL_TIME_MONTHS =
    CASE
    WHEN (EXTRACT(MONTH FROM DT_CLAIM_REPORTED) =
    EXTRACT(MONTH FROM CLIENT_SINCE))
    THEN
        (EXTRACT(YEAR FROM DT_CLAIM_REPORTED) -
        XTRACT(YEAR FROM CLIENT_SINCE))*12
    ELSE
        ROUND((EXTRACT(YEAR FROM DT_CLAIM_REPORTED) -
        EXTRACT(YEAR FROM CLIENT_SINCE))*12 +
        ABS(MONTHS_BETWEEN(CLIENT_SINCE, DT_CLAIM_REPORTED)),3)
    END
    WHERE CD_CLAIM IS NOT NULL;

UPDATE FINAL_DATASET T1
SET SURVIVAL_TIME_DAYS = SURVIVAL_TIME_MONTHS * 31
WHERE CD_CLAIM IS NOT NULL;
```

- **For censored records:**

```
UPDATE FINAL_DATASET T1
SET SURVIVAL_TIME_MONTHS =
CASE
WHEN (EXTRACT(MONTH FROM DT_POLICY_EXPIRATION) =
EXTRACT(MONTH FROM CLIENT_SINCE))
THEN
    (EXTRACT(YEAR FROM DT_POLICY_EXPIRATION) -
    EXTRACT(YEAR FROM CLIENT_SINCE))*12
ELSE
    ROUND((EXTRACT(YEAR FROM DT_POLICY_EXPIRATION) -
    EXTRACT(YEAR FROM CLIENT_SINCE))*12 +
    ABS(MONTHS_BETWEEN(
    CLIENT_SINCE, DT_POLICY_EXPIRATION)),3)
END
WHERE CD_CLAIM IS NULL;
```

- Number of insured vehicles at event time
  - **For censored records:**

```
UPDATE FINAL_DATASET T1
SET EVENT_TIME_INSURED_VEHICLES =
(SELECT COUNT(*) FROM PDW_POLICY T2
WHERE T1.CD_POLICY_NUMBER = T2.CD_POLICY_NUMBER
AND T1.DT_CLAIM_LOSS >= T2.DT_POLICY_EFFECT
AND T1.DT_CLAIM_LOSS < T2.DT_POLICY_EXPIRATION)
WHERE CD_CLAIM IS NOT NULL
```

  - **For uncensored records:**

```
UPDATE FINAL_DATASET T1
SET EVENT_TIME_INSURED_VEHICLES =
(SELECT COUNT(*) FROM PDW_POLICY T2
WHERE T1.CD_POLICY_NUMBER = T2.CD_POLICY_NUMBER
AND T2.DT_POLICY_EXPIRATION > SYSDATE)
WHERE CD_CLAIM IS NULL
```

- Age at event time
  - **For uncensored records:**

```
UPDATE FINAL_DATASET
SET AGE_AT_EVENT_TIME = (
EXTRACT(Year FROM DT_CLAIM_REPORTED) -
EXTRACT(Year FROM BIRTH_DT))
WHERE CD_CLAIM IS NOT NULL
```

  - **For censored records:**

```
UPDATE FINAL_DATASET
SET AGE_AT_EVENT_TIME = (
EXTRACT(Year FROM SYSDATE) - EXTRACT(Year FROM BIRTH_DT))
WHERE CD_CLAIM IS NULL
```

- Time since the policy started
  - **Only for uncensored records:**

```
UPDATE FINAL_DATASET
SET TIME_SINCE_POLICY_OPEN =
ABS(MONTHS_BETWEEN(DT_CLAIM_REPORTED, DT_POLICY_EFFECT);
```

- Time until the policy expires
  - **Only for uncensored records:**

```
UPDATE FINAL_DATASET
SET TIME_TO_POLICY_END =
ABS(MONTHS_BETWEEN(DT_POLICY_EXPIRATION, DT_CLAIM_REPORTED);
```

- Average value of all active insurances at event time
  - **For all records:**

```
UPDATE FINAL_DATASET T1
SET AVG_INSURED_VALUE = (
    SELECT AVG(NM_INSURED_VALUE)
    FROM PDW_PERSONAL_INFO T2
```

```
    WHERE T1.CD_POLICY = T2.CD_POLICY
    AND T1.CD_POLICY_NUMBER = T2.CD_POLICY_NUMBER
);
```

- Average AIA score at event time
  - **For the fraud records with previous claims:**

```
UPDATE FINAL_DATASET T1
SET AVG_AIA_SCORE = 1
WHERE CD_POLICY NOT IN
(SELECT CD_POLICY FROM PDW_CLAIM);
```

  - **For the fraud records with no previous claims:**

```
UPDATE FINAL_DATASET T1
SET AVG_AIA_SCORE = (
SELECT AVG(AIA_SCORE)
FROM TESI_PDW_CLAIM T2
WHERE T1.CD_CLAIM = T2.CD_CLAIM)
WHERE AVG_AIA_SCORE != 0
AND FL_FRAUD = 1;
```

- Number of past claims
  - **With previous opened claims:**

```
UPDATE TESI_DATASET_FINALE_CENSORED T1
SET PREVIOUS_CLAIMS_TOT = (
SELECT COUNT(*)
FROM TESI_PDW_CLAIM T2
WHERE T2.DT_CLAIM_LOSS < T1.DT_CLAIM_LOSS
AND T1.CD_POLICY_NUMBER = T2.CD_POLICY_NUMBER
AND FL_FRAUD = 0)
WHERE CD_CLAIM IS NOT NULL;
```

  - **With no previous opened claims:**

```
UPDATE TESI_DATASET_FINALE_CENSORED T1
```

```
SET PREVIOUS_CLAIMS_TOT = 0
WHERE CD_CLAIM IS NULL;
```

- Number of past frauds
  - **With previous committed fraud:**

```
UPDATE TESI_DATASET_FINALE_CENSORED T1
SET PREVIOUS_FRAUDS_TOT = (
SELECT COUNT(*)
FROM TESI_PDW_CLAIM T2
WHERE T2.DT_CLAIM_LOSS < T1.DT_CLAIM_LOSS
AND T1.CD_POLICY_NUMBER = T2.CD_POLICY_NUMBER
AND FL_FRAUD = 1)
WHERE CD_CLAIM IS NOT NULL;
```

  - **With no previous committed fraud:**

```
UPDATE TESI_DATASET_FINALE_CENSORED T1
SET PREVIOUS_FRAUDS_TOT = 0
WHERE CD_CLAIM IS NULL;
```