

Machine Learning in Forecasting Motor Insurance Claims

Thomas Poufinas, Periklis Gogas * , Theophilos Papadimitriou  and Emmanouil Zaganidis

Department of Economics, Democritus University of Thrace, 69100 Komotini, Greece; tpoufina@econ.duth.gr (T.P.); papadimi@econ.duth.gr (T.P.); ezaganid@econ.duth.gr (E.Z.)

* Correspondence: pgkogkas@econ.duth.gr

Abstract: Accurate forecasting of insurance claims is of the utmost importance for insurance activity as the evolution of claims determines cash outflows and the pricing, and thus the profitability, of the underlying insurance coverage. These are used as inputs when the insurance company drafts its business plan and determines its risk appetite, and the respective solvency capital required (by the regulators) to absorb the assumed risks. The conventional claim forecasting methods attempt to fit (each of) the claims frequency and severity with a known probability distribution function and use it to project future claims. This study offers a fresh approach in insurance claims forecasting. First, we introduce two novel sets of variables, i.e., weather conditions and car sales, and second, we employ a battery of Machine Learning (ML) algorithms (Support Vector Machines—SVM, Decision Trees, Random Forests, and Boosting) to forecast the average (mean) insurance claim per insured car per quarter. Finally, we identify the variables that are the most influential in forecasting insurance claims. Our dataset comes from the motor portfolio of an insurance company operating in Athens, Greece and spans a period from 2008 to 2020. We found evidence that the three most informative variables pertain to the new car sales with a 3-quarter and 1-quarter lag and the minimum temperature of Elefsina (one of the weather stations in Athens) with a 3-quarter lag. Among the models tested, Random Forest with limited depth and XGboost run on the 15 most informative variables, and these exhibited the best performance. These findings can be useful in the hands of insurers as they can consider the weather conditions and the new car sales among the parameters that are considered to perform claims forecasting.



Citation: Poufinas, Thomas, Periklis Gogas, Theophilos Papadimitriou, and Emmanouil Zaganidis. 2023. Machine Learning in Forecasting Motor Insurance Claims. *Risks* 11: 164. <https://doi.org/10.3390/risks11090164>

Academic Editor: Shengkun Xie

Received: 11 August 2023

Revised: 11 September 2023

Accepted: 13 September 2023

Published: 18 September 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: insurance; claims; forecasting; machine learning

JEL Classification: G22; C53

1. Introduction

Insurance is the activity by which an individual or enterprise exchanges an uncertain (financial) loss with a certain (financial) loss. The former is the outcome of an event for which the insured individual or enterprise has received coverage via an insurance policy; the latter is the premium that the insured has to pay to receive this coverage. When such an event occurs, the insured may formally request coverage (monetary or in-kind) in line with the policy terms and conditions, which constitutes the insurance claim.

It is therefore clear that claims are key components of the insurance activity as they essentially comprise the realization of the insurance product/service. Due to the uncertainty of (future) claims occurrence, it is in the interest of the insurers to carefully frame their claims expectations and provisions. Consequently, they pursue claims forecasting. The accurate forecasting of insurance claims is important for several reasons.

First, claims constitute the basis of pricing. In insurance, contrary to other services, the validity of the pricing is confirmed, and the adequacy of the premium is proved only after the experience has been recorded. Traditional pricing is based on historical data; however, it is the occurrence of incidents in the future that determines whether the estimated burning cost was correct or not. Hence, if the claims experience has not been

properly embedded in the pricing models, the (pure) premium may not be sufficient to cover the total claims (incurred or paid) and this could lead to a loss-making activity—if the premium charged is too low. In contrast, it could result in the loss of customers—in the case where the premium charged is too high.

Second, future claims occurrence is important for the compilation of the business plan as claims affect the future profitability of the company. In fact, the claims experience is probably the most significant determinant of the operational profitability of the insurance company. This is due to the fact that when compiling the business plan, an insurance company projects the future premia and the future claims over a period of years. Future premia are based primarily on sales forecasts, the evolution of inflation (ideally the one related to the insurance coverage under examination), as well as the projected claims experience. Expected future claims are based on the historical claims experience as well as on assumptions on the development of claims; this may be decomposed to the development of claims frequency and severity.

Finally, having a forward look in claims is a prerequisite of their risk and solvency assessment process and report, which depicts the risk appetite of the insurer and thus the capital required for the solvency of the insurer. As a matter of fact, it usually requires (one of) the biggest portions of capital (allocations). Indeed, insurers assume the risks that individuals and enterprises want to transfer, hedge, or mitigate. A claim is filed when a covered event (the assumed risk) has occurred. A higher risk appetite indicates the assumption of higher risk and thus higher claim anticipation. This leads to higher (economical) capital required for the absorption of this risk.

The conventional forecasting approaches attempt to either repeat the historical (growth) pattern of claims in the future—with potential seasonality and respective premia considered—or match the claims frequency and severity experience of the insurance company with a known probability distribution function. Smaller claims exhibit higher frequency, whereas large claims have a (much) smaller frequency. To improve the precision of the forecasting, large claims are pooled separately from the small claims and different probability distribution functions are used to best fit the claims frequency and severity of the two pools of claims.

Machine Learning approaches offer an alternative route to claims forecasting. The contribution of ML (artificial intelligence—AI) in insurance globally and in claims prediction specifically has been recognized by practitioners—who have spotted a wide range of ML applications in insurance—spreading over almost all its processes, such as claims processing, claims fraud detection, claims adjudication, claim volume forecasting, automated underwriting, submission intake, pricing and risk management, policy servicing, insurance distribution, product recommendation/personalized offers, assessor assistance, property (damage) analysis, automated inspections, customer lifetime value prediction/customer retention/lapse management, speech analytics, customer segmentation, workstream balancing for agents, and self-servicing for policy management (Seely 2018 and Somani 2021). A report from the Organization for Economic Cooperation and Development (OECD 2020) subscribes to this point of view as it identifies the increasing number of ML (AI) applications in insurance, which are enabled through the widespread collection of big data and their analysis. The report pinpoints marketing, distribution and sales, claims (verification and fraud), pricing, and risk classification as broader areas of ML utilization. It further addresses some attention points, such as policy and regulation with regards to the use of ML in insurance, with emphasis among others in privacy and data protection, market structure, risk classification, and explainability of ML. The implementation of ML (AI) methods in these sectors of the insurance operations, along with the relevant worries on ethical and societal challenges have been recorded by Grize et al. (2020), Banks (2020), Ekin (2020), and Paruchuri (2020). The reports of Deloitte (2017), SCOR (2018), Keller et al. (2018), and Balasubramanian et al. (2021) identify similar applications of ML as they pave the future of insurance.

In this paper, we employ a series of Machine Learning algorithms (Support Vector Machines–SVM, Decision Trees, Random Forests, and Boosting) to forecast the average (mean) insurance claims amount per insured car per quarter and identify a subset of variables that are the most relevant in determining the average claims amount. The claims data come from the motor portfolio of an insurance company (operating in Athens, Greece) for the period between 2008–2020.

This approach is novel as it investigates the impact of two new-to-the-literature sets of variables, namely variables relevant to weather conditions and car sales, on the evolution of motor insurance claims with the use of ML techniques to forecast motor insurance claims. More specifically, insurers attempt to forecast motor insurance claims based on their own experience, which depends on the particulars of the vehicle and the driver. However, there is a third component recorded as “road” (Norman 1962; Dimitriou and Poufinas 2016). “Road” describes (environmental) factors such as time of the day, day of the week, weather conditions, type of road design and surface, lighting and visibility, etc. It is essentially a set of factors that refer to all factors that can affect the incidence of road traffic accidents other than the factors that are relevant to the driver (road user) and the vehicle. “Road” encompasses all factors that are not captured by the driver and the vehicle. Driver, vehicle, and “road” are essentially sets of factors. Consequently, “road” captures (among others) the condition of the terrain, which is impacted by the weather conditions (among other parameters). Furthermore, “road” captures the road usage, which is affected by the number of vehicles using it. This is, in turn, impacted by the new and used car sales. As a result, we feel we unveil the attributes of one important motor accident component, namely, “road”, which is novel in motor insurance claim forecasting.

We trust this is useful in the hands of insurers as they now have an additional set of factors to perform motor claim forecasting. When performing motor claim forecasting, some insurers, among which the insurer that provided the dataset for this study, rely on the address that the insured vehicle is registered; hence, they do not consider the area where the accident took place. As a result, the “road” component is not captured. Our approach offers a way to forecast motor insurance claims with the inclusion of two sets of parameters that impact this component: weather conditions and car sales.

2. Literature Review

The bulk of the literature on the applications of machine learning in insurance is relatively recent (post 2019) and although they cover a wide range of topics relevant to the insurance activity, there is ample room for further research. The main literature strands focus on claims, reserving, pricing, capital requirements–solvency, coverage ratio, acquisition, and retention. We group them into two main categories; actuarial and risk management that incorporates the first four (claims, reserving, pricing, and capital requirements–solvency) and customer management, which incorporates the last three (coverage ratio, acquisition, and retention). As the second category is not relevant to our study, we do not present it in detail. The interested reader may look at Mueller et al. (2018) for the coverage ratio; Boodhun and Jayabalan (2018) and Qazi et al. (2020) for acquisition; and Grize et al. (2020) and Guillen et al. (2021) for retention.

The literature that is relevant to actuarial and risk management issues addresses the main functions of the insurance activity and is thus related to actuarial science and risk management. In fact, insurance is the assumption and management of risks that individuals or enterprises wish to transfer or mitigate. These functions entail the monitoring of the claims/risks evolution, the determination of the required reserves, the estimation of the appropriate tariff rates as well as the calculation of the capital that is required to ensure the solvency of the insurer. The analysis of these literature strands follows.

2.1. Claims/Risks

Fauzan and Murfi (2018) focus on the forecasting of motor insurance accident claims via ML methods with an emphasis on missing data. Rustam and Ariantari (2018) use ML

approaches to predict the occurrence of motor insurance claims based on their claim history (with data stemming from an Indonesian motor insurer). [Pesantez-Narvaez et al. \(2019\)](#) attempt to predict the existence of accident claims with the use of ML techniques on telematics data (coming from an insurance company) with an emphasis on driving patterns (total annual distance driven and percentage of distance driven in urban areas). [Qazvini \(2019\)](#) employs ML methods to predict the number of zero claims (i.e., claims that have not been reported) based on telematics data (on French motor third party liability). [Bermúdez et al. \(2020\)](#) apply ML approaches to model insurance claim counts with an emphasis on the overdispersion and the excess number of zero claims, which may be the outcome of unobserved heterogeneity. [Bärtil and Krummacker \(2020\)](#) attempt to predict the occurrence and the magnitude of export credit insurance claims with the use of ML techniques. The models employed produce satisfactory results for the former but not so satisfactory for the actual claim ratios—with accuracy, Cohen's κ and R^2 were used to assess model performance. [Knighton et al. \(2020\)](#) focused on forecasting flood insurance claims with ML models that applied hydrologic and social demographic data to realize that the incorporation of such data can improve flood claim prediction. [Hanafy and Ming \(2021\)](#) apply ML approaches to predict the occurrence of motor insurance claims (over the portfolio of Porto Seguro, a large Brazilian motor insurer). [Selvakumar et al. \(2021\)](#) concentrated on the prediction of the third-party liability (motor insurance) claim amount for different types of vehicles with ML models (on a dataset derived from Indian public insurance companies).

Some recent articles utilize the data collected through telematics. More specifically, [Duval et al. \(2022\)](#) used ML models to come up with a method that indicates the amount of information—collected via telematics with regards to the policyholders' driving behavior—that needs to be (optimally) retained by insurers to (successfully) perform motor insurance claim classification. [Reig Torra et al. \(2023\)](#) also capitalized on the data provided by telematics and used the Poisson model, along with some weather data, to forecast the expected motor insurance claim frequency over time. They found that weather conditions do affect the risk of an accident. [Masello et al. \(2023\)](#) used the information collected via telematics and employed ML methods to assess the predictive ability of driving contexts (such as road type, weather, and traffic) to driving risks/safety (such as near-misses, speeding, and distraction events), which, in turn, affected the exposure to/occurrence of accidents and thus motor insurance claims.

[Pesantez-Narvaez et al. \(2021\)](#) compared the ability of ML models to detect rare events (on a third-party liability motor insurance dataset) to realize that RiskLogitboost regression exhibits a superior performance over other methods. [Shi and Shi \(2022\)](#) employed ML approaches on property insurance claims to develop rating classes and estimate rating relativities for a single insurance risk; perform predictive modeling for multivariate insurance risks and unveil the impact of tail-risk dependence; and price new products.

In a different direction—that of fraud detection—[Pérez et al. \(2005\)](#) applied ML approaches (on a motor insurance portfolio) in a different context, which still pertained to claims; they focused on the detection of fraudulent claims in motor insurance by properly classifying suspicious claims. [Kose et al. \(2015\)](#) employed ML approaches for the detection of fraudulent claims or abusive behavior in healthcare insurance via an interactive framework that incorporates all the interested parties and materials involved in the healthcare insurance (claim) process. On the same topic, [Roy and George \(2017\)](#) used ML methods to detect fraudulent claims in motor insurance. [Wang and Xu \(2018\)](#) employed ML models that incorporate the (accident) information embedded in the text of the claims to detect potential claim fraud in motor insurance. [Dhieb et al. \(2019, 2020\)](#) applied ML techniques to automatically identify motor insurance fraudulent claims and sort them into different fraud categories with minimal human intervention, along with alerts for suspicious claims.

A series of papers implemented ML approaches in health management/insurance. [Bauder et al. \(2016\)](#) introduced ML approaches to tackle a different topic of insurance claims, thereby allowing them to spot the physicians that post a potentially anomalous behavior (pointing out misuse, fraud, or ignorance of the billing procedures) in health

(medical) insurance claims (with data taken from the USA Medicare system) and for which additional investigation may be necessary.

Hehner et al. (2017) highlighted the merits of the introduction of ML (AI) in hospital claims management, which can be summarized as savings for both the insurers and the insured as ML algorithms result in increased efficiency and well-informed decision-making to the benefit of all interested parties. Rawat et al. (2021) applied ML methods to analyze claims and conclude on a set of factors that facilitate claim filing and acceptance. Cummings and Hartman (2022) propose a series of ML models that provide insurers the ability to forecast Long Term Care Insurance (LTCI) claim rates and thus better their capacity to operate as LTCI providers.

2.2. Reserving

Baudry and Robert (2019) developed a ML method to estimate claims reserves with the use of all policy and policyholder covariates, along with the information pertaining to a claim from the moment it has been reported and compared their results with those generated via chain ladder. Elpidorou et al. (2019) employed ML techniques to introduce a novel Bornhuetter–Ferguson method as a variant of the traditional chain ladder method used for reserving in non-life (general) insurance through which the actuary can adjust the relative ultimate reserves with the use of externally estimated relative ultimate reserves. In the same direction, Bischofberger (2020) utilized ML methods to extend the chain ladder method via the estimated hazard rate for the estimation of non-life claims reserves.

The outperformance (in 4 out of 5 lines of business studied) of ML algorithms over traditional actuarial approaches in estimating loss reserves (future customer claims) is evidenced by the work of Ding et al. (2020). Similarly, Gabrielli et al. (2020) explore the merit of the introduction of ML approaches to traditional actuarial techniques in improving the non-life insurance claims reserving (prediction).

2.3. Pricing

Gan (2013), in a comparatively early work, priced the guarantees (i.e., finds the market value and the Greeks) of a large portfolio of variable annuity policies (generated by the author) via ML techniques. Assa et al. (2019) used ML approaches to study the correct pricing of deposit insurance by improving the implied volatility calibration to avoid mispricing due to arbitrage. Grize et al. (2020) unveiled the role of ML algorithms in (online) motor liability insurance pricing and, at the same time, increased the issue of interpretability. Henckaerts et al. (2021) capitalized on ML methods to price non-life insurance products based on the frequency and severity of claims; their results are superior to the ones produced by the traditionally employed generalized linear models (GLMs).

Kuo and Lupton (2020) explained that the wider adoption of ML techniques (over GLMs) in property and casualty insurance pricing depends very much on their reduced (perceived) transparency. They recommend increased interpretability to overcome this hurdle. These concerns are also addressed in Grize et al. (2020).

Blier-Wong et al. (2020) performed a literature review on the application of ML methods on the property and casualty insurance actuarial tasks and in pricing and reserving. They drafted potential future applications and research in the field and noticed that there can be three main challenges: interpretability, prediction uncertainty, and potential discrimination.

Some practitioner best practices have already been reported in the literature. AXA, for example, has applied ML methods to forecast large-loss car accidents to achieve optimal motor insurance pricing (Sato 2017; Ekin 2020).

2.4. Capital Requirements–Solvency

Díaz et al. (2005)—early enough compared to other studies—employed ML approaches to predict the insolvency of Spanish non-life insurance companies, which was applied on a set of financial ratios.

[Krah et al. \(2020\)](#) focused on the derivation of the solvency capital requirement that life insurers need to honor under the Solvency II directive in the European Union with the use of ML methods, which are alternative to the approximation techniques that insurance companies use.

Finally, [Wüthrich and Merz \(2023\)](#), in their book, presented the (entire) array of traditional actuarial and modern machine learning techniques that can be applied to address insurance-related problems. They explained how they can be applied by actuaries or real datasets and how the derived results may be interpreted.

As can be seen by the aforementioned literature review, our research is closer to the most recent articles of [Reig Torra et al. \(2023\)](#) and [Masello et al. \(2023\)](#), whose work has most likely been done in parallel with ours, as these papers were published in 2023. Still, our work maintains its novelty since (i) we use ML approaches compared to the work of [Reig Torra et al. \(2023\)](#), who employ the Poisson model (even though they also include weather data in their model); and (ii) we use the weather conditions/data in order to forecast the (mean) motor claims, compared to the analysis of [Masello et al. \(2023\)](#) who assess their impact on driving risks/safety.

3. Data and Variables

As the goal of this paper is to forecast the mean motor insurance claim cost, our dataset employs the claims data of a motor insurance portfolio from Athens, Greece. The data spans a period from 2008 to 2020. The frequency of the variables in our dataset is constrained by the availability of the data from the insurance company. Thus, we used a sample with quarterly frequency.

Besides the claims data, our dataset consists of the number of new car sales, imported used car sales in the greater region of Athens, the weather conditions as described by the maximum and minimum temperatures, the number of days that the temperature was below zero (Celsius), and the number of rainy days for three geographical areas, where weather stations are located in the broader region of Athens (Elefsina, Tatoi, and Spata). The choice of the three locations was dictated by the availability of data; the weather is recorded in several more areas within the broader Athens region, though we discovered large periods with no recorded values and, consequently, we were unable to include data from these areas in our dataset.

We have also included in our dataset four lags of each independent variable, as well as the moving averages of order four (MA(4)) for the target variable, the number of new cars, and the number of imported used cars sold. The total number of observations is 48, while the total number of explanatory variables is 79 (16 meteorological variables with four lags, the target variable, the number of new cars, and the number of imported used cars sold with their 4 lags and their moving averages).

The weather conditions data came from the Hellenic National Meteorological Service—HNMS (2022); the new and imported used car sales came from the Association of Motor Vehicles Importers Representatives—AMVIR (2022); and the claims data came from the motor insurance portfolio of the insurance company in Athens, Greece (who prefers not to be disclosed). All data were retrieved by their providers after a formal request.

Consequently, the dependent–target variable of our models is the mean (motor) insurance claims amount per car per quarter. The independent variables are presented in Table 1 below:

Table 1. Independent Variables.

Independent Variables				Definition	Source
Mean Insurance Claims/car $t - 1$	Mean Insurance Claims/car $t - 2$	Mean Insurance Claims/car $t - 3$	Mean Insurance Claims/car $t - 4$	The average claim amount per insured car	The motor insurance portfolio of the insurance company
New cars $t - 1$	New cars $t - 2$	New cars $t - 3$	New cars $t - 4$	New car sales	AMVIR (2022)
Used cars $t - 1$	Used cars $t - 2$	Used cars $t - 3$	Used cars $t - 4$	Imported used car sales	AMVIR (2022)
Max Temp Elefsina $t - 1$	Max Temp Elefsina $t - 2$	Max Temp Elefsina $t - 3$	Max Temp Elefsina $t - 4$	The maximum temperature recorded at the weather station of Elefsata	HNMS (2022)
Min Temp Elefsina $t - 1$	Min Temp Elefsina $t - 2$	Min Temp Elefsina $t - 3$	Min Temp Elefsina $t - 4$	The maximum temperature recorded at the weather station of Elefsata	HNMS (2022)
Mean Temp Elefsina $t - 1$	Mean Temp Elefsina $t - 2$	Mean Temp Elefsina $t - 3$	Mean Temp Elefsina $t - 4$	The average temperature recorded at the weather station of Elefsata	HNMS (2022)
Max Temp Tatoï $t - 1$	Max Temp Tatoï $t - 2$	Max Temp Tatoï $t - 3$	Max Temp Tatoï $t - 4$	The maximum temperature recorded at the weather station of Tatoï	HNMS (2022)
Min Temp Tatoï $t - 1$	Min Temp Tatoï $t - 2$	Min Temp Tatoï $t - 3$	Min Temp Tatoï $t - 4$	The maximum temperature recorded at the weather station of Tatoï	HNMS (2022)
Mean Temp Tatoï $t - 1$	Mean Temp Tatoï $t - 2$	Mean Temp Tatoï $t - 3$	Mean Temp Tatoï $t - 4$	The average temperature recorded at the weather station of Tatoï	HNMS (2022)
Max Temp Spata $t - 1$	Max Temp Spata $t - 2$	Max Temp Spata $t - 3$	Max Temp Spata $t - 4$	The maximum temperature recorded at the weather station of Spata	HNMS (2022)
Min Temp Spata $t - 1$	Min Temp Spata $t - 2$	Min Temp Spata $t - 3$	Min Temp Spata $t - 4$	The maximum temperature recorded at the weather station of Spata	HNMS (2022)
Mean Temp Spata $t - 1$	Mean Temp Spata $t - 2$	Mean Temp Spata $t - 3$	Mean Temp Spata $t - 4$	The average temperature recorded at the weather station of Spata	HNMS (2022)
No of rainy days Elefsina $t - 1$	No of rainy days Elefsina $t - 2$	No of rainy days Elefsina $t - 3$	No of rainy days Elefsina $t - 4$	The number of days with rain recorded at the weather station of Elefsina	HNMS (2022)
No of rainy days Spata $t - 1$	No of rainy days Spata $t - 2$	No of rainy days Spata $t - 3$	No of rainy days Spata $t - 4$	The number of days with rain recorded at the weather station of Spata	HNMS (2022)
No of rainy days Tatoï $t - 1$	No of rainy days Tatoï $t - 2$	No of rainy days Tatoï $t - 3$	No of rainy days Tatoï $t - 4$	The number of days with rain recorded at the weather station of Tatoï	HNMS (2022)
No of days below zero Tatoï $t - 1$	No of days below zero Tatoï $t - 2$	No of days below zero Tatoï $t - 3$	No of days below zero Tatoï $t - 4$	The number of days with a temperature below zero recorded at the weather station of Tatoï	HNMS (2022)
No of days below zero Elliniko $t - 1$	No of days below zero Elliniko $t - 2$	No of days below zero Elliniko $t - 3$	No of days below zero Elliniko $t - 4$	The number of days with a temperature below zero recorded at the weather station of Elliniko	HNMS (2022)
No of days below zero Elefsina $t - 1$	No of days below zero Elefsina $t - 2$	No of days below zero Elefsina $t - 3$	No of days below zero Elefsina $t - 4$	The number of days with a temperature below zero recorded at the weather station of Elefsina	HNMS (2022)
No of days below zero Spata $t - 1$	No of days below zero Spata $t - 2$	No of days below zero Spata $t - 3$	No of days below zero Spata $t - 4$	The number of days with a temperature below zero recorded at the weather station of Spata	HNMS (2022)
Moving Average Mean Insurance Claims/car	Moving Average New Cars	Moving Average Used Cars		The moving average of the aforementioned variables	

Note: Data and variables are on a quarterly basis. The time lag notation is as follows: $t - 1$ denotes a 1-year lag; $t - 2$ denotes a 2-year lag; $t - 3$ denotes a 3-year lag and $t - 4$ denotes a 4-year lag. Source: Created by the authors.

Figure 1 depicts the evolution of the mean insurance claims amount per insured car per quarter. One can observe a declining trend from 2010 to 2016 (with some seasonality on peaks and troughs, especially after 2012), which is most likely attributed to a significant reduction in car activity during this period. This was the result of the Greek sovereign debt crisis that started in 2010 and resulted in strict austerity measures that greatly negatively impacted household income, consumption, and the GDP. Fuel prices increased significantly after a new tax on fuel was introduced, and car sales reached a minimum for the decade.

After 2017, the trend is slightly increasing until the end of 2019, which coincides with the recovery of the Greek economy from the debt crisis. In 2020 the trend is decreasing again, without the seasonal rebound at the end of the year, as noted in previous years. This is most probably due to the effect of the pandemic, although we need more recent data to determine whether this assumption is valid. On the same figure we also illustrate the situation of the Greek economy: Unshaded areas represent periods of real GDP growth, while shaded areas represent periods of negative real GDP growth (real output contractions). There is a positive correlation between the insurance claims and real GDP. The relevant Pearson correlation coefficient is $\rho_{i,j} = 0.49$. This correlation statistic is significant even at the 0.01 significance level, with a p -value of 0.000368.

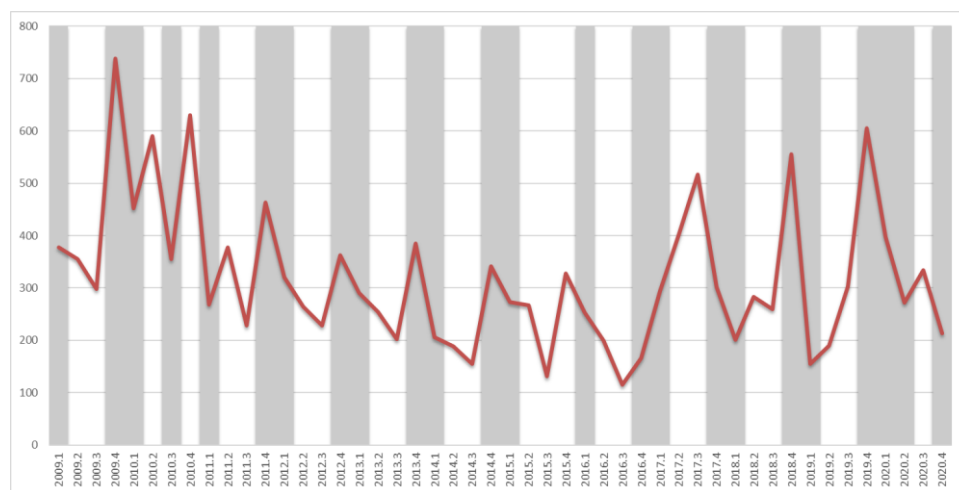


Figure 1. The time series of the mean insurance claims per insured car on a quarterly basis. In the background we depict the situation of the Greek economy: Unshaded areas represent periods of real GDP growth, while shaded areas represent periods of negative real GDP growth (real output contractions). Source: Based on authors estimates with data from the motor insurance portfolio.

We observe that the mean insurance claims exhibit some seasonality. More specifically, there is a peak (local maximum) noted on an annual basis during the 4th quarter of each year. In fact, there is a V-shaped formation starting from the peak of the 4th quarter of the previous year, dropping to reach a trough (local minimum) during the 3rd quarter and rising to reach a peak during the 4th quarter of the year. This is most likely attributed to the fact that the insured tend to declare their claims towards year-end and that the insurers tend to settle/pay—even the claims that were declared earlier in the year—towards year-end. The only exception is 2009, which is most likely due to the financial crisis that hit the country in 2009 and because of which the pattern may have been disrupted. The peak has shifted towards the 1st quarter of 2010. A second, lower peak is observed in the 2nd quarter in 2010, which subscribes to this point of view. After that, the pattern resumes until 2017, where the peak appears a bit earlier, towards the end of the 3rd quarter, which is a small deviation from the seasonality observed.

4. Methodology

Machine Learning was established in the 1950s to deliver the “Learning” component on the Artificial Intelligence (AI) systems. The basic concept of Machine Learning is the automated analytical model building; it is the idea that systems can learn from the data, identify patterns, and make decisions with minimal human intervention. They can also automatically improve their performance through experience. This is achieved by learning patterns and relationships in the data.

Historically, Machine Learning has relied on large datasets (Gogas and Papadimitriou 2021). This is the reason Machine Learning in economics was mainly applied to financial data,

the subfield of economics with an abundance of data mainly due to the availability of very high frequencies—daily, hourly, or even seconds or tick-to-tick. Towards the end of the 20th century, new algorithms were introduced, such as the Support Vector Machines and Random Forest coupled with Boosting and Bagging techniques, which achieved high accuracies with even small datasets. For this reason, we will base our forecasting study on these algorithms.

All variables in our dataset were normalized to a zero mean and unit variance (see [Ah-san et al. 2021](#)). The dataset was then split into two subsamples: the larger part of 38 observations was used for the training/testing step and the smaller subsample, the out-of-sample subset consisting of 10 observations, was kept outside the training process and was only used to evaluate the model’s performance to unseen–unknown data. Before splitting the dataset, the observations were shuffled to remove the temporal dimension from our training subset.

The training of the models was performed in a cross-validation framework. Cross-validation is the standard set-up to avoid overfitting during the training of the model (overfitting happens when the model learns to treat the data patterns in the training dataset but fails to generalize to new data). In cross-validation, the dataset (training/testing part) is divided into k equal folds (subsets) and the training/testing is performed k times. In every iteration, a different fold is used for the testing of the model, while the remaining $k - 1$ folds are used for training the model. The overall performance of each model is calculated as the mean performance over all the iterated k subsets. In Figure 2, we present a graphical representation of the cross-validation procedure with three folds. In our tests, we used a 4-fold configuration.

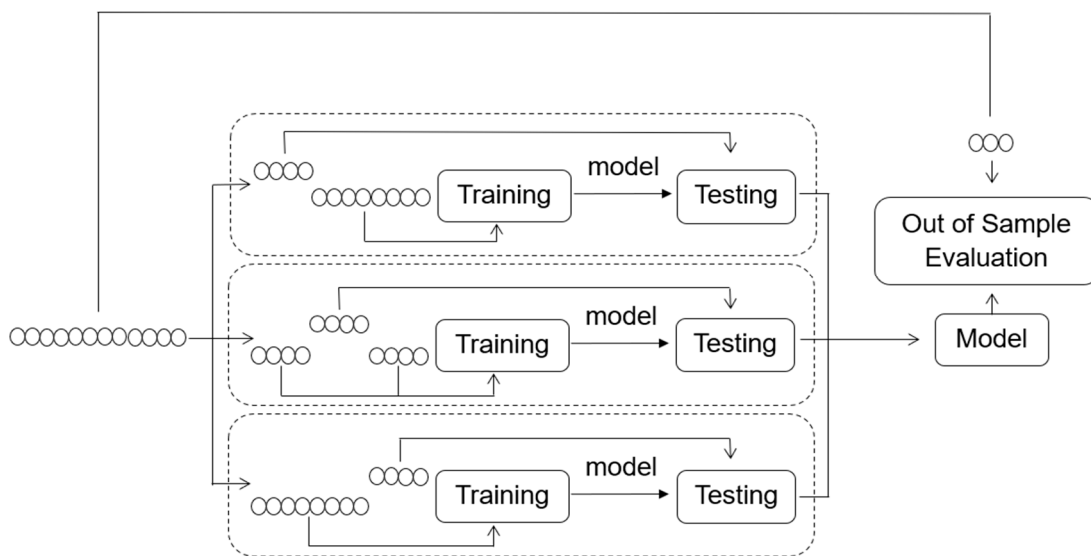


Figure 2. A graphical representation of cross-validation with three folds. Source: Created by the authors.

4.1. Support Vector Machines

Support Vector Machines (SVM) is a supervised machine learning algorithm that is used for both classification and regression tasks (Support Vector Regression–SVR). In classic regression (Ordinary Least Squares, for example), the main objective is to minimize the sum of the least squared errors. If, for example, we try to estimate the target y_i , using the data points x_i , the goal is to find the regression coefficients w that:

$$\min_w \sum_i (y_i - wx_i)^2$$

The basic concept of SVR is the creation of a tolerance band of width ϵ around the regression line. All the points in our dataset are expected to lie inside this tolerance band;

in mathematics, this means that we tolerate all the predicted values to fall within a $\pm\epsilon$ of the true values. The objective of SVR is to minimize the coefficients through the l2-norm of the coefficient vector. The errors are handled in the constraints, where we restrain the absolute error to the maximum error ϵ . When training the model, the ϵ is tuned according to the desired model accuracy. The objective function and constraints are:

$$\min_w \|w\|^2 \text{ subject to} \\ |y_i - wx_i| \leq \epsilon \text{ for all } i$$

The presented model is an unrealistic model since it cannot tolerate any error outside the ϵ -tolerance band. To allow larger than ϵ errors, we add the slack parameter ζ_i in our model as follows:

$$\min_{w, \zeta} \|w\|^2 + C \sum_i |\zeta_i|^2 \text{ subject to} \\ |y_i - wx_i| \leq \epsilon + \zeta_i \text{ for all } i$$

The constraint $\leq \epsilon + \zeta_i$ allows points to lie out of the band, though the addition of the slack parameter ζ_i in the objective function ensures that we want them to remain as close to the band as possible. Figure 3 gives an illustration of the SVR paradigm with slack variables.

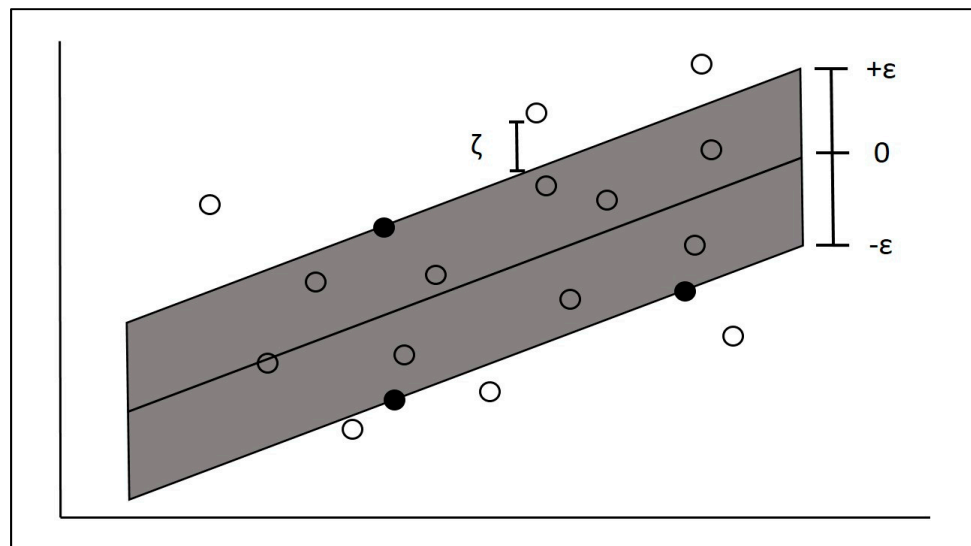


Figure 3. The Support Vector Paradigm. The gray area is the ϵ tolerance band around the regression line. Any point inside the gray area does not affect the objective function. Every point outside the gray zone adds ζ_i to the objective function of the minimization. The marginal black points in the gray zone are the Support Vectors.

The points in the margin of the ϵ -tolerance zone are called Support Vectors and they define the position of the regression line.

When the problem at hand cannot be treated using a linear classifier, then we use the kernel trick to change the dimensionality of the space where the optimization is performed during the training. The data points from the initial variable space of n -dimensions (in our illustrated example it is two dimensions) are projected into a higher dimension space, called the feature space, where the regression hyperplane creates the acceptable error; see Figure 4. When the kernel function is non-linear, the produced SVR model is also non-linear.

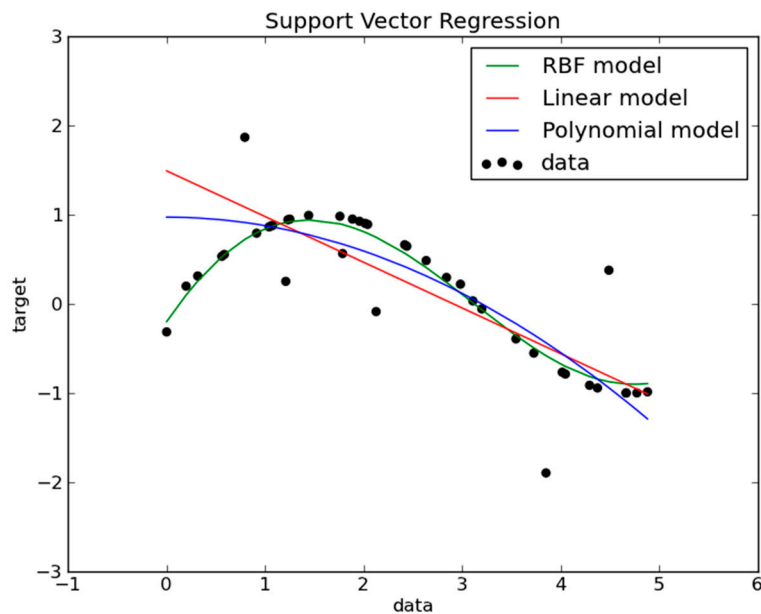


Figure 4. The kernel tricks in a non-linear case, taken from the Scikit-learn page. The black data points are approximated using the standard SVR (linear kernel), the polynomial kernel, and the RBF (Radial Basis Function) kernel. For the two kernel cases, the data are projected in another dimension where the regression is linear. When the regression line is returned in the initial data space it takes the form of the blue (polynomial) and green (RBF) line. https://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/auto_examples/svm/plot_svm_regression.html (accessed on 12 September 2023).

4.2. Decision Trees

Decision Trees are a supervised machine learning algorithm that is used for both classification and regression tasks. It works by recursively partitioning the data based on the most informative features. They are flowchart-like top-down structures of nodes and branches; see Figure 5. In regression tasks, the decision tree predicts the value of the target variable by averaging the values of the training data points that fall into the same leaf node.

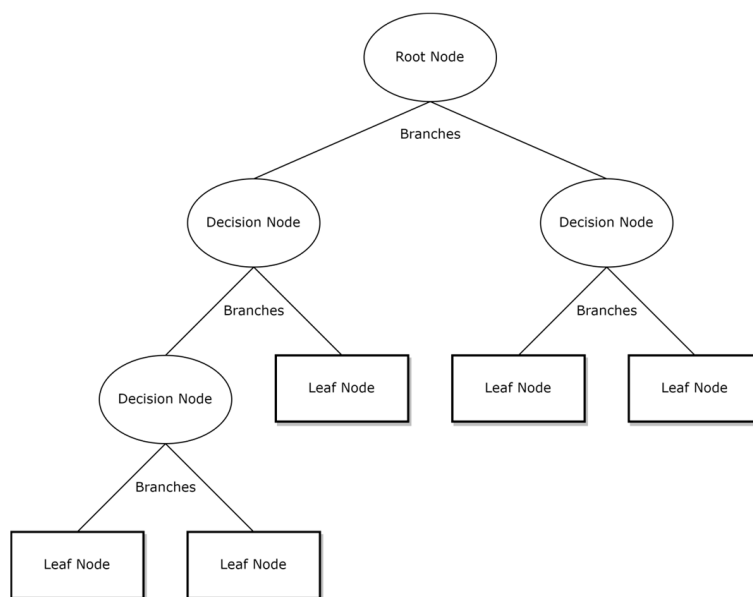


Figure 5. A graphical representation of Decision Trees. Every decision node is an if statement regarding one of the variables on the set. Every leaf node corresponds to a final value for the regression.

The top node is the root node representing the complete dataset. The decision nodes are if statements on one variable of the dataset. The two branches below each decision node split the dataset into two subsets: a subset where the decision node is true and a subset where the decision node is false. The nodes that do not split any further are called leaf (or terminal) nodes and depict the final outcomes of the decision-making process (in our case, a value for the regression procedure).

4.3. Random Forests

The Random Forests model explores the idea of many decision trees combined through a bootstrapping–aggregating algorithm, called bagging; see Figure 6. In each tree, a different randomly selected replacement subsample equal to the size of the initial dataset is used and a randomly selected subset of the initial independent variables is used for training the tree. The model’s ability to perform in unknown data is estimated using the observations not selected in the training step (this data part is often called out-of-bag in ML terminology). Thus, in the Random Forests algorithmic procedure, a stochastic process is implemented in choosing both the observations used for training (rows of the data matrix) and the independent variables (columns of the data matrix). In regression tasks, the prediction is made by calculating the average value of the target variable from the data points within every leaf node. In our experiments, we used two strategies: Random Forests with unlimited splits (partitionings) of the dataset and Random Forests with limited splits (8 splits).

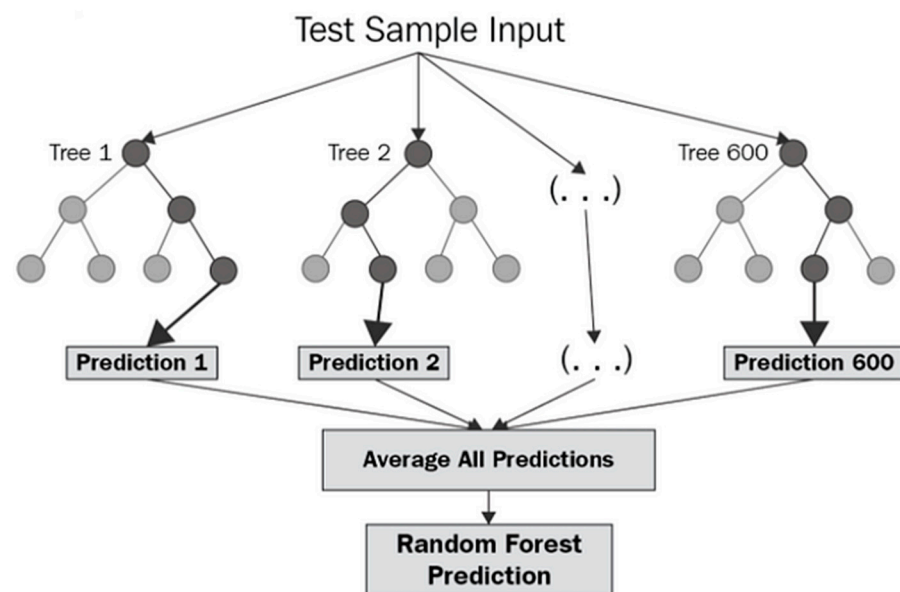


Figure 6. The depiction of a 600-tree Random Forest. The average of the 600 predictions is the final prediction of the Random Forest. The graphic was taken from <https://levelup.gitconnected.com/random-forest-regression-209c0f354c84> (accessed on 12 September 2023).

4.4. Boosting

Boosting is based on the idea of accumulating sequentially weak learners, each correcting its predecessors. The weak learners in our case are shallow decision trees, i.e., decision trees with few partitions of the dataset. Each added decision tree corrects the regression and improves the performance of the overall model. The combination of all the shallow decision trees constructs the boosted model (see Figure 7). In our experiments, we used the XGBoost algorithm to perform the boosting procedure.

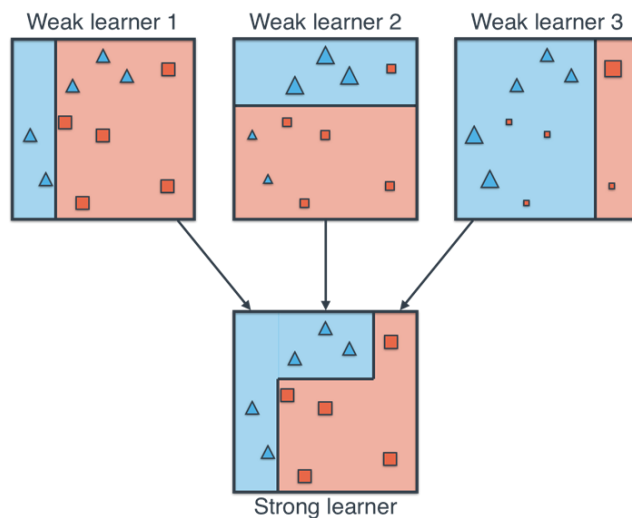


Figure 7. The basic concept of the combination of many weak learners to create a strong one using boosting. The information from the previous weak learners is incorporated by the size of the data points. The large data points correspond to points that were not correctly classified by the previous learners; the small ones describe the opposite case. Each of the three weak learners is unable to correctly classify the two classes, though their combination into a strong learner is successful. (<https://livebook.manning.com/book/grokking-machine-learning/chapter-10/v-9/45>) (accessed on 12 September 2023).

4.5. Forecast Evaluation

All models were evaluated using the Mean Absolute Percentage Error metric, defined as:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100$$

where x_i and \hat{x}_i are the actual and the forecasted values of the target variable, respectively, and n is the sample size. The MAPE metric measures the mean absolute distance between the actual and the forecasted values in percentages.

5. Results and Discussion

We trained an arsenal of ML-based models, including Support Vector Machines, Random Forests, and XGBoost, and tested it in two setups: (a) one using all the independent variables of our dataset, and (b) one using only the top-15 most informative independent variables¹ (Table 2). From these results we can see that all weather variables sum up to 32.96% relevance, while all car sales variables add to 30.55%. Thus, of the 15 most important variables, a total of 63.51% relevance is attributed to either the weather or car sales.

Table 2. The top-15 variables according to the Impurity Metric.

Top 15 Variables	Relevance
New Cars t – 3	10.49%
New cars t – 1	8.45%
Min Temp Elefsina t – 3	6.39%
New Cars t – 2	6.00%
Moving Average New Cars	5.61%
Min Temp Tatoi t – 1	4.60%
Min Temp Elefsina_t – 2	4.40%
Mean Temp Elefsina t – 3	3.97%
Rain Spata t – 4	3.89%
Mean Temp Elefsina t – 1	3.52%

Table 2. *Cont.*

Top 15 Variables	Relevance
Mean Insurance Claims/car t – 4	3.44%
Mean Insurance Claims/car t – 3	3.27%
Min Temp Elefsina t – 1	3.14%
Moving Average Mean Insurance Claims	3.08%
Min Temp Spata t – 3	3.05%

Source: Based on authors estimates with data from the motor insurance portfolio, the HNMS (2022) and the AMVIR (2022).

For each model, we calculated the MAPE for the training and testing subsets, and for the out-of-sample part. Nonetheless, due to the small size of our dataset (48 observations comprises a small dataset even for the selected methodologies), we choose to evaluate the performance of every model using the MAPE metric on the whole dataset. In Table 3, we present the results from all the subsets and the full dataset.

Table 3. The performance of our models in cross-validation with 4 folds.

Model	Independent Variables	Train Set	Test Set	Out-of-Sample	Full Data
Random Forest Limited Depth	all	24.47%	27.71%	29.61%	25.54%
Random Forest Full Grown Trees	all	17.11%	31.91%	32.27%	20.27%
Support Vector Machine	all	22.34%	28.76%	34.72%	24.92%
XGBoost	all	21.69%	33.11%	37.60%	25.00%
Random Forest Limited Depth	top 15	15.47%	30.08%	28.77%	18.24%
Random Forest Full Grown Trees	top 15	21.02%	27.41%	29.71%	22.83%
Support Vector Machine	top 15	23.73%	29.01%	32.47%	25.55%
XGBoost	top 15	16.36%	28.54%	31.72%	19.56%

Source: Based on authors estimates with data from the motor insurance portfolio, the HNMS (2022) and the Association of Motor Vehicles Importers Representatives (2022).

Overall, the best model was the Random Forests model with limited depth, which was fed with the top-15 most relevant variables and reached a MAPE of 18.24%. The second-best model was the XGBoost with the top-15 most relevant variables, which reached a MAPE of 19.56%. In Figure 8, we show the graphical representation of the actual and the forecasted values from the best model.

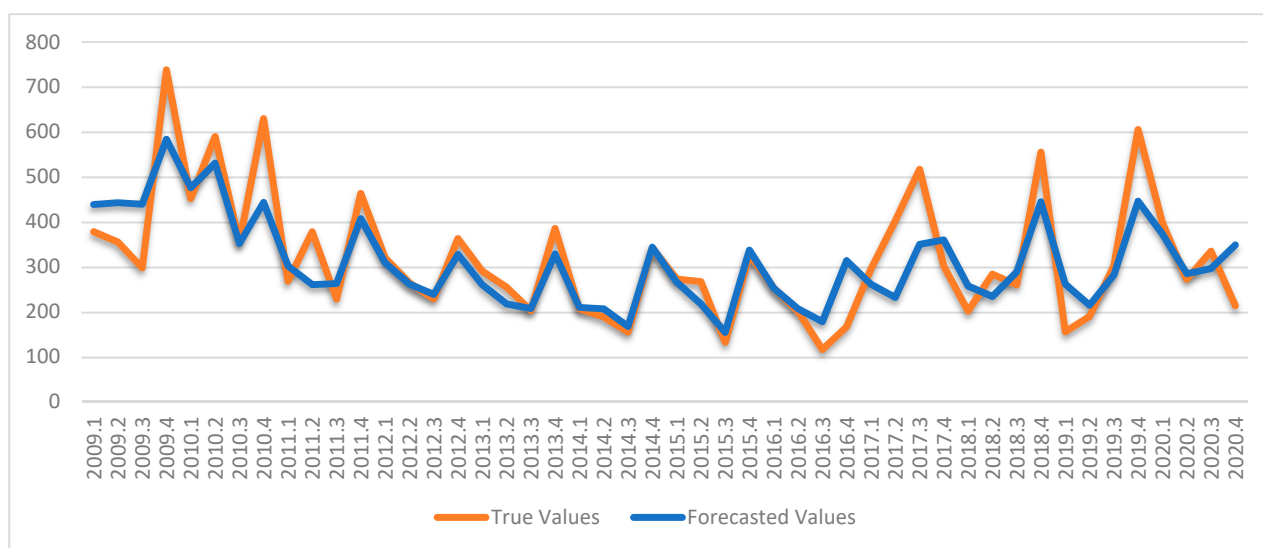


Figure 8. The actual and the forecasted values from the best model. Source: Based on authors estimates with data from the motor insurance portfolio, the HNMS (2022) and the Association of Motor Vehicles Importers Representatives (2022).

It is easy to verify that in most cases the forecasted values are visually close to the actual ones. Indeed, the peaks and troughs of both time series in Figure 8 coincide. Despite the model's failure to detect the peaks during the second quarter of 2011 and the third quarter of 2017, the mean absolute percentage error (MAPE) between 2011 and 2016 was calculated to be 7.6%.

The analysis performed indicates that the five variables with the most relevance, or, in other words, the best predictors of insurance claims, are: the number of new cars with a time lag of 3 quarters and 1 quarter, respectively, followed by the minimum temperature recorded at the Elefsina station 3 quarters ago, the number of new cars 2 quarters ago, and the moving average of new car registrations. The next five variables with respect to forecasting importance are: the minimum temperatures at the Tatoi station 1 quarter ago, the minimum temperatures at the Elefsina station 2 quarters ago as well as the mean temperatures at Elefsina 3 quarters ago, the rain at Spata station 4 quarters ago, and the mean temperature at Elefsina 1 quarter ago. The last set of five variables among the top 15 predictors are the mean insurance claims per car 4 quarters ago, the mean insurance claims per car 3 quarters ago, the minimum temperature in Elefsina 1 quarter ago, the moving average of the mean insurance claims, and the minimum temperature in Spata 3 quarters ago.

According to these results, the registration of new cars appears to be one of the most significant predictors of insurance claims. This may be attributed to the fact that as more new cars circulate, the number of accidents increases and thus the total claims cost increases, yielding a higher mean claims amount per insured vehicle in the motor portfolio under investigation. The time lag (of one to three quarters) is possibly justified, as drivers tend to be more careful when their car is brand new, and they stop paying that much attention as their car ages.

Furthermore, the lowest temperature can impact the overall claims expense because when such a temperature is recorded, the overall claims expense increases (potentially influenced by the number of accidents), leading to a consistently higher mean claims amount per insured vehicle in our motor portfolio. The time lag seems to be consistent with the new car time lag. The same is observed for the mean temperature and the rain, which is most likely due to a similar reason.

Finally, the mean insurance claims (per car with a time lag or moving average) affect the mean insurance claims per vehicle, but with a lower predictive capacity, indicating that the historical experience of the average claim amount is important for the determination of the current average claim amount. Indeed, one can expect that for a rather mature motor insurance portfolio, past claims can be (and are in practice) used to project future claims.

One potential explanation for the time lag is that claims are not immediately reported or paid. Consequently, they may be paid later in time. Moreover, when there are bodily injuries, the total claims amount increases and the time period over which the claim is paid is longer.

According to Table 2, which shows the top-15 variables in terms of importance, the weather variables (Min Temp Elefsina $t - 3$, Min Temp Tatoi $t - 1$, Min Temp Elefsina $t - 2$, Mean Temp Elefsina $t - 3$, Rain Spata $t - 4$, Mean Temp Elefsina $t - 1$, Min Temp Elefsina $t - 1$, and Min Temp Spata $t - 3$) have a significance of 32.96%, while the car sales variables (New Cars $t - 3$, New Cars $t - 1$, New Cars $t - 2$, and Moving Average New Cars) contribute 30.55%. Therefore, out of all the variables available, the 12 variables mentioned above measure a combined relevance of 63.51%, which is attributed either to the weather or to car sales. Additionally, there is a positive correlation between the total amount of insurance claims and the Greek real GDP. The relevant Pearson correlation coefficient is $\rho_{i,j} = 0.49$, which is highly significant. The high correlation between the total insurance claims and the real GDP may be due to the fact that the increasing economic activity leads to increased transportation needs both in distance and time travelled. Thus, firms are in need of being supplied by more raw materials; more goods are produced, traded, and distributed to the retail stores; consumers and employees spend more time driving to shopping centers

and to work; traffic jams are more frequent; and, also, usually more people use their private vehicles instead of public mass transportation when their income increases.

Overall, given the limited dataset that was supplied to us for this study, the algorithms employed seem to produce quite satisfactory results in terms of the MAPE that approximately ranges between 18% to 25%. Moreover, the Random Forests limited depth algorithm exhibits the best performance, especially during the time period from 2011 to 2016; one can see that the distance between the peaks and troughs in that period is comparatively smaller. This indicates that the models employed can be quite reliable when used by insurers for forecasting their motor insurance claims evolution.

6. Conclusions and Further Research

In this paper, we offer an innovative approach for insurance claims forecasting. The innovations of our approach are threefold: (a) we introduce two novel arrays of variables, i.e., weather conditions and car sales; (b) we managed to get the permission to use a proprietary dataset from an actual insurance company; and (c) we employ an arsenal of Machine Learning (ML) algorithms (Support Vector Machines, Decision Trees, Random Forests, and Boosting). We forecast the mean insurance claims amount per insured car per quarter and also identify the variables that are the most relevant in this forecast. The results show that the three most relevant variables are the new car sales with a 3-year and 1-year lag and the minimum temperature of Elefsina (one of the weather stations in Athens) with a 3-year lag. Random Forests limited depth and XGBoost, which were run on the top-15 variables in terms of relevance, are the best performers. Overall, weather variables sum up to 32.96% relevance, and car sales variables sum up to 30.55%.

Insurance companies may take advantage of these results when they attempt to forecast their future claims evolution by using weather conditions and new car sales as variables that affect claims. Furthermore, they can employ ML techniques when performing these forecasts, instead of the traditional actuarial approaches, as they seem to deliver quite reliable results.

Future research venues pertain to the forecasting of the claims frequency, as well as a deeper analysis of the findings of this paper, potentially using larger motor portfolios to train more accurate models, better test the validity of the results, and provide more thorough interpretations of the results.

As a next step, future research incorporates the simultaneous study of the new sets of variables reflecting weather conditions and car sales and the traditional set of variables capturing vehicle characteristics (such as type of vehicle—e.g., passenger car, truck, motorcycle, etc.; vehicle use—e.g., private car, commercial car, etc.; horsepower; weight; car make and model; manufacturing year; etc.) and driver characteristics (such as age, gender, driving experience, lifestyle, marital status, etc.).

Author Contributions: Methodology, T.P. (Thomas Poufinas), P.G., T.P. (Theophilos Papadimitriou) and E.Z.; Writing—original draft, T.P. (Thomas Poufinas), P.G., T.P. (Theophilos Papadimitriou) and E.Z.; Writing—review & editing, T.P. (Thomas Poufinas), P.G., T.P. (Theophilos Papadimitriou) and E.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data are confidential as per the NDA with the insurance company.

Conflicts of Interest: The authors declare no conflict of interest.

Note

- ¹ The relevance of the variables was estimated using the impurity feature during the random forest training. The Impurity features in brief counts how many times a variable is used in a decision tree of a random forest, how many times the variable is essential for the classification or regression of the random forest.

References

- Ahsan, Md Manjurul, M. A. Parvez Mahmud, Pritom Kumar Saha, Kishor Datta Gupta, and Zahed Siddique. 2021. Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. *Technologies* 9: 52. [CrossRef]
- Assa, Hirbod, Mostafa Pouralizadeh, and Abdolrahim Badamchizadeh. 2019. Sound deposit insurance pricing using a machine learning approach. *Risks* 7: 45. [CrossRef]
- Balasubramanian, R., Ari Libarikian, and Doug McElhaney. 2021. Insurance 2030—The Impact of AI on the Future of Insurance. McKinsey and Company. Available online: <https://www.mckinsey.com/industries/financial-services/our-insights/insurance-2030-the-impact-of-ai-on-the-future-of-insurance> (accessed on 28 December 2022).
- Banks, David. 2020. Discussion of “Machine learning applications in non-life insurance”. *Applied Stochastic Models in Business and Industry* 36: 538–40. [CrossRef]
- Bärtl, Mathias, and Simone Krummacker. 2020. Prediction of claims in export credit finance: A comparison of four machine learning techniques. *Risks* 8: 22. [CrossRef]
- Bauder, Richard A., Taghi M. Khoshgoftaar, Aarion Richter, and Matthew Herland. 2016. Predicting medical provider specialties to detect anomalous insurance claims. Paper presented at the 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), San Jose, CA, USA, November 6–8; pp. 784–90.
- Baudry, Maximilien, and Christian Y. Robert. 2019. A machine learning approach for individual claims reserving in insurance. *Applied Stochastic Models in Business and Industry* 35: 1127–55. [CrossRef]
- Bermúdez, Lluís, Dimitris Karlis, and Isabel Morillo. 2020. Modelling unobserved heterogeneity in claim counts using finite mixture models. *Risks* 8: 10. [CrossRef]
- Bischofberger, Stephan M. 2020. In-sample hazard forecasting based on survival models with operational time. *Risks* 8: 3. [CrossRef]
- Blier-Wong, Christopher, Hélène Cossette, Luc Lamontagne, and Etienne Marceau. 2020. Machine learning in PandC insurance: A review for pricing and reserving. *Risks* 9: 4. [CrossRef]
- Boodhun, Noorhannah, and Manoj Jayabalan. 2018. Risk prediction in life insurance industry using supervised learning algorithms. *Complex and Intelligent Systems* 4: 145–54. [CrossRef]
- Cummings, Jared, and Brian Hartman. 2022. Using Machine Learning to Better Model Long-Term Care Insurance Claims. *North American Actuarial Journal* 26: 470–83. [CrossRef]
- Deloitte. 2017. From Mystery to Mastery: Unlocking the Business Value of Artificial Intelligence in the Insurance Industry. Deloitte Digital. Available online: <https://www.coursehero.com/file/36465601/Artificial-Intelligence-in-Insurance-Whitepaper-deloitte-digitalpdf/> (accessed on 28 December 2022).
- Dhieb, Najmeddine, Hakim Ghazzai, Hichem Besbes, and Yehia Massoud. 2019. Extreme gradient boosting machine learning algorithm for safe auto insurance operations. Paper presented at the 2019 IEEE International Conference on Vehicular Electronics and Safety (ICVES), Cairo, Egypt, September 4–6; pp. 1–5.
- Dhieb, Najmeddine, Hakim Ghazzai, Hichem Besbes, and Yehia Massoud. 2020. A secure ai-driven architecture for automated insurance systems: Fraud detection and risk measurement. *IEEE Access* 8: 58546–58. [CrossRef]
- Díaz, Zuleyka, Maria Jesus Segovia, and Jose Fernández. 2005. Machine learning and statistical techniques. An application to the prediction of insolvency in Spanish non-life insurance companies. *The International Journal of Digital Accounting Research* 5: 1–45. [CrossRef]
- Dimitriou, Dimitrios, and Thomas Poufinas. 2016. Cost of road accident fatalities to the economy. *International Advances in Economic Research* 22: 433–45. [CrossRef]
- Ding, Kexing, Baruch Lev, Xuan Peng, Ting Sun, and Miklos A. Vasarhelyi. 2020. Machine learning improves accounting estimates: Evidence from insurance payments. *Review of Accounting Studies* 25: 1098–134. [CrossRef]
- Duval, Francis, Jean-Philippe Boucher, and Mathieu Pigeon. 2022. How Much Telematics Information Do Insurers Need for Claim Classification? *North American Actuarial Journal* 26: 570–90. [CrossRef]
- Ekin, Tahir. 2020. Discussion of “Machine learning applications in nonlife insurance”. *Applied Stochastic Models in Business and Industry* 36: 541–44. [CrossRef]
- Elpidorou, Valandis, Carolin Margraf, María Dolores Martínez-Miranda, and Bent Nielsen. 2019. A likelihood approach to Bornhuetter-Ferguson analysis. *Risks* 7: 119. [CrossRef]
- Fauzan, Muhammad Arief, and Hendri Murfi. 2018. The accuracy of XGBoost for insurance claim prediction. *International Journal of Advances in Soft Computing and Its Applications* 10: 159–71.
- Gabrielli, Andrea, Ronald Richman, and Mario Wüthrich. 2020. Neural network embedding of the over-dispersed Poisson reserving model. *Scandinavian Actuarial Journal* 2020: 1–29. [CrossRef]
- Gan, Guojun. 2013. Application of data clustering and machine learning in variable annuity valuation. *Insurance: Mathematics and Economics* 53: 795–801.
- Gogas, Periklis, and Theofilos Papadimitriou. 2021. Machine Learning in Economics and Finance. *Computational Economics* 57: 1–4. [CrossRef]
- Grize, Yves-Laurent, Wolfram Fischer, and Christian Lützelshwab. 2020. Machine learning applications in nonlife insurance. *Applied Stochastic Models in Business and Industry* 36: 523–37. [CrossRef]
- Guillen, Montserrat, Catalina Bolancé, Edward W. Frees, and Emiliano A. Valdez. 2021. Case study data for joint modeling of insurance claims and lapsation. *Data in Brief* 39: 107639. [CrossRef]

- Hanafy, Mohamed, and Ruixing Ming. 2021. Machine learning approaches for auto insurance big data. *Risks* 9: 42. [CrossRef]
- Hehner, Steffen, Boris Körs, Manuela Martin, Elke Uhrmann-Klingen, and Jack Waldron. 2017. Artificial Intelligence in Health Insurance: Smart Claims Management with Self-Learning Software. McKinsey and Company. Available online: <https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/artificial-intelligence-in-health-insurance-smart-claims-management-with-self-learning-software> (accessed on 28 December 2022).
- Henckaerts, Roel, Marie-Pier Côté, Katrien Antonio, and Roel Verbelen. 2021. Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal* 25: 255–85. [CrossRef]
- Keller, Benno, Martin Eling, Hato Schmeiser, Markus Christen, and Michele Loi. 2018. Big Data and Insurance: Implications for Innovation, Competition and Privacy. The Geneva Association. Available online: https://www.genevaassociation.org/sites/default/files/research-topics-document-type/pdf_public/big_data_and_insurance_-_implications_for_innovation_competition_and_privacy.pdf (accessed on 28 December 2022).
- Knighthon, James, Brian Buchanan, Christian Guzman, Rebecca Elliott, Eric White, and Brian Rahm. 2020. Predicting flood insurance claims with hydrologic and socioeconomic demographics via machine learning: Exploring the roles of topography, minority populations, and political dissimilarity. *Journal of Environmental Management* 272: 111051. [CrossRef] [PubMed]
- Kose, Ilker, Mehmet Gokturk, and Kemal Kilic. 2015. An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. *Applied Soft Computing* 36: 283–99. [CrossRef]
- Krah, Anne-Sophie, Zoran Nikolić, and Ralf Korn. 2020. Machine learning in least-squares Monte Carlo proxy modeling of life insurance companies. *Risks* 8: 21. [CrossRef]
- Kuo, Kevin, and Daniel Lupton. 2020. Towards explainability of machine learning models in insurance pricing. *arXiv* arXiv:2003.10674.
- Masello, Leandro, German Castignani, Barry Sheehan, Montserrat Guillen, and Finbarr Murphy. 2023. Using contextual data to predict risky driving events: A novel methodology from explainable artificial intelligence. *Accident Analysis and Prevention* 184: 106997. [CrossRef]
- Mueller, Erik, J. S. Onésimo Sandoval, Srikanth Mudigonda, and Michael Elliott. 2018. A cluster-based machine learning ensemble approach for geospatial data: Estimation of health insurance status in Missouri. *ISPRS International Journal of Geo-Information* 8: 13. [CrossRef]
- Norman, Leslie George. 1962. *Road Traffic Accidents—Epidemiology, Control and Prevention*. Geneva: World Health Organization, pp. 47–60.
- OECD. 2020. The Impact of Big Data and Artificial Intelligence (AI) in the Insurance Sector. Available online: www.oecd.org/finance/Impact-Big-Data-AI-in-the-Insurance-Sector.htm (accessed on 28 December 2022).
- Paruchuri, Harish. 2020. The Impact of Machine Learning on the Future of Insurance Industry. *American Journal of Trade and Policy* 7: 85–90. [CrossRef]
- Pérez, Jesus Maria, Javier Muguerza, Olatz Arbelaitz, Ibai Gurrutxaga, and Jose Ignacio Martín. 2005. Consolidated tree classifier learning in a car insurance fraud detection domain with class imbalance. Paper presented at the International Conference on Pattern Recognition and Image Analysis, Bath, UK, August 22–25; Berlin/Heidelberg: Springer, pp. 381–89.
- Pesantez-Narvaez, Jessica, Montserrat Guillen, and Manuela Alcañiz. 2019. Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. *Risks* 7: 70. [CrossRef]
- Pesantez-Narvaez, Jessica, Montserrat Guillen, and Manuela Alcañiz. 2021. Risklogitboost regression for rare events in binary response: An econometric approach. *Mathematics* 9: 579. [CrossRef]
- Qazi, Maleeha, Kaya Tollas, Teja Kanchinadam, Joseph Bockhorst, and Glenn Fung. 2020. Designing and deploying insurance recommender systems using machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10: e1363. [CrossRef]
- Qazvini, Marjan. 2019. On the validation of claims with excess zeros in liability insurance: A comparative study. *Risks* 7: 71. [CrossRef]
- Rawat, Seema, Aakankshu Rawat, Deepak Kumar, and A. Sai Sabitha. 2021. Application of machine learning and data visualization techniques for decision support in the insurance sector. *International Journal of Information Management Data Insights* 1: 100012. [CrossRef]
- Reig Torra, Jan, Montserrat Guillen, Ana M. Pérez-Marín, Lorena Rey Gámez, and Giselle Aguer. 2023. Weather Conditions and Telematics Panel Data in Monthly Motor Insurance Claim Frequency Models. *Risks* 11: 57. [CrossRef]
- Roy, Riya, and K. Thomas George. 2017. Detecting insurance claims fraud using machine learning techniques. Paper presented at the 2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT), Kollam, India, April 20–21; pp. 1–6.
- Rustam, Zuherman, and Ni Putu Ayu Audia Ariantari. 2018. Support Vector Machines for classifying policyholders satisfactorily in automobile insurance. *Journal of Physics: Conference Series* 1028: 012005. [CrossRef]
- Sato, Kaz. 2017. Using Machine Learning for Insurance Pricing Optimization. Google Cloud. Available online: <https://cloud.google.com/blog/products/gcp/using-machine-learning-for-insurance-pricing-optimization> (accessed on 28 December 2022).
- SCOR. 2018. The Impact of Artificial Intelligence on the (Re)Insurance Sector. Focus SCOR. Available online: https://www.scor.com/sites/default/files/focus_scor-artificial_intelligence.pdf (accessed on 28 December 2022).
- Seely, S. 2018. Eight Use Cases for Machine Learning in Insurance. Azure. Available online: <https://azure.microsoft.com/en-us/blog/eight-use-cases-for-machine-learning-in-insurance/> (accessed on 28 December 2022).
- Selvakumar, V., Dipak K. Satpathi, Pravan P. Kumar, and Venkata Vajjha Haragopal. 2021. Predictive modeling of insurance claims using machine learning approach for different types of motor vehicles. *Accounting and Finance* 9: 1–14. [CrossRef]

- Shi, Peng, and Kun Shi. 2022. Non-Life Insurance Risk Classification Using Categorical Embedding. *North American Actuarial Journal* 27: 579–601. [CrossRef]
- Somani, Shymam. 2021. 17 Disruptive AI and Machine Learning Use Cases in Insurance World—AI and ML in Insurance Industry. Birlasoft. Available online: <https://www.birlasoft.com/articles/17-ai-and-ml-use-cases-insurance> (accessed on 28 December 2022).
- Wang, Yibo, and Wei Xu. 2018. Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decision Support Systems* 105: 87–95. [CrossRef]
- Wüthrich, Mario V., and Michael Merz. 2023. *Statistical Foundations of Actuarial Learning and Its Applications*. Berlin/Heidelberg: Springer Nature, p. 605.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.