mathematics

MDPI

Article

A Comparison between Explainable Machine Learning Methods for Classification and Regression Problems in the Actuarial Context

Catalina Lozano-Murcia [1,2,†], Francisco P. Romero [1,*], Jesus Serrano-Guerrero [1] and Jose A. Olivas [1]

[1] Department of Information Systems and Technologies, University of Castilla La Mancha, 13071 Ciudad Real, Spain; catalina.lozano@alu.uclm.es (C.L.-M.); jesus.serrano@uclm.es (J.S.-G.); joseangel.olivas@uclm.es (J.A.O.)
[2] Master Program in Actuarial Science, Escuela Colombiana de Ingeniería Julio Garavito, Bogota D.C. 205, Colombia
* Correspondence: franciscop.romero@uclm.es
† Current address: School of Computer Engineering, University of Castilla La Mancha, Paseo de la Universidad, 4, 13071 Ciudad Real, Spain.

Abstract: Machine learning, a subfield of artificial intelligence, emphasizes the creation of algorithms capable of learning from data and generating predictions. However, in actuarial science, the interpretability of these models often presents challenges, raising concerns about their accuracy and reliability. Explainable artificial intelligence (XAI) has emerged to address these issues by facilitating the development of accurate and comprehensible models. This paper conducts a comparative analysis of various XAI approaches for tackling distinct data-driven insurance problems. The machine learning methods are evaluated based on their accuracy, employing the mean absolute error for regression problems and the accuracy metric for classification problems. Moreover, the interpretability of these methods is assessed through quantitative and qualitative measures of the explanations offered by each explainability technique. The findings reveal that the performance of different XAI methods varies depending on the particular insurance problem at hand. Our research underscores the significance of considering accuracy and interpretability when selecting a machine-learning approach for resolving data-driven insurance challenges. By developing accurate and comprehensible models, we can enhance the transparency and trustworthiness of the predictions generated by these models.

Keywords: machine learning; artificial intelligence; explainable machine learning; accuracy; interpretability

MSC: 62R07

Citation: Lozano-Murcia, C.; Romero, F.P.; Serrano-Guerrero, J.; Olivas, J.A. A Comparison between Explainable Machine Learning Methods for Classification and Regression Problems in the Actuarial Context. *Mathematics* **2023**, *11*, 3088. https://doi.org/10.3390/math11143088

Academic Editors: Eric Ulm and Budhi Surya

Received: 1 May 2023
Revised: 8 July 2023
Accepted: 10 July 2023
Published: 13 July 2023

1. Introduction

Actuarial science, seeking risk modeling through mathematical and statistical techniques, faces new challenges every day, both in the volume of existing information to improve its modeling capacity and the nature of the different problems. The techniques associated with artificial intelligence (AI) and machine learning (ML) provide a series of tools whose purpose is to improve the processes of product design, pricing, reservations and the establishment of market niches practically and realistically [1]. However, a significant limitation exists in the practical application of complex models. The insurance industry is essential to the global economy when managing risks and public money, so its management is highly regulated. Owing to the stringent regulatory environment, non-replicable and non-auditable models pose considerable challenges, notwithstanding the benefits they confer regarding process efficiency or their robust predictive capabilities. Therefore, the main problems related to applying AI and ML techniques tend to be concentrated in the review and audit processes. As they are considered black boxes, it is impossible to guarantee that

a review process is adequate and presents an accessible development. Another common complaint concerns explaining the relationships to a non-technical person or an expert in the developed system and why the model is adequate. However, we cannot establish what relationships were created or why besides the model. Then, it is challenging to convey or communicate these complex models' benefits effectively.

In general, explainable artificial intelligence (XAI) defines methods and techniques to explain and understand the results or solutions proposed by ML models. In addition, XAI provides mechanisms that help identify the relationship's importance or strength between the variables input and target. The development of the XAI allows different users to formalize the definition of interpretability and reliability of AI models, which are an intermediate objective to verify various criteria in the end users of the results in the models made and informed decision making. The potential applications of explainable AI techniques would make it possible to understand and evaluate the capacity of the proposed models' results, thus facilitating the understanding and monitoring of the adequacy of these models. Therefore, these techniques allow us to translate into simple words that the general relationships establish the model or at least the relevance of the variables in the solution of the solved problem.

In actuarial science, conventional models and techniques are firmly established and still in use. However, an increasing trend towards integrating artificial-intelligence-based methods was observed in recent studies. Machine learning, particularly in the insurance sector, has expanded, encompassing everything from ANOVA methodologies to classification models using ensemble model techniques [2]. Tools such as the Markov decision process (MDP) [3] and artificial neural networks [4] are increasingly applied to boost accuracy and computational efficiency.

The interest in utilizing explainability strategies as a concluding phase in constructing machine learning models is increasing. These strategies encompass a range of techniques, including, but not limited to, the implementation of graphical analysis tools, such as LIME [5], SHAP [6], and partial dependence plots (PDPs) [7]. They present a visual representation of how variables affect the model's predictions. These approaches contribute towards a comprehensive understanding of the model's overall structure based on a combination of localized explanations for individual predictions [8]. Machine learning techniques applied to real-world finance and insurance issues must be transparent and repeatable to withstand audits and reviews. A significant challenge is creating a framework that simplifies the explainability analysis for complex or black-box models [9]. By enhancing the understandability of machine learning models, explainability techniques are instrumental in creating more transparent, accountable, and trustworthy AI systems. Their use could pave the way for the greater acceptance and utilization of machine learning models across various sectors and industries. Furthermore, this expansion in the use of explainability techniques signifies a critical evolution in machine learning, highlighting the need for models that not only predict well but are also interpretable and explainable [10]. Emphasis should be placed on the alternatives grounded in model-agnostic approaches, which enable the assessment of relationships between variables in an aggregated manner, thereby fostering a comprehensive understanding of the models under consideration [11]. However, less investigation was performed for other sensitive domains, such as the judicial system, finance and academia, in contrast with the domains of healthcare [12], industry [13] or other domains [14].

Our study introduces an innovative evaluation framework within the actuarial context, thoroughly assessing prevalent methods for explaining machine learning algorithms. Unlike previous studies, we examine these explainability techniques across various classification and regression problems and under diverse data scenarios. Our framework provides a comprehensive understanding of the implications of these algorithms in actuarial science. This new approach represents a significant advancement in the clarity, transparency, and utility of machine learning explainability techniques in the actuarial

domain, thereby outperforming the capabilities offered by conventional methods and previous scientific investigations.

The paper is organized as follows. In Section 2, the datasets are described, together with the machine learning and the XAI approach employed. Finally, in Section 3, we summarize and discuss the obtained results. Finally, Section 4 collects some conclusions and outlines future work perspectives.

2. Materials and Methods

This paper addresses the problem of identifying variables that may affect a machine learning model's decision (segmentation, pricing, and forecasting) in an actuary problem. We use XAI techniques to rank variables according to their relevance in decision making. The proposed approach is summarized in Figure 1.

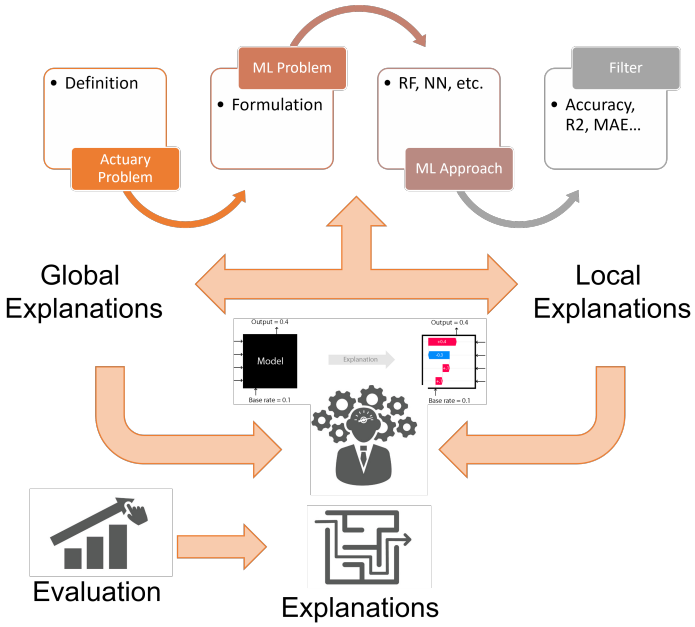

Figure 1. Approach outline.

The actuarial context poses unique challenges in defining and formulating problems. The first step is to clearly define the problem at hand. Once the problem is defined, machine learning (ML) can be leveraged to formulate the problem and build several datasets that represent different situations. A range of ML models can then be applied to solve the formulated problem, with an evaluation conducted according to the selected metrics. The filter process involves the removal of models with inadequate accuracy. To aid with global explanations, feature relevance is carefully considered in all predictions.

Evaluating machine learning algorithms under alternative scenarios is vital for understanding their functionality and potential enhancements. Detailed experimentation aids in feature determination and eliminating redundancies, leading to precision. Data normalization, feature selection, and outlier removal enhance model accuracy and robustness. A systematic examination across different scenarios informs the development of domain-specific models, such as those in actuarial science. This analysis also helps form best practices for real-world applications, supporting critical decision-making processes.

2.1. Datasets

This proposal is validated based on three insurance datasets. The first dataset is the Prudential Life Insurance [15] dataset, an anonymized collection of policyholder information used in a 2016 Kaggle competition to predict risk categories for insurance applicants. It contains various features, such as age, gender, medical history, and employment infor-

mation, to aid in building accurate predictive models. The second one is the Insurance dataset [16]; this dataset contains examples of beneficiaries currently enrolled in a health insurance plan, with features indicating the insured individual's characteristics and the total medical expenses charged for the calendar year. Finally, the Actuarial Loss Prediction Competition 2020/21 dataset [17] is used, which contains anonymized insurance policy and claim data for predicting claim payments and loss ratios. The dataset is intended to support the development of models that accurately forecast insurance losses, helping insurers make informed pricing and underwriting decisions.

2.1.1. Prudential Dataset

This dataset comes from the "Prudential Life Insurance Assessment" competition, published on the Kaggle platform in November 2005. Prudential is an insurer specializing in the life insurance segment worldwide and promotes the development of techniques to streamline the pricing process. Specifically, this contest sought to facilitate the rating of customers who purchase life insurance linked to products or purchases. In these processes, customers provide information for risk classification and insurance eligibility, considering variables such as age, gender and, in many cases, medical information that is corroborated by examinations.

The problem focuses on developing a classification model that accurately predicts the risk level of the potential client, using a more automated approach, allowing for greater efficiency in the risk selection and pricing process. The dataset provided for the development of the contest contains a training set and a testing set, on which the outcome is blindly evaluated. For our analysis, we made exclusive use of the dataset provided for training, which includes 128 variables describing the attributes of life insurance applicants (See Table 1), with an extension of 59,381 records. The task consists of predicting the variable "Response", an ordinal measure of risk with eight levels for each Id in the dataset.

Table 1. Prudential dataset variables.

Variable	Description
Id	Unique identifier associated with an application
Product_Info_1-7	Set of normalized variables related to the requested product
Ins_Age	Standardized age of the applicant
Ht	Applicant's standard height
Wt	Applicant's standardized weight
BMI	Applicant's normalized BMI
Employment_Info_1-6	Set of normalized variables related to the applicant's employment history
InsuredInfo_1-6	Set of normalized variables that provide information about the requester
Insurance_History_1-9	Set of normalized variables related to the applicant's insurance history
Family_Hist_1-5	Set of normalized variables related to the applicant's family history
Medical_History_1-41	Set of standardized variables relating to the applicant's medical history
Medical_Keyword_1-48	Set of dummy variables related to the presence or absence of a medical keyword associated with the application
Response	Target variable, an ordinal variable related to the final decision associated with a request

In order to understand the structure of the database, some approaches were proposed, such as identifying the number of products for which the request is made and their weight in the database. As shown in Table 2, the most requested product is D3, followed by D1 and D2, with 64% of the requests concentrated in type D.

Table 2. Product frequency.

Product	Count	% Data	Product	Count	% Data
A1	2363	3.98%	C1	285	0.48%
A2	1974	3.32%	C2	160	0.27%
A3	977	1.65%	C3	306	0.52%
A4	210	0.35%	C4	219	0.37%
A5	775	1.31%	D1	6554	11.04%
A6	2098	3.53%	D2	6286	10.59%
A7	1383	2.33%	D3	14,321	24.12%
A8	6835	11.51%	D4	10,812	18.21%
B1	54	0.09%	E1	2647	4.46%
B2	1122	1.89%			
Total	59.381				

The response variable has about 33% of the observations in category 8, followed by category 6 with 19% and 7 with 13%, with the lowest participation being in category 3 with less than 2% (see Figure 2).

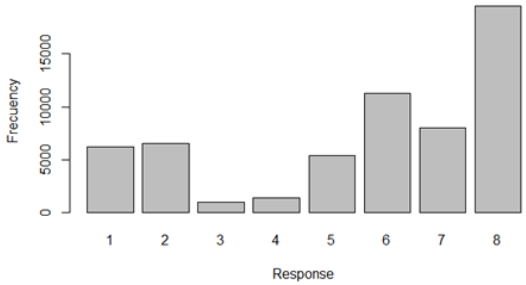

Figure 2. Predicted variable's behavior.

Finally, to understand the potential linear relationships that could exist between the continuous variables, a correlation analysis was developed, which corroborates the high relationship (0.85) between the variables *Weight* and *BMI*. In addition, as part of the exploratory analysis, the *NAs* were evaluated for each variable. The highest rates were obtained for the variables Medical_History{32,25,15} and Family_Hist_5. There are no missing values in the continuous variables, and it is only striking that Employment Info, which intuitively should be a reported field, has *NAs*.

Furthermore, four distinct working scenarios were devised to tackle the issue at hand, considering four separate datasets derived from the original testing base. These scenarios encompass variable selection as well as the transformation of the response variable, thereby offering a comprehensive approach to address the problem under investigation.

1. Dataset 1: The dataset is treated as numeric fields considering dummy variables for $Product\_info$. All missing values are replaced by zero.
2. Dataset 2: The dataset is treated as numeric fields considering dummy variables for $Product\_info$ and the creation of the variable Sum Medical_Keyword that sums the word count to reduce the processing fields. All missing values are replaced by zero.
3. Dataset 3: Dataset treated as numeric fields considering dummy variables for Product info2, creation of the variable *Sum_Medical_Keyword* that sums the word count, to reduce processing fields. Variables with missing values higher than 50% are eliminated. All missing values are replaced by zero.
4. Dataset 4: The complete base is treated as numeric fields considering dummy variables for Product info2, the creation of the variable *Sum Medical_Keyword* that sums the word count, to reduce processing fields. Variables with missing values greater than 50% are eliminated. All missing values are replaced by zero.

Thus, the problem can be categorized into two primary groups: datasets 1 and 4 represent an eight-category classification challenge, while datasets 2 and 3 embody a binary classification problem, distinguishing between favorable and unfavorable outcomes.

2.1.2. Health Insurance Dataset

The main purpose of this problem is to predict surcharges related to a health insurance dataset. The original dataset has 1338 observations. It has six variables in addition to the target variable. Of the explanatory variables, three categorical and three numerical variables were identified, of which three observations were eliminated due to inconsistencies in the value of the objective variable: age of the insured, gender of the insured (female or male), body mass index (BMI), number of children, smoker (yes) or non-smoker (no), and place of residence (northeast, northwest, southeast, and southwest). Figure 3 shows the predicted variable's behavior, with a minimum of 1.121, a mean of 13.270, a median equivalent of 9.382 and a maximum of 63.770. Among the descriptive variables, the participation of insured persons without children and non-smokers stands out as an equivalence between men and women. A relationship is also identified between the value of the surcharge and the number of children as well as smokers.

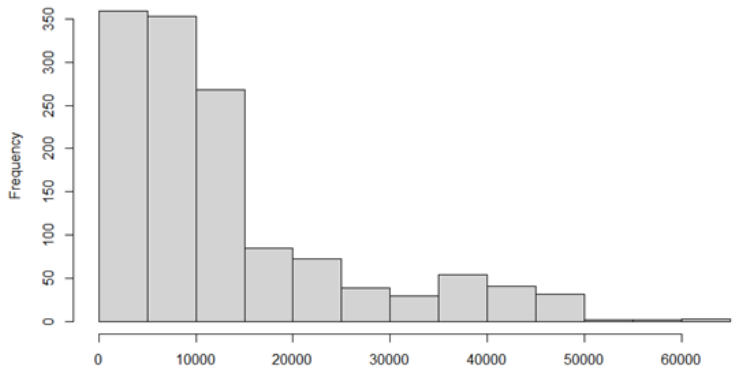

Figure 3. Predicted variable's behavior.

Furthermore, four distinct working scenarios were devised to tackle the issue at hand, taking into account four separate datasets derived from the original testing base in the same manner as the Prudential Life Insurance dataset.

- Dataset 1: considering age, gender (as dummy), body mass index, number of children, smoking (as dummy), and region (common-1 dummies).
- Dataset 2: taking the basis of dataset 1, excluding the variables where no significant parameters were identified in a linear regression exercise.
- Dataset 3: taking the basis of dataset 1, normalizing the numerical variables.
- Dataset 4: taking the basis of dataset 3, excluding the variables where no significant parameters were identified in a linear regression dataset.
- Dataset 5: based on dataset 1, excluding the values that were detected as outliers in the cost of surcharges.

2.1.3. Claims Dataset

The main objective of this problem is to predict the final total cost per claim of a group of occupational risk policies. We use a subset of an insurance dataset from OpenML ((ID 42876) https://www.openml.org/d/42876 (accessed on 1 March 2023)) synthetically generated by Colin Priest. It describes workers' compensation claims regarding their ultimate loss, the initial claim amount, and other information. However, this exercise does not seek to deepen or contextualize the problem in greater detail since the objective focuses on the models' prediction and interpretability.

The original database, comprising 90,000 records, was developed without regional or legal context for the 2020/21 claims prediction competition, promoted by the Institute of

Actuaries of Australia, the Institute and Faculty of Actuaries, and the Singapore Actuarial Society. It encompasses 14 explanatory variables (see Table 3) in addition to the target variable, and contains 36,176 observations. A total of 105 observations were excluded due to missing values and an extreme value four times higher than the subsequent maximum. Among the explanatory variables, two are dates, five are categorical or textual, and five are numerical. The final dataset consists of 36,176 observations, with three instances removed because of inconsistencies in the objective variable's value.

Table 3. Claims dataset variables.

Variable	Description
ClaimNumber	Unique policy identifier.
DateTimeOfAccident	Date and time of accident.
DateReported	Date that accident was reported.
Age	Age of worker.
Gender	Gender of worker.
MaritalStatus	Martial status of worker.
DependentChildren	The number of dependent children.
DependentsOther	The number of dependants excluding children.
WeeklyWages	Total weekly wage.
PartTimeFullTime	Binary (P) or (F).
HoursWorkedPerWeek	Total hours worked per week.
DaysWorkedPerWeek	Number of days worked per week.
ClaimDescription	Free text description of the claim.
InitialIncurredClaimCost	Initial estimate by the insurer of the claim cost.
UltimateClaimCost	Total claim payments by the insurance company.

In addition to the original variables, three new features are created to help explain the target variable (total value of the claim): (1) days elapsed between the date of the claim and the date of notice of claim, (2) year in which the loss occurred, and (3) month in which the loss occurred. It should be noted that the categorical variables were processed using *OneHotEncoding* for inclusion in the models.

As can be seen in Figure 4, a growth in the increase in the initial costs of claims over time is identified, as well as an increase in the dispersion of these in recent years. The average cost in thousands has grown significantly, which is economically related to the fact that health inflation in the last decade has grown above the average inflation in most of the world economies.

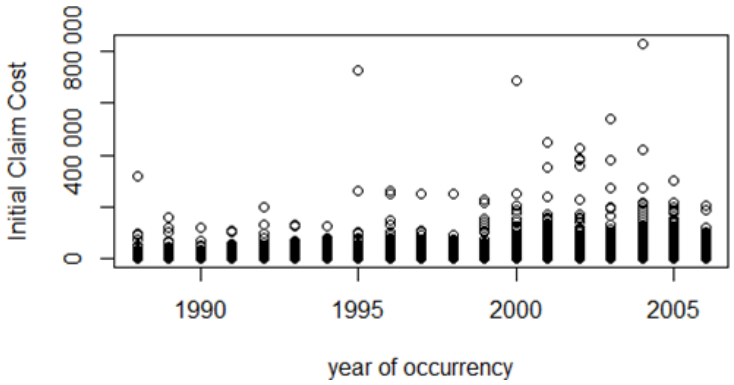

Figure 4. InitialClaimCost evolution.

Four distinct working scenarios were conceived with similar objectives as the previous datasets, yielding four separate resultant datasets as outcomes.

- Dataset 1: considering age, gender (M), marital status (single and married), dependent children, other, weekly salary, part-time work (yes), hours worked per week, the

month of occurrence, year of occurrence and days elapsed between occurrence and notice of loss.
- Dataset 2: taking the basis of dataset 1, the variables that in a linear regression exercise were not found to have significant parameters are excluded.
- Dataset 3: considering the normalized dataset excluding variables, no significant parameters were identified in a linear regression exercise.
- Dataset 4: considering dataset 1 and excluding the values that were detected as outliers in the cost of the claim.

2.2. Methods

In the following, we briefly present some well-established and popular techniques for solving classifications and regression problems.

2.2.1. Machine Learning Methods

Many different machine learning methods could be used to identify the source of a dataset and classify the source correctly. For our purposes, we chose the following ones:

- *General linear models*: general linear models (GLMs) are a family of statistical models (linear regression, ANOVA, and logistic regression) that analyze continuous outcomes while accounting for the effects of one or more predictor variables [18].
- *Decision trees* [19]: Establishes monotonous transformations of independent predictive variables with a determination of a recursive algorithm, divided into hyperrectangles, where each observation contained will have the same estimated value. The transformation results in a set of independent, monotonous variables, with the same estimated values for each observation contained in the hyper rectangles [20].
- *Artificial neural networks*: A neural network is a computational, parallel model composed of adaptive processing units with a high interconnection in them [21]. These computational models use processing units called neurons that process the information received through a connection called synapses, all in hidden layers that finally generate an output of either prediction or classification.
- *Ensemble methods*: bagging, random forests, and boosting form powerful machine learning tools that integrate multiple models to enhance predictive performance. Bagging, or bootstrap aggregation, ref. [22] uses various data samples to train multiple models, and final predictions are obtained by averaging the individual models' predictions, often reducing variance and improving robustness. Random forest, an instance of bagging, employs multiple decision trees, each casting a vote for a given instance's predicted label, enhancing the model's overall accuracy through the combination of diverse tree predictions [23]. Boosting, on the other hand, is an additive method that progressively improves a model's predictive capability by combining numerous weak predictors to form a stronger one, reducing prediction error [24].

2.2.2. Explainability Methods

Improving model transparency involves two approaches, model-specific explanations, which detail the factors used in predictions, and model-agnostic explanations, which provide general insights into how any model operates as the following:

- *Variable importance* (*VI*) [25] is a technique that measures the global contribution of each feature in a prediction model by analyzing the absolute value of feature weights. While determination methods may vary, making scores incomparable, the order of variables can still be used to compare established relationships.
- *Permutation-based variable importance* (*perm*) [26] is a model-agnostic method for evaluating feature significance in machine learning models. By randomly shuffling individual feature values and measuring the resulting decrease in model performance, this technique determines the relative importance of each feature based on the performance drop, revealing the most crucial features for accurate predictions.

- *SHAP* [27] is a game-theoretic approach for explaining machine learning model outputs using Shapley values. It assigns a SHAP value to each feature, representing its contribution to the model output. The method is model agnostic, consistent, and fast, making it suitable for various models and large datasets.
- *Feature importance ranking measure* (*FIRM*) [28] is a univariate feature selection technique that ranks features based on their variances, assuming a linear relationship between features and the target variable. While computationally efficient and easy to implement, FIRM does not consider feature interactions and may not be suitable for all data types or models.

2.3. Evaluation

There exist several attributes that, when present within a computational model, can substantially enhance its interpretability and elucidate the underlying decision-making processes for actuaries in the context of risk assessment and management [29]. These key characteristics include but are not limited to the following:

- *Feature importance*: Determine which features significantly impact the predictions and whether this feature's importance aligns with domain knowledge and expert intuition. To calculate the score based on the intersection of rankings, we can define the equation as

$$FI_{score}(E) = |\{f_i \mid f_i \in R_E \text{ and } f_i \text{ is among the top k values in } R_D\}| \quad (1)$$

  where E is the explainability method, R_E is the ranking of features provided by the explainability method E, represented as $f_1, f_2, ..., f_n$, and R_D is the ranking of features provided by domain experts.
- *Consistency*: The explanations should be consistent across different instances and similar inputs. This helps build trust in the explanations provided by the ML models. This score is based on the correlation of the explainability method results after its applications to different algorithms in the same scenario (Equation (2)):

$$C_{score}(E) = \frac{1}{m(m-1)/2} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \text{corr}_{\text{Spearman}}(E(A_i), E(A_j)) \quad (2)$$

  where $C_{score}(E)$ is the score of the explainability technique E, m is the number of different machine learning algorithms applied, and $E(A_i)$ and $E(A_j)$ are the results of applying E to algorithms A_i and A_j, respectively. $corr_{\text{Spearman}}()$ is the Spearman correlation coefficient between these results. The sums iterate over all unique pairs of different machine learning algorithms, and the whole equation averages over the number of these pairs.
- *Stability and robustness*: Check the stability of the explanations and the overall model across different training samples or data perturbations. Robust models should produce consistent results, even when the input data changes. Then, the correlation of the explainability technique among the different scenarios is computed (Equation (3)):

$$R_{score}(E) = \frac{1}{n(n-1)/2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{corr}(E(S_i), E(S_j)) \quad (3)$$

  where $Score(E)$ is the score of the explainability technique E, n is the number of different scenarios, and $E(S_i)$ and $E(S_j)$ are the results of applying E to the scenarios S_i and S_j, respectively. $corr()$ is the Pearson correlation coefficient between these results. The sums iterate over all unique pairs of different scenarios, and the whole equation averages over the number of these pairs.

- *Computation time and efficiency*: Some explainable ML techniques can be computationally expensive. The trade-offs between computation time and the quality of the explanations provided has to be evaluated using Equation (4):

$$\text{TO} = w_1 \cdot T + w_2 \cdot (1 - F_{score}) \tag{4}$$

where T denotes the computation time, F_{score} the quality of the explanation, and the weights $w1$ and $w2$ can be adjusted according to specific requirements. This equation assumes that a lower TO score is better (since lower computation time and higher F_{score} are desired). If F_{score} is perfect (equals 1), its contribution to TO is zero. If F_{score} is poor (approaching zero), its contribution to TO increases.

- *Fairness and bias*: It is necessary to evaluate the models for potential biases or the unfair treatment of certain groups, especially regarding sensitive attributes [30]. In this study, the general idea is to identify a set of fairness criteria (age, gender and race) and then evaluate whether the explanations provided satisfy these criteria for each sensitive feature (Equation (5)):

$$FA_{score} = \sum_{i=1}^{n} w_i \cdot F_i \tag{5}$$

where n is the total number of relevant fairness criteria, and w_i is the weight assigned to the ith fairness criterion based on its relevance or importance. F_i is a binary variable indicating whether the model satisfies the ith fairness criterion (1 for satisfying the criterion, 0 for not satisfying it).

- *Regulatory compliance*: Ensure that the explainable ML techniques and models you choose adhere to relevant actuarial and insurance regulations, such as the GDPR and CCPA. Measuring regulatory compliance with an explainability technique is largely a qualitative and procedural process rather than a quantitative one. It will depend heavily on the regulation specifics and the explainability technique's technical aspects. The number of regulations complied with from a specific setcould be considered a basic measure of regulatory compliance. However, it is important to remember that not all regulations are of equal importance or relevance, and not complying with a single critical regulation could potentially have significant consequences. Thus, weighting the regulations according to their importance or relevance to the problem is more informative when computing the compliance score:

$$C_{score} = \sum_{i=1}^{n} w_i \cdot C_i \tag{6}$$

where n is the total number of relevant regulations. w_i is the weight assigned to the ith regulation, based on its relevance or importance. C_i is a fuzzy variable indicating compliance with the ith regulation (1 for compliance, 0 for non-compliance).

After assessing each of the factors individually, we employed a ranking aggregation approach to provide an overall comparison of the models. This is based on the concept that while raw scores can vary across different metrics, the relative ranking of models within each criterion provides valuable insight into their performance. The models were ranked for each factor, with the best-performing model receiving the highest rank. Specifically, if there were n models under consideration, the best model for a given criterion received a rank of n, the second best a rank of $n - 1$, and so on, with the worst model receiving a rank of 1. These ranks were then aggregated across all the factors to provide an overall ranking of models. This ranking aggregation approach allows us to integrate insights from multiple factors into a single comparative framework. According to our multi-factor evaluation, the model with the highest overall rank score is considered the best model. This method assumes that all factors are equally important. If some factors are more important than others, it is necessary to assign weights to the ranks before aggregating them.

3. Results

This section presents the findings from our comprehensive analysis of the three distinct datasets, as previously described. The examination of these datasets allows us to draw meaningful conclusions and insights that contribute to our understanding of the underlying patterns and relationships. By evaluating the outcomes from each dataset, we aim to showcase the implications of our research and facilitate a better comprehension of the subject matter under investigation.

3.1. Prudencial Dataset

The optimal result for dataset 1 is achieved using boosting. In contrast, for datasets 2 and 3, random forest yields the most favorable outcomes, while dataset 4 exhibits the least desirable results, with no model attaining even 50% accuracy. It appears that the elimination of certain variables in this scenario directly impacts the categorization of the *Response*, as categories such as 1, 3, and 8 are devoid of cases in the testing exercises. The results are compared using the accuracy indicator from the testing outcomes for each model, taking into account the four proposed work scenarios (refer to Table 4). The best model for each scenario is highlighted in bold, while the second best is presented in italics.

Table 4. Machine learning models accuracy on the test set. The best model in bold.

Model	1	2	3	4
GLM	0.093	0.762	0.759	0.205
Decision Trees	0.482	0.781	0.781	**0.474**
Random Forest	0.236	**0.803**	**0.802**	0.213
XgBoost	**0.523**	0.772	0.773	0.156

As anticipated, datasets 2 and 3 exhibit better response levels since they involve binary classification problems, as opposed to the more complex prediction capacity required for multi-class classification problems. The boosting model delivers the best result for the original problem, aligning with the competition outcomes, and appears to be the most suitable model, despite not being directly comparable to the problem statement.

Notably, dataset 4 demonstrates a considerably low accuracy level, with no model identifying all categories during the testing phase, typically capturing only 5 to 6 categories. The least represented categories, 1 and 3, were not identified in this phase, suggesting that some variables eliminated due to missing values may hold crucial information for the classification process.

Interestingly, a model such as the classification tree model, which allows for variable importance evaluation through coefficients, ranks second for datasets 1 to 3, outperformed by more challenging-to-interpret models, like random forest and XGBoost.

As expected, accuracy results for binary classification are favorable in more interpretable models like decision tree classifiers, though surpassed by more sophisticated techniques. In the case of 8-category classification bases, specifically for dataset 1, acceptable and consistent results are achieved, with the best technique aligning with those obtained by the competition winners, from which the base was extracted.

3.1.1. Explainable Results

All models, excluding GLM, generally exhibit a predominance of shared variables, such as BMI, WT, Medical_History_4, and Product_info 4, albeit in varying orders. The best-performing models demonstrate a similar ranking, which is further elaborated upon below. Concerning variable relevance, shared variables prevail for dataset 1, including BMI, WT, Medical_History_4, and Product_info 4, with BMI being the most significant for models other than GLM (see Figure 5).

Concerning variable importance, common variables dominate for dataset 1, such as BMI, WT, *Medical_History*_4, and *Product_info*_4. Notably, BMI is the most crucial factor for models excluding GLM.

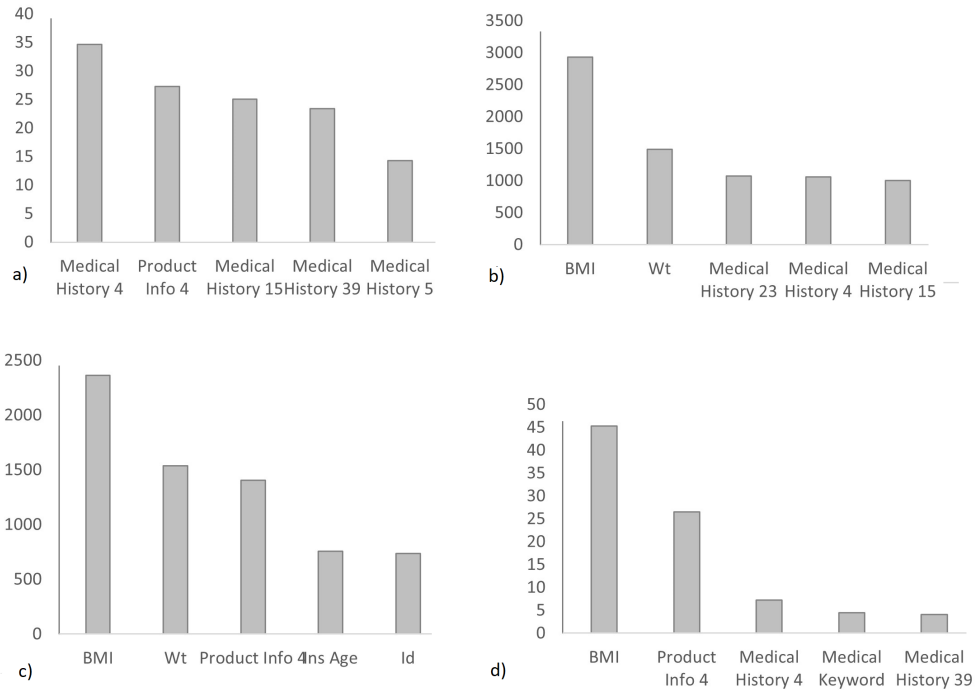

Figure 5. VIPs dataset 1. Left to right, up to down. (**a**) VIP GLM; (**b**) VIP decision tree classifier; (**c**) VIP random forest; and (**d**) XG boost.

Related to the importance of the variables in the different models of dataset 2, BMI, WT and Medical_History_4 and Product_info 4 predominate, being the most important for models other than GLM, BMI, with the participation of additional variables such as Medical_History_39 and Ins_Age.

The most relevant variables in dataset 3 are like those described in dataset 2, considering the same order in each model. There are no novelties regarding the importance of the variables in the dataset 4 models, considering BMI, WT Medical_History_4, and Product_info as the most important for models other than GLM and BMI, with the participation of additional variables such as Medical_History 39 and Ins_Age.

In general, and considering the limitation of information related to product, medical history, family information, whose actual content is unknown, about variables relevant to life coverage, such as BMI and weight, which, together with the age of a person, could give a signal of good or bad health, closely related to the risk of mortality. On the other hand, there are three variables whose content is unknown, but they seem to have information characteristic of the population analyzed in any of the approaches, such as Medi-cal_History_4, Product_info 4 and Medical_History 39, being common variables in terms of the analysis of the alternatives provided by means of the databases constructed.

3.1.2. Evaluation

According to feature importance, the first dimension in our evaluation framework (see Section 2.3), at least three of the five most relevant variables per model and per scenario are within the a priori most relevant groups (see Table 5).

Table 5. FI_{score} Evaluation of each model according to VI explainability measure.

Scenario	GLM	Decision Tree	Random Forest	Boosting
1	0.2	0.6	0.8	0.6
2	0.2	0.6	0.6	0.8
3	0.2	0.6	0.6	0.6
4	0.2	0.6	0.8	0.8

Regarding consistency, only boosting achieves a correlation of at least 0.5 among two or more explainability techniques across all datasets. For instance, decision trees reach this consistency level in two datasets (3 and 4), while GLM and random forest attain it in only one dataset (2 and 4, respectively).

Concerning stability and robustness, the between-variable importance (VI) of each model is computed for the four scenarios (see Table 6). A similar analysis was conducted by comparing the results using dataset 1. VI with AI techniques yields high correlations, exceeding 75%, whereas GLM models produce notably lower VI values.

Table 6. FI_{score} Evaluation of each model according to VI explainability measures.

	GLM	Decision Tree	Random Forest	Boosting
Minimum Correlation	0.25	0.78	0.72	0.86
Maximum Correlation	0.59	0.92	0.96	0.93

The *model* measure attains the best results regarding computational cost and availability. While the *firm* and *shap* techniques apply to every ML technique, their execution is slow. In contrast, the execution of perm is fast, but it is not available for all models and data scenarios.

From a regulatory compliance perspective, this dataset adheres to the GDPR, as it is anonymized and precludes the association of characteristics for inferring personal information. Moreover, in insurance pricing ease, we have a model for defining homogeneous groups that, although only partially replicable, would yield similar results when repeated. This is further reinforced by the XAI analysis, which facilitates understanding and review by the regulator.

Regarding fairness and bias, all analyses reveal that BMI, Wt, and certain health-related variables influence the classification process. For example, if we consider that the rating process aims to evaluate the allocation of insurance risk associated with financial products, the identified relevance reaffirms these characteristics, despite their seemingly discriminative nature. Furthermore, the dataset does not include variables penalized for discrimination in insurance pricing, such as sex.

3.2. Health Insurance Results

After running the parameter optimization process for each group of techniques, the parameterization that gave the best results for each of the models in each dataset is selected. The results obtained can be seen in the following Table 7.

In dataset 2 (see Table 8), the algorithms were tested with the same dataset after a previous feature selection process. As a result, their results remained the same. They even worsened slightly, except for the neural networks, which significantly improved. However, reducing the dataset, even more, does not bring significant improvements for a problem with a small dataset.

The results are similar in the scenarios with normalized data with (Table 9) or without outliers (Table 10). However, the predictive power of the neural networks is even more outstanding, obtaining the best MAE values in the scenario without outliers.

Table 7. Health insurance machine learning results. Dataset 1.

Method	Cross Validation		Test	
	R^2	MAE	R^2	MAE
Random Forests	0.85	2447.96	0.88	2326.47
Bagging	0.84	2593.46	0.88	2295.47
ANN	0.82	2587.77	0.87	2378.30
Decision Tree	0.78	3318.35	0.85	3027.99
Boosting Trees	0.74	3471.32	0.84	2888.02
Linear Reg.	0.68	4267.66	0.77	4194.54

Table 8. Health insurance machine learning results. Dataset 2.

Method	Cross-Validation		Test	
	R^2	MAE	R^2	MAE
Random Forests	0.84	2660.93	0.88	2403.53
Bagging	0.85	2595.70	0.88	2294.15
ANN	0.83	2239.63	0.87	2208.04
Decision Tree	0.79	3305.79	0.86	2993.71
Boosting Trees	0.74	3353.79	0.82	2933.78
Linear Reg.	0.70	4597.26	0.72	4933.10

Table 9. Health insurance machine learning results. Dataset 3.

Method	Cross-Validation		Test	
	R^2	MAE	R^2	MAE
Random Forests	0.85	0.211	0.88	0.19
Bagging	0.85	0.210	0.88	0.19
ANN	0.82	0.182	0.86	0.18
Decision Tree	0.78	0.274	0.85	0.25
Boosting Trees	0.75	0.268	0.81	0.25
Linear Reg.	0.69	0.351	0.77	0.25

Table 10. Health insurance machine learning results. Dataset 4.

Method	Cross-Validation		Test	
	R^2	MAE	R^2	MAE
Random Forests	**0.83**	0.213	**0.85**	0.20
Bagging	0.82	0.213	0.86	**0.19**
ANN	0.82	**0.210**	0.83	0.20
Decision Tree	0.81	0.246	0.84	0.23
Boosting Trees	0.79	0.206	0.82	0.25
Linear Reg.	0.69	0.365	0.71	0.34

The performance results obtained in the scenario without outliers (Table 11) are worse regarding R^2 in both the cross validation and test. They are better in MAE in cross validation (slight overlearning) but worse in the test step. The best results are obtained by neural networks that take advantage of not having to handle these extreme cases to obtain good results.

Table 11. Health insurance machine learning results. Scenario 5.

Method	Cross-Validation		Test	
	R^2	MAE	R^2	MAE
Random Forests	0.68	2267.03	0.58	2641.42
Bagging	0.67	2320.13	0.60	2435.21
ANN	0.64	1879.13	0.57	2046.13
Decision Tree	0.60	2878.58	0.57	2971.81
Boosting Trees	0.55	2865.61	0.45	3054.34
Linear Reg.	0.61	2743.45	0.58	2601.31

In most cases, decision forests (random or bagging) achieve the best results both in cross validation and evaluation. The optimized models that achieve the best results are complex models without a reduced capability to provide explanations. Artificial neural networks achieve good predictive results with the best MAE in the dataset with feature selection (Scenario 2) or without outliers (Scenario 5). These techniques present outstanding predictive performance (MAE) with lower performance in description power (R^2) and difficulty explaining the results. In addition, self-explainable techniques like linear regression or decision trees achieve the worst results, showing poor performance in any scenario. However, the poor performance of the boosting algorithms is remarkable; these methods need a considerable volume of data to achieve good performance.

Explainability Results

According to this method, if we study the correlation between the results of the variables' relevance and the results of the different algorithms (see Table 12), the result is very similar, presenting very high correlations (average = 0.97). From another point of view, analyzing the relevance of the variables for the same algorithm but for different datasets, the result is also very high, with the highest differences compared to *scenario 5*. Artificial neural networks are the most stable in terms of the correlation of results.

Table 12. Correlation of feature importance provided by the same algorithm in different datasets.

Scenarios	RF	Bagging	ANN	DT	Boosting
Esc 1 vs. Esc 2	0.9997	0.9997	0.9948	1.0000	0.9659
Esc 1 vs. Esc 3	0.9995	0.9996	0.9946	1.0000	0.9743
Esc 1 vs. Esc 4	0.9984	0.9998	0.9912	0.9999	0.9919
Esc 1 vs. Esc 5	0.8842	0.8544	0.9601	0.8721	0.9184
Esc 2 vs. Esc 3	0.9999	0.9999	0.9891	1.0000	0.9984
Esc 2 vs. Esc 4	0.9986	0.9998	0.9807	0.9999	0.9783
Esc 2 vs. Esc 5	0.8749	0.8527	0.9424	0.8711	0.8936
Esc 3 vs. Esc 4	0.9990	1.0000	0.9972	0.9999	0.9847
Esc 3 vs. Esc 5	0.8725	0.8581	0.9664	0.8721	0.9091
Esc 4 vs. Esc 5	0.8754	0.8582	0.9815	0.8768	0.8886
Average	0.9502	0.9422	0.9798	0.9492	0.9503
Std. Dev	0.0633	0.0744	0.0181	0.0656	0.0429

Table 13 reflects different results for the models obtained. However, the relevance of *smoker.no* is as the most important in all models, and *BMI* and *age* are third place in several of the models.

Analyzing the features that significantly impact predictions, it is noteworthy that at least three of the five most relevant variables per model fall within the a priori relevant groups. Moreover, the explanations are consistent across different instances and similar inputs, as all correlations between various techniques in each scenario exceed 85%. This demonstrates that the explainability techniques align with domain knowledge in this case.

Table 13. Variable importance by model.

Model	Children	Age	bmi	smoker.no	Southeast	Female
Random Decision Forest	1.25%		11.37%	14.63%	68.27%	
Bootstrap Decision Forest	1.19%	11.45%	17.88%	66.96%		
Neural Network 1		16.14%	12.82%	51.73%	3.99%	4.31%
Neural Network 2	3.31%	17.21%	12.04%	56.23%		3.26%
Decision Tree	3.40%	16.38%	24.56%	39.94%		

The *model* measure achieves the most optimal results in terms of computational cost and availability. While the *firm* and *shap* techniques apply to all machine learning techniques, their execution is slow. In contrast, the *perm* technique executes quickly but is unavailable for all models and data scenarios.

All models concur that BMI, age, and smoking condition influence the fitting process (see Figure 6). Considering that the rating process aims to evaluate the allocation of insurance risk associated with financial products, the identified relevance reaffirms these characteristics, despite their seemingly discriminative nature. Conversely, the dataset includes variables such as sex and paternity, which are penalized as discriminatory for insurance pricing, but reflect actual conditions according to experience.

The database adheres to the GDPR, as it is anonymized and precludes the association of characteristics for inferring personal information.

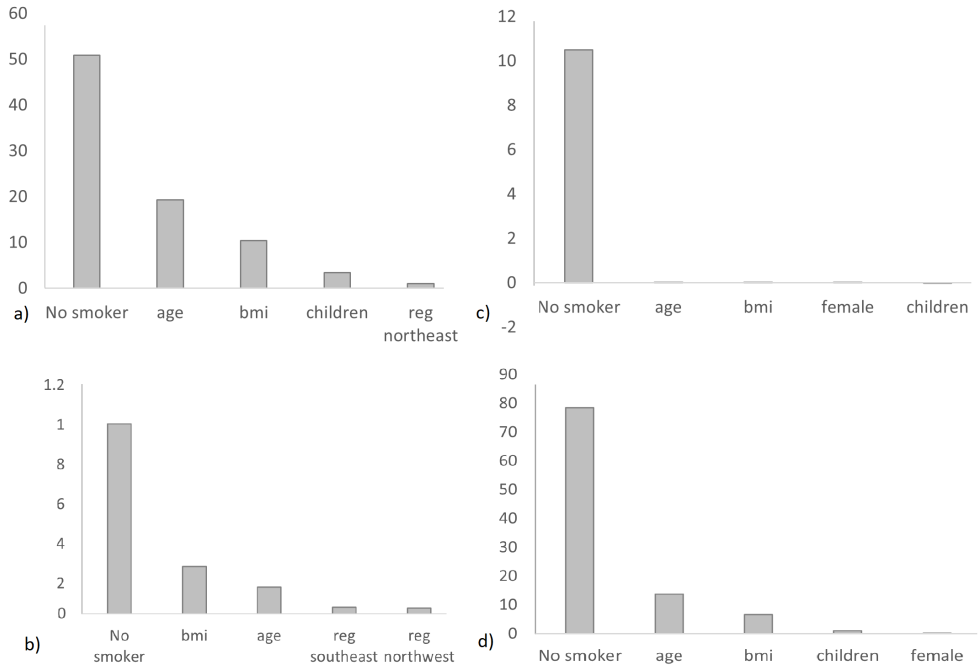

Figure 6. VIPs dataset 1. Up to down, Left to right (**a**) VIP regression; (**b**) VIP artificial neural network; (**c**) VIP random forest; and (**d**) VIP boosting.

3.3. Claim Results

The results of our analysis, as presented in Table 14, reveal some interesting insights. Firstly, it is clear that the performance of the machine learning algorithms varies across the different datasets. For example, regression trees performed best in dataset 1, with an error rate of 11,538.02, while Bagging achieved the lowest error rate in dataset 3, with an error rate of 0.423. On the other hand, boosting performed the best in dataset 4, with an error rate of 2960.106.

It is worth noting that neural networks consistently performed well across all datasets. This highlights the versatility and robustness of neural networks, making them a viable

option for a wide range of applications. Another interesting observation is the relatively high error rates in datasets 1 and 2 compared to datasets 3 and 4. This can be attributed to datasets 1 and 2 being multi-class classification problems, which are generally more challenging than binary classification problems, such as those in datasets 3 and 4.

Our results demonstrate the importance of carefully considering the problem and selecting an appropriate machine learning algorithm to achieve optimal results. While there is no one-size-fits-all solution, the versatility of neural networks and the varied strengths of different algorithms in different datasets highlight the importance of conducting thorough experimentation and analysis to identify the best solution for each problem.

Table 14. MAE results by model.

Dataset	Regression Trees	Neural Networks	Bagging	Boosting
1	11,538.02	11,680.67	11,476.85	11,314.6
2	12,002.64	11,648.35	11,958.65	11,710.33
3	0.424	0.414	0.423	0.414
4	3857.927	2974.121	3022.765	2960.106

Explainability Results

The analysis of the prediction techniques used in the study show that the ANN model with linearly significant and untransformed variables had a preponderance for the insured's age and marital status, whereas work-related variables were less relevant. However, this model included a variable that was previously discarded in the other datasets: the year of occurrence, which is potentially related to inflation or the growth of the average cost of the claim over time. Additionally, the boosting model, which considered the normalized base without outliers or variables with no linear relationship, reinforced the importance of the most relevant variables from the previous relationships, such as weekly income, year of occurrence, age, and gender.

It should be noted that the incidence of the variables in the models cannot be understood solely from the results obtained from each algorithm. Despite the limitations of information, such as the type of work or cause of the accident, the study was able to clearly identify the influence of weekly income, age, and year of occurrence. The cost of an accident at work is directly related to the injured worker's salary (income), which is usually related to their experience level, making sense that age matters. Furthermore, the year provides a reference to the influence of the value of money over time or the inflationary effect in wages. The results of the study are presented in Figure 7.

The boosting machine learning (ML) algorithm demonstrated its effectiveness in generating consistent and robust explanations through various explainability techniques. The high level of correlation among the results of these techniques further strengthens the credibility of the explanations. Moreover, the variable selection process achieves high consistency and robustness across different datasets, indicating the reliability of the approach.

In terms of computational cost and availability, the *model* measure outperforms the *firm* and *shap* techniques, which are slow. In contrast, the *perm* technique executes quickly but is not universally available for all models and datasets.

From a regulatory compliance perspective, the model provides a suitable fit that may only be partially replicable, but repetition would lead to similar results. The XAI analysis further facilitates understanding any review by the regulator, reinforcing the model's reliability. The analysis of the results reveals the significance of the weekly income, age, and year of occurrence as influential variables in solving the problem. These variables align with the criteria of a claims specialist, as the cost of the accident claim is directly related to the salary (income) of the injured worker, which in turn is usually associated with the worker's experience level. It makes sense that age matters in such cases, and the year of

occurrence refers to the influence of the value of money over time or the inflationary effect on salaries. Therefore, no weight is given to any discriminant variable.

Figure 7. VIP results.

4. Conclusions

There is a lack of consensus on the best approach to developing explainable machine-learning methods for data-driven insurance problems. Therefore, this paper addresses this issue by comparing different explainable machine learning methods for solving classification and regression problems in the actuarial context. The goal is to develop accurate and interpretable models, enhancing the transparency and trustworthiness of the predictions generated by these models. Three insurance datasets were used to validate the proposed approach and assess the quality of the explanations provided by each method. XAI techniques were employed to facilitate an understanding of the relationships established by the models and their intuitive contrast depending on the context of the problem. We further enhanced the robustness of our approach by applying these algorithms in various data scenarios, including contexts with and without outliers, as well as binary classification problems. This exhaustive and versatile application under diverse conditions underscores our framework's adaptability and potential utility in elucidating the functionality of machine learning within actuarial science. Overall, the machine learning algorithms showed varying performance across the different datasets. However, boosting techniques effectively generated consistent and robust explanations through various explainability techniques. For example, the variable selection process achieved high consistency and robustness across different datasets, indicating the reliability of the approach.

The research shows that different XAI methods vary in accuracy, implying that organizations must carefully select the appropriate method for each problem. An accurate model can significantly improve decision making, ranging from risk assessment to customer segmentation. Moreover, by employing explainable AI (XAI) methods, insurance companies can ensure higher transparency and trust in their predictive models. This could lead to more acceptance from stakeholders, such as regulators and customers.

The results demonstrate that the most noticeable method is SHAP, which provides insights into feature relevance through local and global feature importance measures due to its consistency, interpretability, and model-agnostic nature. Firstly, it aligns with actuarial practice by providing a fair and consistent allocation of a prediction across its features, similar to how actuaries distribute risk. Secondly, SHAP facilitates interpretability and transparency by explaining the influence of each feature on the output of any machine learning model, which is crucial in actuarial work that often requires clear explanations for stakeholders and regulatory authorities. Lastly, being model-agnostic, SHAP can be applied to any machine learning model, offering invaluable flexibility in actuarial contexts, where different models might be used depending on the problem.

The findings of this study contribute to the ongoing development of explainable machine-learning methods for data-driven insurance problems and provide a roadmap for future research in this area. However, the choice of an explainability technique should be context dependent, considering the model's complexity, data characteristics, computational

resources, and specific task needs. Implementing a process to continuously evaluate the performance of different XAI methods as new data become available and as the business context changes will help ensure that the selected method remains optimal over time.

Author Contributions: Conceptualization, J.S.-G.; Software, C.L.-M.; Validation, J.S.-G.; Data curation, C.L.-M.; Writing—original draft, F.P.R.; Writing—review & editing, F.P.R.; Supervision, J.A.O.; Funding acquisition, J.A.O. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partially supported by FEDER and the State Research Agency (AEI) of the Spanish Ministry of Economy and Competition under grant SAFER: PID2019-104735RB-C42 (AEI/FEDER, UE).

Data Availability Statement: Publicly available datasets were analyzed in this study.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
MAE	Mean Absolute Error
ML	Machine Learning
XAI	Explainable Artificial Intelligence

References

1. Shapiro, A.F. Fuzzy logic in insurance. *Insur. Math. Econ.* **2004**, *35*, 399–424. [CrossRef]
2. Henckaerts, R.; Cote, M.P.; Antonio, K.; Verbelen, R. Boosting insights in insurance tariff plans with tree-based machine learning methods. *N. Am. Actuar. J.* **2021**, *25*, 255–285. [CrossRef]
3. Krasheninnikova, E.; García, J.; Maestre, R.; Fernández, F. Reinforcement learning for pricing strategy optimization in the insurance industry. *Eng. Appl. Artif. Intell.* **2019**, *80*, 8–19. [CrossRef]
4. Kovalnogov, V.N.; Fedorov, R.V.; Generalov, D.A.; Chukalin, A.V.; Katsikis, V.N.; Mourtas, S.D.; Simos, T.E. Portfolio insurance through error-correction neural networks. *Mathematics* **2022**, *10*, 3335. [CrossRef]
5. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
6. Van den Broeck, G.; Lykov, A.; Schleich, M.; Suciu, D. On the tractability of SHAP explanations. *J. Artif. Intell. Res.* **2022**, *74*, 851–886. [CrossRef]
7. Wadoux, A.M.C.; Molnar, C. Beyond prediction: Methods for interpreting complex models of soil variation. *Geoderma* **2022**, *422*, 115953. [CrossRef]
8. Alonso Robisco, A.; Carbó Martínez, J.M. Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction. *Financ. Innov.* **2022**, *8*, 1–35. [CrossRef]
9. Blier-Wong, C.; Cossette, H.; Lamontagne, L.; Marceau, E. Machine learning in P&C insurance: A review for pricing and reserving. *Risks* **2020**, *9*, 4.
10. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef]
11. Kshirsagar, R.; Hsu, L.Y.; Greenberg, C.H.; McClell, M.; Mohan, A.; Shende, W.; Tilmans, N.P.; Guo, M.; Chheda, A.; Trotter, M.; et al. Accurate and Interpretable Machine Learning for Transparent Pricing of Health Insurance Plans. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 15127–15136. [CrossRef]
12. Du, Y.; Rafferty, A.R.; McAuliffe, F.M.; Wei, L.; Mooney, C. An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus. *Sci. Rep.* **2022**, *12*, 1170. [CrossRef] [PubMed]
13. Islam, M.R.; Ahmed, M.U.; Barua, S.; Begum, S. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Appl. Sci.* **2022**, *12*, 1353. [CrossRef]
14. Clement, T.; Kemmerzell, N.; Abdelaal, M.; Amberg, M. XAIR: A Systematic Metareview of Explainable AI (XAI) Aligned to the Software Development Process. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 78–108. [CrossRef]
15. Prudential. February 2016. Available online: https://www.kaggle.com/c/prudential-life-insurance-assessment/data (accessed on 15 December 2022).
16. Lantz, B. *Machine Learning with R: Expert Techniques for Predictive Modeling*; Packt Publishing Ltd.: Birmingham, UK. 2019.

17. Priest, C. (2021, November) Actuarial Loss Prediction Competition 2020/21. Available online: https://www.kaggle.com/competitions/actuarial-loss-estimation/overview (accessed on 15 December 2021).
18. Tomasevic, N.; Gvozdenovic, N.; Vranes, S. An overview and comparison of supervised data mining techniques for student exam performance prediction. *Comput. Educ.* **2020**, *143*, 103676. [CrossRef]
19. Ben-Haim, Y.; Tom-Tov, E. A Streaming Parallel Decision Tree Algorithm. *J. Mach. Learn. Res.* **2010**, *11 849-872* .
20. Xie, Y.; Schreier, G.; Chang, D.C.; Neubauer, S.; Redmond, S.J.; Lovell, N.H. Predicting number of hospitalization days based on health insurance claims data using bagged regression trees. In Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, IL, USA, 26–30 August 2014; pp. 2706–2709.
21. Goundar, S.; Prakash, S.; Sadal, P.; Bhardwaj, A. Health Insurance Claim Prediction Using Artificial Neural Networks. *Int. J. Syst. Dyn. Appl.* **2020**, *9*, 40–57. [CrossRef]
22. Yao, J.; Yu, S.; Wang, C.; Ke, T.; Zheng, H. Medicare fraud detection using a bagging algorithm. In Proceedings of the 2021 7th International Conference on Computer and Communications (ICCC), Chengdu, China, 10–13 December 2021; pp. 1515–1519.
23. Lin, W.; Wu, Z.; Lin, L.; Wen, A.; Li, J. An ensemble random forest algorithm for insurance big data analysis. *IEEE Access* **2017**, *5*, 16568–16575. [CrossRef]
24. Fauzan, M.A.; Murfi, H. The accuracy of XGBoost for insurance claim prediction. *Int. J. Adv. Soft Comput. Appl.* **2018**, *10*, 159–171.
25. Greenwell, B.M.; Boehmke, B.C.; Gray, B. Variable Importance Plots—An Introduction to the vip Package. *R J.* **2020**, *12*, 343. [CrossRef]
26. Molnar, C.; Casalicchio, G.; Bischl, B. Interpretable machine learning—A brief history, state-of-the-art and challenges. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* ; Springer International Publishing: Cham, Switzerland, 2021; pp. 417–431.
27. Sohail, M.; Peres, P.; Li, Y. Feature importance analysis for customer management of insurance products. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–8.
28. Scholbeck, C.A.; Molnar, C.; Heumann, C.; Bischl, B.; Casalicchio, G. Sampling, intervention, prediction, aggregation: A generalized framework for model-agnostic interpretations. In *Machine Learning and Knowledge Discovery in Databases, Proceedings of the International Workshops of ECML PKDD 2019, Würzburg, Germany, 16–20 September 2019*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 205–216.
29. Saranya, A.; Subhashini, R. A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. *Decis. Anal. J.* **2023**, *7*, 100230.
30. Angerschmid, A.; Zhou, J.; Theuermann, K.; Chen, F.; Holzinger, A. Fairness and explanation in ai-informed decision making. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 556–579. [CrossRef]