JOURNAL OF INTERNATIONAL CRISIS AND RISK COMMUNICATION RESEARCH
ISSN: 2576-0017
2024, VOL 7, NO S12

Validation of Health Insurance Customers Using XGBoost Modeling

Saeed Shouri [1,*]

[1] Ph.D. Candidate in Economics, Ferdowsi University of Mashhad, Iran; saeedshouri63@gmail.com
*Correspondence: saeedshouri63@gmail.com

Abstract: Given the inadequate risk assessment of policyholders in the insurance industry, particularly in health insurance, there is a significant emphasis on the validation modeling for customers' creditworthiness. Therefore, the current study aimed to provide the modeling for health insurance customers validation, with a specific focus on individuals covered by health insurance, particularly employees of the East Iran Oil Company. In this study, method XGBoost using machine learning were employed as the top artificial intelligence methods for customer validation.
Notably, the validation process identified approximately 1.78% of the population as "unhealthy." This seemingly small group accounts for a disproportionately high 17.47% of the company's total health insurance claims. After the training process, the designed model was evaluated using various metrics, including standard metrics, such as Accuracy, Precision, Recall, and F-measure, each examining specific features of the model. The values of these metrics were 0.999, 0.992, 1, and 0.996, respectively. These values were indicative of the very high accuracy, precision, and efficiency of the model. This type of validation model is one of the most practical modeling approaches that insurance companies can use to validate their customers in order to pay an insurance premium in proportion to the level of risk.
Keywords: Validation, health insurance, Claims-Based Risk, XGBoost

1. Introduction

The insurance industry is considered as a tool for transferring uncertainty and risk in a society and an efficient and strong financial intermediary. Therefore, insurance in a society is generally considered an economic and social desirable phenomenon [1,2]. One of the most important categories of the insurance industry is health insurance, which holds particular significance among other insurance topics as it is directly related to the health and well-being of society [3].

An increasing rise in healthcare costs puts significant pressure on the economies of developed and developing countries, a challenge worsened by population aging and advancements in health technology [4]. Moreover, Hamid (2024) also discusses that there are many benefits of healthcare insurance programs but fraud in healthcare continues to be a significant challenge in the insurance industry [5,6,7]. However, determining health insurance premiums based on accurate risk assessment can help reduce the abuse of insurance companies [8].

Recently, health insurance claims have regained attention in healthcare research and quality improvement as a social reality. However, few studies have examined the validity of health insurance claims so far [9].

Keyur et al. (2014) define Process validation can also be defined as the collection and evaluation of data, from the process design stage, that provides scientific evidence that a process is capable of consistently delivering a quality product [10,11]. Types of validation can be distinguished: prospective validation, concurrent validation, Retrospective validation [11,12].

Prospective validation refers to all the activities carried out before the distribution of new products to ensure compliance with the initial conditions (legal / recommended / etc.) by the product features [11,13]. Concurrent validation is issued to create documentary evidence during the actual assignment of the process to show that the process is in control. Retrospective process validation is based on the review of historical production and test data, and the analysis of accumulated results from past production to assess the consistency of a process [10,13].

In this study, the validation is based on a scientific approach that includes the analysis of current information and the history of insurance applicants to assign a credit score based on the level of health risk, and therefore provides the possibility of classification. To check the validity of a computational model, the validity of the model determines the degree to which the model is an accurate representation of the real phenomenon. Of note, model validation evaluates the accuracy of a computational model and improves the model's process based on the validation results [14]. Validation ensures that a product, service, or system (or a part thereof) leads to another product, service, or system (or a part thereof) that meets operational needs [15]. In other words, validation is a proper system, which is required to have high reliability, a specific process, and predefined specifications [16].

Vabalas (2019) offers a reliable way to validate the performance of a Machine learning (ML) model is to train a model with existing data and evaluate its classification performance using newly collected data or a separate dataset. Another reliable method, commonly called train/test splitting, is to isolate a portion of the data before developing an ML model and use that data only for validation [17]. Machine learning is a powerful tool for gleaning knowledge from massive amounts of data [18].

The current study aimed to provide a health insurance customer validation for National Iranian Oil Company employees through an assessment based on the characteristics of insured individuals, using an Extreme Gradient Boosting (XGBoost) algorithm.

The structure of the article is as follows: The first section undertakes an analysis of health insurance validation. The second section reviews the research background, while the third section presents the theoretical foundations of the research. The fourth section deals with the introduction and analysis of research data and the fifth section details the design and modeling of the validation process. The concluding section summarizes the findings and offers recommendations.

2. Review of the related literature

The concept of validation was first introduced in the mid-1970s to improve the quality of pharmaceutical products [19]. The concept of validation has expanded over the past few years in a wide range of activities, from analytical methods used to control the quality of materials and drugs to computerized systems, the validation process has become an important and integral part of manufacturing [20]. Validation is a method with applications in various fields of medicine, economics, psychology, chemistry, biology, etc. In fact, the concept of validation can be used in most fields.

Various studies have been conducted to compare various statistical techniques for prediction and credit validation challenges in different domains. According to Abdou (2011), these studies can be categorized as [21]:

• Health and Medicine (Behrman et al., 2007; Nguyen et al., 2002; Warner & Misra, 1996)

• Accounting and Finance (Landajo et al., 2007; Pendharkar, 2005; Baestaens, 1999; Altman et al., 1994; Richardson, 1996; Duliba, 1991; Long, 1973)

• Marketing (Chiang et al., 2006; Teime et al., 2000; Kumar et al., 1995; Dasgupta et al., 1994; Feng and Wang, 2002; Smith and Mason, 1997)

• Public Goods (Nikolopoulos et al., 2007; Usha, 2005; Hardgrave et al., 1994)

Numerous studies have been conducted on validation, specifically within the field of credit scoring, classification, and identification of risk factors related to the health insurance sector. Christian-Alexander Behrendt et al. (2019) examined the foundations of validation for health insurance claims. They introduced two approaches for health insurance credit scoring: a model-based approach and a classification-based approach. For the primary assumption of the first approach, the focus was not on the validity of the data itself but on the features and interpretations of the analyses performed on the data. Therefore, multi-level models with varying complexities using global and local indicators were considered ideal for the hierarchical structure of the data (patients clustered within hospitals). The fundamental principle underlying the second approach was that the results of descriptive or complex methods in health care research typically concentrated on comparable subgroups [9].

Dionne et al. (2012) classified risk using observable and objective characteristics, which insurers use to group insurance applicants with similar expected losses. This classification helps calculate the corresponding insurance premiums, thereby reducing asymmetric information. Risk classification can be employed to reduce adverse selection and improve the insurance market [22].

Mariner (2013) examined the role of insurance in defining the responsibility of healthcare risk and its associated costs. The obtained results indicated that factors, such as diseases, health status, age, and gender, could be risks that influence health insurance premiums and coverage level [23].

Upon reviewing studies in the field of health and healthcare in Iran, it is observed that these studies can be broadly categorized into three levels. The first group includes studies at the macroeconomic level, the second group comprises studies at the microeconomic level, and the third group involves comparative studies.

Given the relevant literature, the current study develops to the existing literature in that based on the characteristics of the studied community, the validation and classification of individuals within the studied population into healthy and unhealthy groups.

3. Theoretical framework

As mentioned, health insurance claims have recently gained concerted attention of the scientific community as a source of real-world evidence towards the improvement and further development of healthcare, and pragmatic experiments [24]. It is worth mentioning that few studies have been conducted in the field of health insurance claims validation. Validation is a statistical method used to determine the risk level of insurance customers, and it must be applied within insurance companies. Validation is the cornerstone of risk management since it is difficult to make

accurate decisions regarding insurance coverage, especially complementary medical insurance without an accurate estimate of an individual's creditworthiness.

As mentioned, one of the most important criteria in validation is determining the risk level of insurance customers. Several definitions have been provided for the term "risk," some of which are used in everyday language, while others are used in a more specialized context like the insurance business. A relatively general definition can be presented in mathematical terms. Risk can be defined as a random number X, the actual outcome (or realization) of which is unknown [25].

The risk variable is denoted by X, which is nonnegative random variable and includes all possible claims that may fall on the insurance company. The variable X is shown as a relation (1), $and X_i$ has a Poisson distribution:

$$X = S_n = \sum_{i=1}^{n} X_i \ , N \sim Poissn(\lambda), \lambda > 0. \tag{1}$$

That is, S is the result of adding X_i one random number N of times, X_i independent from N, which means that the number of claims occurred and the monetary amounts of which one are not related [26].

Lima Ramos (2017) states the expected utility principle for the insurer and specifies that an insurer with utility function U and capital W, should accept a contract against risk X and with a premium of $\pi[x]$ if and only if:

$$E[u(w + \pi[x] - x)] \geq u(w). \tag{2}$$

Definition 1: According to the expected utility, the minimum value of $\pi[x]$ that an insurer with initial capital W must spend to cover risk X is the solution of the following equation:

$$u(w) = E[u(w + \pi[X]^- - x)]. \tag{3}$$

When the utility function is concave, this principle is called zero utility principle. Suppose the insurer has the utility function $u(x) = -\alpha e^{-\alpha x}$ and $\alpha > 0$ and its purpose is to calculate the minimum insurance premium according to the desired utility theory $\pi[x]$ that the insurance company must accept the risk X that $\beta > 0 \, , \theta > 0 \, , X \sim Gama(\theta, \beta)$. First, the generating function of the moment X must be calculated, that is, the function that assigns $a \in R$ to each t, in the interval [– a, a], the value of the function Moment $M_X(t) = E[e^{Xt}]$ as long as the expected value is bounded.

On the other hand, by applying definition 1, the minimum insurance premium will be as follows:

$$\begin{aligned} u(w) = E[u(w + \pi[X]^- - X)] &\Leftrightarrow -\alpha e^{-\alpha w} = E\left[-\alpha e^{-\alpha(w + \pi[X]^- - X}\right] \\ &\Leftrightarrow -\alpha e^{-\alpha w} = E\left[-\alpha e^{-\alpha w} e^{-\alpha \pi[X]} e^{\alpha X}\right] \\ &\Leftrightarrow \boldsymbol{e^{\alpha \pi[X]^- = M_X(\alpha)}} \\ &\Leftrightarrow \pi[X]^- = \frac{1}{\alpha} \ln\left(M_X(\alpha)\right) \end{aligned} \tag{4}$$

Consider the relation 4. The minimum premium value does not depend on the initial capital of the insurer according to the principle of expected utility using the final utility function. Instead, it depends on the parameter α and the risk distribution function X [26]. In the following, we will explain the risk factors.

Ermanno Pitacco (2012) classified risk factors as objective, subjective, observable, and non-observable. Objective risk factors include the physical characteristics of the insured, particularly age, gender, health records, and occupation. Subjective risk factors are the personal attitude towards health, which determines the individual demand for medical treatments and, consequently, the application for insurance benefits. Another relevant classification is observable vs non-observable factors.Observable risk factors are those factors whose impact on claim frequency and claim severity can be assessed during the underwriting phase. Typical examples are age, gender, occupation, etc. Objective risk factors are usually observable factors. Other risk factors are non-observable factors (at least at the time of policy issue). A typical example is given by the personal attitude towards health [27].

In advanced economies, risks are traded as commodities in financial markets. Insurance companies cover risks as an activity in exchange for receiving an insurance premium. As mentioned, insurance companies cover risks as an activity in exchange for receiving insurance premiums. However, risk is a random variable defined in a probability measure space. One of the simplest methods for studying and evaluating a random risk is summarizing it into a number. Therefore, this section discusses important risk measurement tools, examining the analysis of the conditional Poisson distribution function based on auxiliary variables.

Boucher et al. (2014) performed a study on the ranking of losses in insurance using cross-sectional data. For the parametric modeling of the number of losses or claims conditional on auxiliary variables, the actuary must opt for a counting distribution. Generally, the Poisson distribution serves as the initial choice for modeling count data. The probability mass function of the Poisson distribution is as follows:

$$Pr[N_i = n_i|X_i] = \frac{\lambda_i^{n_i} e^{-\lambda_i}}{n_i!} \quad (5)$$

The characteristics of the insured individuals that affect their insurance premium are included as regressors in the parameter of the mean of the counting distribution. These exogenous data can be encoded using binary variables.

In insurance, an exponential function is commonly used as $\lambda_i = t_i \exp(x_i'\beta)$. Where, t_i is the exposure of the insured individual i to the risk. The reason is that $E[N_i|X_i] = \lambda_i$ with the given characteristics of the insured, actuaries can calculate the insurance premium for an insured individual. This information includes the age and gender of the insured, specific diseases, the number of visits to healthcare organizations, and the amount paid by the insurance company. This data structure is designed to ensure the independence of each contract. An example of a cross-sectional database includes claim or loss amounts, age, gender, frequency and number of claims, and health status [28].

In the current study, machine learning algorithms were employed to evaluate the credibility of individual health insurance customers. Therefore, this section provides the theoretical foundations of Data mining in the form of machine learning algorithms.

Koh et al. (2020) define data mining as the process of discovering patterns and previously unknown trends in databases and utilizing this information to construct predictive models.

Cubillas (2023) believed, Data mining in business domains collectively contribute to the field of data-driven applications and predictive modeling across various domains, like a healthcare [29,30].

Data mining provides the methodology and technology to transform this massive data into useful information for decision-making [31]. In this study, XGBoost is used as the most practical machine learning methods.

3.1. XGBoost

Extreme Gradient Boosting (XGBoost) model was first recommended by Tianqi Chen and Carlos Guestrin in 2011 and has been continuously optimized and improved in the follow-up study of many scientists [32]. The XGBoost model is a learning framework based on Boosting Tree models.

Wei Li (2019) discusses, merge the tree model with addition method, assuming a total of K classification and regression trees (CARTs), and use F to delegate the basic tree model, then:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \;\; f_k \in F \quad (6)$$

where the x_i are members of the training data sets and y_i are the corresponding class labels, f_k is the leaf score for the kth tree and F is the set of all K scores for all CARTs. Regularization is applied to improve the final result:

$L = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$ (7)

where L is the loss function, which represents the error between target y_i and the predictive $\hat{y}_i$; Ω is the function used for regularization to prevent overfitting:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (8)$$

where γ, λ are constants controlling the regularization degree, T is the number of leaves in the tree and w is the weight of each leaf [33,34].

Thongsuwan (2021) also discusses, Gradient boosting (GB) is effective in regression and classification problems [34]. Gradient boosting refers to a method in which new models are trained with the aim of predicting the residuals of previous models.

In method XGboost, according to Figure 1, each new model is trained with the aim of correcting the errors caused by previous models. Models are added sequentially until there is no further development possible [35].

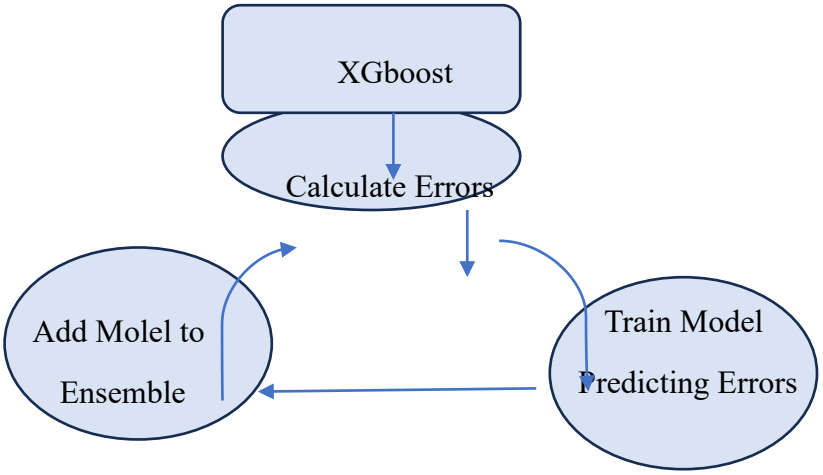

Fig 1. predicting the residuals of previous models [35]

4. Research variables:

As mentioned, in this study, validation was performed based on risk factors in the healthcare domain. As noted by Ellili (2023), big data plays a crucial role in the insurance sector, leading to the adoption of advanced processing technologies like machine learning and artificial intelligence [36]. Therefore, explanatory variables included the amount of insurance claim or payment by the insurance company for medical expenses reimbursement (Iranian Rials), age, gender, number of visits to healthcare organizations, and specific diseases. The statistical population of this research consists of 112,485 data belonging to 22,497 insured employees of the oil company in the northeastern region of Iran. [37].

Considering that the purpose of this study is to validate people into 2 groups of healthy (1) and unhealthy (0) using the XGBoost, a Label must be specified to diagnose healthy or unhealthy. For this purpose, we use the Current health expenditure per capita index. Jaworeck (2022) examines Health Care Index indices in a study and Current health expenditure index is one of the indices introduced and examined [38]. Current health expenditure per capita index for Iran is 104630262 Rials.

5. Modelling

5.1. Validation using XGBoostmethod

This stage involves categorizing health insurance customers as healthy or unhealthy based on a series of characteristics using the XGBoost algorithm and Python software. The implementation of the XGBoost algorithm begins with loading as a set of packages from Python libraries is required. The steps for modeling using the XGBoost algorithm in Python are as follows:

Stage 1: Data Preparation

In the first step, the data, including the target and explanatory variables, are imported from an Excel file.

- *X = data['gender', 'age', 'visits', 'cost', 'catastrophes'] # Features*
- *Y = data['healthy'] # Target variable*

Table 1 lists the explanatory and target variables along with their definition.

Table 1: Main Variables Used in Analysis

	Variable	Definition
Target Variable	healthy	A binary variable equal to 1 if healthy and 0 if unhealthy
Explanatory Variables	gender	A binary variable equal to 1 for males and 0 for females
	age	Age of the individuals.
	visits	Number of visits the person made to doctor/ healthcare centers
	cost	The payout amount by the health insurance company
	specific diseases	A binary variable equal to 1 if the person suffers from a catastrophic illness and 0 otherwise

Notes: The variable *specific diseases* include severe conditions that can lead to major expenses during a person's lifetime, such as end-stage renal failure, major organ transplant, stroke, diabetes, coronary artery disease, vascular disease, and cancer.

Stage 2: Data Training and Testing

The data is split for training and testing using machine learning algorithms, with 80% allocated for training to enable the model to learn and make predictions. The remaining 20% is reserved for testing and evaluation, allowing the comparison of predicted outcomes with actual results to assess the model's accuracy.

Stage 3: XGBoost Algorithm Implementation

The XGBoost algorithm is defined as follows and executed on the data:

- *xgb =XGBRegressor(n_estimators=100, max_depth=3, random_state=42)*
- *xgb.fit(X_train_normalized, y_train)*

Stage 4: Prediction

The model was created in the past four stages, and stage 4 was where the model was used for prediction.

- *y_test_pred = model.predict(X_test)*

The model can now be used to classify and analyze healthy and unhealthy individuals in any new population. To demonstrate its applicability, new data including 148,925 data of 29,785 individuals from the same statistical population for 2023 is entered into the model for analysis. This process involves the following step:

- *y_test_pred = model.predict(X_test)*

Table 2 presents the analysis of health insurance claims payment data based on the validation of healthy and unhealthy individuals using this updated data.

Table 2: Classification based on health insuranceclaims after validation

Classification	ALL	Females	Males
Total Population	29785	15063	14722
Healthy Population	29254	14819	14435
Unhealthy Population	531	244	287
Percentage of healthy Population	98.22%	98.38%	98.5%
Percentage of unhealthy Population	1.78%	1.62%	1.95%
Percentage of the cost of unhealthy Population to the total cost	17.48%	20.31%	15.1%
Percentage of the cost of healthy Population to the total cost	82.52%	79.69%	84.89%

Source: The study's findings.

Table No. 2 shows the results of the validation for the entire population and also separately for the population of women and men. Notably, the validation process identified approximately 1.78% of the population as "unhealthy." This seemingly small group accounts for a disproportionately high 17.48% of the company's total health insurance claims, despite currently being classified and charged premiums as healthy individuals. Validation results for men and women show that although the percentage of unhealthy people in men (1.95%) is more than women (1.62%), but the percentage of women's health insurance claims (20.31%) It is more than men (15.1%).

By applying this classification unhealthy individuals would pay higher premiums, ensuring a fairer and more efficient pricing structure.

5.2. Regression Coefficients

In the following, the coefficients of explanatory variables of the model are estimated using the XGBoost method and the results are according to Table 3. In addition, the instructions related to calculating the coefficients are as follows:

- *for i, importance in enumerate(feature_importances):print (f"Feature {X.columns[i]}: {importance:.4f}")*
  plt.figure(figsize=(8, 6))

Table 3: Coefficients of variables

	Variable	Coefficients
Explanatory Variables	Gender	0.0011
	age	0.0026
	visits	0.0019
	cost	0.9935
	specific diseases	0.0007

Source: The study's findings.

As the coefficients results show, the variable of cost of health insurance claims has the highest coefficient and the greatest impact on the direct unhealthy status of individuals. This result is natural because the higher the cost, the unhealthier it becomes.

The second variable affecting the "health" target variable negatively is age. Naturally, as age increases, the body undergoes changes that can lead to decreased health.

The third variable is Visit, which is directly related to the treatment budget, and the higher the Number of visits to healthcare organizations, the higher the insurance company's payment amount andtheunhealthier the person, and vice versa.

The next variable, which has a negative impact on individuals' health, is "Gender." Table No. 3 shows that the percentage of unhealthy people's per capita cost is higher in women than in men.

The fifth factor influencing the healthy or unhealthy status is specific disease.

5.3. Model evaluation

For issues related to validation and classification, there are usually several standard metrics used. It might be assumed that there are only two cases: either the algorithm has correctly identified the situation, or it has misidentified it. However, the problem is not that simple. Before calculating these values, four main parameters, TP, FP, TN, and FN, needed to be calculated. For a binary classification problem, the Confusion Matrix was a 2×2 matrix that included the mentioned parameters. To clarify, assume the problem is the classification of healthy individuals from unhealthy individuals.For this purpose, Vujovic (2021) introduces standard criteria for classification:

Table 4. Standard metrics for classification

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN) TYPE II ERROR
	Negative	False Positive (FP) TYPE I ERROR	True Negative (TN)

To understand Table 4, imagine that the classification algorithm is subjected to a test set after learning and creating the model [39]. In Table 3, the rows represent the actual labels, and the columns show the algorithm's predictions. Based on this, there are four cases:

True Positive (TP): Some unhealthy individuals are correctly identified as unhealthy.

False Negative (FN): Some unhealthy individuals are mistakenly identified as healthy.

False Positive (FP): Some healthy individuals are mistakenly identified as unhealthy.

True Negative (TN): Healthy individuals are correctly identified as healthy.

Finally, each data sample would fall into one of these two "classes" (Class). Therefore, for each data sample, one of the four scenarios mentioned above may occur.

Evaluation Metrics:

Powers (2011) introduces standard metrics for evaluating the quality of machine learning systems, including Accuracy, Precision, Recall and the F-measure [40].

Accuracy: Indicates the ratio of truly positive cases to all predictions.

Precision: Indicates the ratio of correctly predicted positive cases to the total predicted positive cases.

Recall: Indicates the ratio of truly positive cases correctly predicted as positive. It measures the coverage of truly positive cases.

Since Precision and Recall are inversely related, meaning an increase in one lead to a decrease in the other, a combined metric called the F-measure is introduced. The F-measure is the geometric mean of Precision and Recall.

Udurume (2024) defines the classification measures are all based on four elements: True Positives (TPs), True Negatives (TNs), False Positives (FPs), and False Negatives (FNs) [41]. The representations of the utilized metrics are as Table 5:

Table 5. Metrics of model evaluation

Evaluation Metrics	Formula
Accuracy	$\frac{TP+TN}{TP+FN+TN+FP}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F-measure	$\frac{2(Precision)(Recall)}{Precision+Recall}$

In this section, the confusion matrix should be created. To achieve this, following command was performed: Create a confusion matrix:

- $cm = confusion_{matrix(y_{test}, y_{pred})}$

Finally, the matrix below was created. Figure 2 and Table 6 show the confusion matrix.

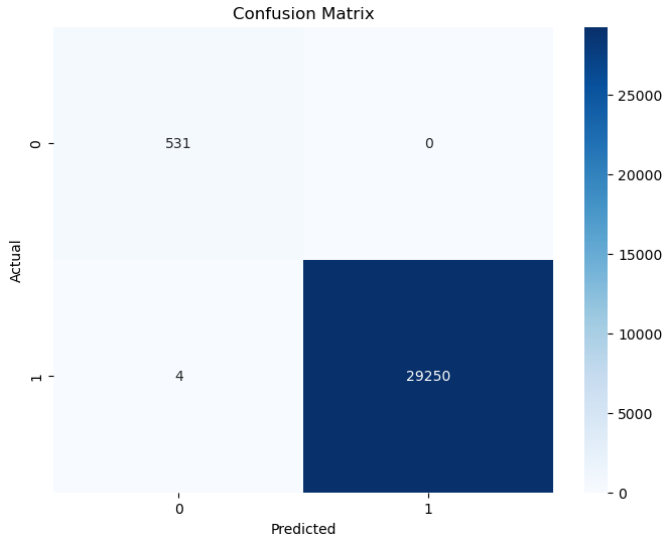

Fig 2. confusion matrix

Table 6. confusion matrix (XGBoost)

TP=531	FN=0
FP=4	TN=29250

Source: The study's findings.

Based on the confusion matrix of methods XGBoost algorithm, the Metrics of model evaluation are shown in Table 7.

Table 7. Metrics of model evaluation

Evaluation Metrics	XGBoost
Accuracy	0.999865
Precision	0.992523
Recall	1.0
F1_score	0.996247

Source: The study's findings.

As can be seen in Table 7, the Accuracy metric was 0.9998, indicating a 99.98% probability of correctly predicting the health status of individuals using this validation. However, the problem with Accuracy is that it fails to differentiate between false negatives and false positives. Despite the very low error rate in this validation, three other metrics, namely precision, recall, and F-measure, were used. The precision metric was 0.9925, signifying a 99.25% probability of accurate identification when the algorithm predicted an individual as unhealthy. The recall metric, focusing on truly healthy data, was at 1, implying a 100% accurate prediction when the algorithm identified an individual as healthy. In certain situations, such as health insurance validation, the recall metric is more important than precision. For instance, if a truly unhealthy person is mistakenly identified as healthy, it incurs a high cost to the insurance company since the unhealthy individual pays the premium for a healthy person. The F1 metric, as the weighted harmonic average of precision and recall, was 0.9962, indicating a 99.62% probability of correct predictions in the performed health insurance validation.

6. Conclusion

In this study, validation was conducted using the XGBoost method for employees of the East of Iran Oil Company. The basis for model validation was risk assessment based on potential damages. The detailed steps of the modeling and evaluation process, along with relevant coding, are presented in stage 4. The most crucial stage is stage 3 and 4, involving the model execution, and ultimately, prediction.

After completing step 3 and performing the training process, the designed model was evaluated using various criteria, including standard criteria such as accuracy, precision, recall, and F-measure, each of which examines specific features of the model. The values of these criteria were 0.999, 0.9925, 1 and 0.996, respectively. These values indicated the very high precision, accuracy and efficiency of the model.

After modeling and verifying the acceptable model, in the 4th step, the model was used to validate the new society, and the results in the studied society indicate the fact that 1.78% of people are unhealthy, so that this small group includes a high cost equal to 17.48% of the total claims.

The two variables that had the most impact on the selection of an unhealthy outcome or validation results indicating unhealthiness, were the cost of health insurance claims variable (The estimated coefficient is equal to 0.9935) and the age variable (The estimated coefficient is equal to 0.0026).

This type of validation model is one of the most practical modeling approaches, which is recommended in the first stage, insurance companies can use it every year for customer validation. In the second step, the health insurance premium should be calculated based on the modeled validation. In the third stage, it is suggested to continue using health insurance validation and insurance premium calculation (based on artificial intelligence validation), to develop flexible health insurance contracts.

References

[1] Defever KM. Insurance Is Accelerating Economic and Social Change in the United States: A Legal and Sociological Perspective. *InCross-Disciplinary Impacts on Insurance Law: ESG Concerns, Financial and Technological Innovation*.2024; 23:29-52. doi:10.1007/978-3-031-38526-1_2

[2] Fotova Čiković K, Cvetkoska V, Mitreva M. Investigating the Efficiency of Insurance Companies in a Developing Country: A Data Envelopment Analysis Perspective. *Economies*. 2024;12(6): 128.doi: 10.3390/economies12060128

[3] Barman A, Barman MS. Health Insurance and Subjective Well-being. 2014.

[4] Shin H, Park H, Lee J, Jhee WC. A scoring model to detect abusive billing patterns in health insurance claims. *Expert Systems with Applications*. 2012;39:7441-50.doi: 10.1016/j.eswa.2012.01.105

[5] Hamid Z, Khalique F, Mahmood S, Daud A, Bukhari A, Alshemaimri B. Healthcare insurance fraud detection using data mining. *BMC Medical Informatics and Decision Making*. 2024;24:112.doi:10.1186/s12911-024-02512-4

[6] Fabrikant R, Kalb PE, Bucy PH, Hopson MD. Health care fraud: Enforcement and compliance. *Law Journal Press*, 2024.
[7] Devaguptam S, Gorti SS, Akshaya TL, Kamath SS. Automated Health Insurance Processing Framework with Intelligent Fraud Detection, Risk Classification and Premium Prediction. *SN Computer Science*. 2024;5:1-4. doi:10.1007/s42979-024-02801-9
[8] Varadarajan V, Kakumanu VK. Evaluation of risk level assessment strategies in life Insurance: A review of the literature. *Journal of Autonomous Intelligence*. 2024;7. doi:10.32629/jai. v7i5.1147
[9] Behrendt CA, Schwaneberg T, Hischke S, Müller T, Petersen T, Marschall U, Debus S, Kriston L. Data privacy compliant validation of health insurance claims data: the IDOMENEO approach. *Das Gesundheitswesen*. 2020;82: S94-100. doi:10.1055/a-0883-5098
[10] Ahir KB, Singh KD, Yadav SP, Patel HS, Poyahari CB. Overview of validation and basic concepts of process validation. *Sch. Acad. J. Pharm*, 2014;3:178.
[11] Glod-Lendvai AM. Validation–a brief introduction. 2018. doi:10.5719/GeoP.3.1/2
[12] Chaitanya Kumar G, Rout RP, Ramtake S, Bhattacharya S. Process Validation. *The Indian Pharmacist*. 2005, 14-19.
[13] Suriyaprakash TN, Ruckmani K, Thirumurugan R. Concepts of Process Validation in Solid Dosage Form [Tablet]–An Overview. *SAJ Pharmacy and Pharmacology*. 2014;1:1.
[14] Lee G, Kim W, Oh H, Youn BD, Kim NH. Review of statistical model calibration and validation—from the perspective of uncertainty structures. *Structural and Multidisciplinary Optimization*. 2019;60:1619-44.doi:10.1007/s00158-019-02270-2
[15] IEEE Computer Society. Software Engineering Standards Committee. IEEE Standard for Software Verification and Validation. *IEEE*; 1998.
[16] Huber L. Validation of computerized analytical systems. *CRC Press*. 2023.
[17] Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. *PloS one*. 2019;14: e0224365.doi: 10.1371/journal.pone.0224365
[18] Polyzotis N, Zinkevich M, Roy S, Breck E, Whang S. Data validation for machine learning. *Proceedings of machine learning and systems*. 2019;1:334-47.
[19] Sarvani V, Elisha RP, Nama S, Pola LM, Rao CB. Process validation: An essential process in pharmaceutical industry. *International Journal of Medicinal Chemistry and Analysis*. 2013;3:49-52.
[20] Kaur H, Singh G, Seth N. Pharmaceutical process validation: a review. *Journal of Drug Delivery and Therapeutics*. 2013;3:189-94.
[21] Abdou HA, Pointon J. Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent systems in accounting, finance and management*. 2011;18:59-88. doi:10.1002/isaf.325
[22] Dionne G, Rothschild C. Risk classification and health insurance. *Encyclopedia of Health Economics*. 2014; 3:272-80. doi:10.2139/ssrn.2134190
[23] Mariner WK. The affordable care act and health promotion: the role of insurance in defining responsibility for health risks and costs. *Duq. L. Rev*. 2012; 50:271.
[24] Choudhry NK. Randomized, controlled trials in health insurance systems. New England Journal of Medicine. 2017 Sep 7;377(10):957-64. doi:10.1056/NEJMra1510058
[25] Olivieri A, Pitacco E. Introduction to insurance mathematics: technical and financial features of risk transfers. Springer; 2015 Sep 30. Doi:10.1007/978-3-319-21377-4
[26] Lima Ramos P. Premium calculation in insurance activity. Journal of Statistics and Management Systems. 2017 Jan 2;20(1):39-65. doi:10.1080/09720510.2016.1187927
[27] Pitacco E. Health insurance. *Basic Actuarial Models, Cham, Switzerland: Springer Verlag*, 2014.doi:10.1007/978-3-319-12235-9
[28] Boucher JP, Inoussa R. A posteriori ratemaking with panel data. *ASTIN Bulletin: The Journal of the IAA*, 2014;44:587-612. doi:10.1017/asb.2014.11
[29] Cubillas JJ, Ramos MI, Feito FR. Use of Data Mining to Predict the Influx of Patients to Primary Healthcare Centres and Construction of an Expert System. *Applied Sciences*. 2022;12:11453. doi:10.3390/app122211453
[30] Amrit C, Abdi A. Methods and Applications of Data Mining in Business Domains. *Applied Sciences*. 2023;13:10774.doi.org/10.3390/app131910774
[31] Koh HC, Tan G. Data mining applications in healthcare. *Journal of healthcare information management*. 2011;19:65.

[32] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. *InProceedings of the 22nd acmsigkdd international conference on knowledge discovery and data mining*. 2016: 785-794. doi.org/10.1145/2939672.293978
[33] Li W, Yin Y, Quan X, Zhang H. Gene expression value prediction based on XGBoost algorithm. *Frontiers in genetics*. 2019;10:1077.doi.org/10.3389/fgene.2019.01077
[34] Ellili N, Nobanee H, Alsaiari L, Shanti H, Hillebrand B, Hassanain N, Elfout L. The applications of big data in the insurance industry: A bibliometric and systematic review of relevant literature. *The Journal of Finance and Data Science*. 2023:100102.doi: 10.1016/j.jfds.2023.100102
[35] Sachan, S. Mastering XGBoost: Your Ultimate Guide to Boosting Machine Learning Performance. https://medium.com, 2024.
[36] Thongsuwan S, Jaiyen S, Padcharoen A, Agarwal P. ConvXGB: A new deep learning model for classification problems based on CNN and XGBoost. *Nuclear Engineering and Technology*. 2021;53:522-31. doi.org/10.1016/j.net.2020.04.008
[37] The East of Iran Oil Company, Statistical report of health insurance data.
[38] Jaworeck S. A New Approach for Constructing a Health Care Index including the Subjective Level. *International Journal of Environmental Research and Public Health*. 2022;19:9686. doi.org/10.3390/ijerph19159686
[39] Vujović Ž. Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*. 2021; 12:599-606.DOI: 10.14569/IJACSA.2021.0120670
[40] Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv*, 2020. Doi:10.48550/arXiv.2010.16061
[41] Udurume M, Shakhov V, Koo I. Comparative Analysis of Deep Convolutional Neural Network—Bidirectional Long Short-Term Memory and Machine Learning Methods in Intrusion Detection Systems. *Applied Sciences*. 2024;14:6967. doi.org/10.3390/app14166967