proceedings

Proceedings

Predicting Health Care Costs Using Evidence Regression [†]

Belisario Panay [1,*], Nelson Baloian [1], José A. Pino [1], Sergio Peñafiel [1], Horacio Sanson [2] and Nicolas Bersano [2]

[1] Department of Computer Science, Universidad de Chile, 8370456 Santiago, Chile; nbaloian@dcc.uchile.cl (N.B.); jpino@dcc.uchile.cl (J.A.P.); spenafie@dcc.uchile.cl (S.P.)

[2] Allm Inc., Tokyo 150 0002, Japan; horacio@allm.net (H.S.); n.bersano@allm.net (N.B.)

[*] Correspondence: bpanay@dcc.uchile.cl

[†] Presented at the 13th International Conference on Ubiquitous Computing and Ambient Intelligence UCAmI 2019, Toledo, Spain, 2–5 December 2019.

Published: 21 November 2019

Abstract: People's health care cost prediction is nowadays a valuable tool to improve accountability in health care. In this work, we study if an interpretable method can reach the performance of black-box methods for the problem of predicting health care costs. We present an interpretable regression method based on the Dempster-Shafer theory, using the Evidence Regression model and a discount function based on the contribution of each dimension. Optimal parameters are learned using gradient descent. The k-nearest neighbors' algorithm was also used to speed up computations. With the transparency of the evidence regression model, it is possible to create a set of rules based on a patient's vicinity. When making a prediction, the model gives a set of rules for such a result. We used Japanese health records from Tsuyama Chuo Hospital to test our method, which includes medical checkups, exam results, and billing information from 2016 to 2017. We compared our model to an Artificial Neural Network and Gradient Boosting method. Our results showed that our transparent model outperforms the Artificial Neural Network and Gradient Boosting with an R^2 of 0.44.

Keywords: health care costs prediction; expert systems; Dempster-Shafer Theory; evidence regression; supervised learning

1. Introduction

Health care expenditure is one of the most critical issues in today's society. World Health Organization (WHO) statistics showed that global health care expenditure was approximately 7.5 trillion US$, equivalent to 10% of the global GDP in 2016 [1]. One of the reasons for these high expenses in care are the low accountability in health care, such as unnecessary procedures or drugs used on patients, or excessive charges for patient treatments.

If we could predict health care costs for each patient with high certainty, problems such as accountability could be solved, enabling control over all parties involved in patients' care. It could also be used for other applications such as risk assessment in the health insurance business, allowing competitive premium charges, or for the application of new policies by governments to improve public health.

With the now-common use of electronic health records (EHR), an interest has emerged in solving accountability problems using data mining techniques [2]. There have been various approaches to predict health care costs for large groups of people [3,4]. On the contrary, prediction for an individual patient has rarely been tackled. Initially, rule-based methods [5] were used for trying to solve these problems requiring domain knowledge as if-then rules. The downside of this method is the requirement of a domain expert to create the rules, thus making the solution expensive and limited to the dataset being used. In the current state of the art, statistical and supervised learning are preferred with supervised methods having better performance. The reason for better performance is the skewed and heavy right-hand tail with a spike at zero of the cost distribution in health care [6].

Supervised learning methods can be evaluated by performance and interpretability; usually, the more sophisticated methods are the ones that have a better performance, sacrificing interpretation (e.g., Random Forest, Artificial Neural Networks, and Gradient Boosting). A drawback of these high performing machine learning algorithms in health care is their black-box nature, especially in critical use cases. Even though health care cost prediction is not a critical use case, using patients' personal and clinical information for this problem could suffer biased results without an interpretable method. Interpretable methods would allow patients, physicians, and insurers to understand the reasoning behind a prediction, giving them the option to accept or reject the knowledge the method is providing. In this work, we present an interpretable regression method applied to the health care cost prediction problem based on the Dempster-Shafer theory, also known as the theory of belief function. We based our work on Petit-Renaud and Denœux evidence regression model [7] using a discount function related to the importance of each dimension. Each dimension importance is learned during the training phase in two different approaches. The first approach uses a variable for each dimension, and the other one uses an Artificial Neural Network (ANN) to obtain the weights of the dimensions. In both approaches, the optimal parameters are learned using gradient descent. Given the transparency of the evidence regression model, we create a set of rules for each patient in the training set based on their vicinity, and when a prediction is made, we give the set of rules with their importance. Our research question is whether it is possible to develop an interpretable method that has a performance similar to black-box methods for the health care cost prediction problem. To test our answer, we used Japanese health records from Tsuyama Chuo Hospital, which include medical checkups, exam results, and billing information from 2013 to 2018, and compare our method performance with less interpretable methods such as Random Forest, ANN, and Gradient boosting (GB). Our results show that our transparent model outperforms the ANN and GB models in the health care cost prediction with an R^2 of 0.44.

2. Related Work

2.1. Health Care Cost Prediction

Statistical methods (e.g., linear regression) suffer from the spike at zero and skewed distribution with a heavy right-hand tail of health care costs [8] in small to medium sample sizes [9]. Advanced methods have been proposed to address this problem, for example, Generalized Linear Models (GLM) where a mean function (between the linear predictor and the mean) and a variance function (between the mean and variance on the original scale) are specified and the parameters are estimated given these structural assumptions [10]. Another example is the two-part and hurdle model, where a Logit or Probit model is used in the first instance to estimate the probability of the cost been zero, and then if it is not, a statistical model is applied, such as Log-linear [11] or GLM. The most complex statistical method used to solve this problem is the Markov Chain model; an approach based on a finite Markov chain suggested estimating resource use over different phases of health care [12]. Mihaylova et al. [8] present a detailed comparison of statistical methods in health care cost prediction.

Supervised learning methods have been vastly used to predict health care costs; the data used for these methods vary. While a few works use only demographic and clinical information (e.g., diagnosis groups, number of admissions and number of laboratory tests) [13], the majority have incorporated cost inputs (e.g., previous total costs, previous medication costs) as well [14–17], obtaining better performance. GB [18] excels as the method with the best performance for this problem [17], which is an ensemble-learning algorithm, where the final model is an ensemble of weak regression tree models, which are built in a forward stage-wise fashion. The most essential attribute of the algorithm is that it combines the models by allowing optimization of an arbitrary loss function, in other words, each regression tree is fitted on the negative gradient of the given loss function, which is set to the least absolute deviation [19]. ANNs come close to the performance of GB, ANNs are an extensive collection of processing units (i.e., neurons), where each unit is connected with many others; ANNs typically consist of multiple layers, and some goal is to solve problems in the same way that the human brain would do it [20]. Another type of model with good results is the M5 Tree [16]; this algorithm is also a Regression Tree, where a Linear Regression Model is used for building the model and calculating the sum of errors as opposed to the mean [21].

In health care, the majority of the expenses of a population are originated from a small group, as Bertsimas et al. [14] showed in their dataset: 80% of the overall cost of the population originates from only 20% of the most expensive members. Therefore, to improve the performance of the methods listed above, a classification phase is suggested to classify patients in a risk bucket. Morid et al. [6] reported that for low-risk buckets, GB obtains the best results, but for higher ones, ANN is recommended.

2.2. Interpretability

In machine learning, interpretability is the ability of a model to explain or present its prediction in an understandable way. The techniques used for interpretability fall into two categories [22]. The first is model transparency, which means to fully comprehend a model, understanding the parts of the model, input values, parameters, and calculation; there is an intuitive explanation, and it can prove that training will converge to a unique solution. The second category is post-hoc explanations, usually applied to black-box models, where a prediction is presented in a comprehensible way with visual or textual artifacts.

In some domains (e.g., medical domain), method interpretability can be as important as its accuracy or even more, given legal or ethical reasons. Interpretability also helps to gain confidence from its end-users; this is why for their ease of interpretation, some problems use simple, transparent models with less accuracy instead of complex, more accurate ones. For example, Cuarana et al. [23] used generalized additive models with pairwise interactions applied to predict pneumonia risk and hospital 30-day readmission, Ustun et al. [24] created a data-driven scoring system called a Super-sparse Linear Integer Model to create a highly tailored scoring system for sleep apnea screening. Naive-Bayes has been used to create a prediction system for heart disease [25]. Regression and decision trees have been applied to a variety of problems [16,26,27]. Some authors have used the Dempster-Shafer theory, also called the theory of belief functions, which is a generalization of the Bayesian theory, like Maseleno et al. [28], who created an expert system for the detection of skin diseases and Peñafiel et al. [29] who associated the risk of getting a stroke with health checkup data.

Today with the heavy adoption of black-box methods (e.g., ANN), there has been a need to interpret these models to improve their performance and make them more reliable. The complex relations between its attributes make them unintelligible, but they usually outperform transparent models. Some of these approaches are focused exclusively on interpreting ANN models [30,31]. Others treat the models as black-box functions developing model-agnostic approaches that produce post-hoc explanations [32]. One of these approaches is to use transparent models to create an approximate representation of the black-box method [33,34]. Another approach is using perturbation-based methods, which consists of making perturbations in individual inputs and observe the variation of the output [35–37]. This approach has similarities with a sensitivity analysis process of a model. Ribeiro et al. [38] used these two approaches and created LIME, an algorithm that can explain the prediction presenting textual or visual artifacts that provide a qualitative understanding of the relationship between the components of the instance and the model prediction of any classifier or regressor by approximating it locally with an interpretable method.

2.3. Dempster Shafer Theory

The Dempster-Shafer Theory (DST) [39] is a generalization of the Bayesian theory that is more expressive than classical Bayesian models since it allows to assign "masses" to multiple outcomes measuring the degree of uncertainty of the process.

Let X be the set of all states of a system called frame of discernment. A mass assignment function m is a function that satisfies:

$$m : 2^X \rightarrow [0,1], \quad m(\phi) = 0, \quad \sum_{A \subseteq X} m(A) = 1 \tag{1}$$

The term $m(A)$ can be interpreted as the probability of getting precisely the outcome A, and not a subset of A.

Multiple evidence sources expressed by their mass assignment functions of the same frame of discernment can be combined using the Dempster Rule (DR) [40]. Given two mass assignment functions m_1 and m_2, a new mass assignment function m_c can be constructed by the combination of the other two using the following formula:

$$\begin{aligned} m_c(A) &= m_1(A) \oplus m_2(A) \\ &= \frac{1}{1-K} \sum_{B \cap C = A \neq \phi} m_1(B) m_2(C) \end{aligned} \tag{2}$$

where K is a constant representing the degree of conflict between m_1 and m_2 and is given by the following expression:

$$K = \sum_{B \cap C = \phi} m_1(B) m_2(C). \tag{3}$$

Petit-Renaud and Denœux were the first ones to introduce a regression analysis based on a fuzzy extension of belief functions [7], called evidence regression (EVREG). Given an input vector **x**, they predict a target variable **y** in the form of a collection of evidence associated with a mass of belief. This evidence can be fuzzy sets, numbers, or intervals, which are obtained from a training set based on a discounting function that takes their distance to the input vector **x** and is pooled using the Dempster combination rule (2). They showed that their methods work better than similar standard regression techniques such as the Nearest Neighbors using data of a simulated impact of a motorcycle with an obstacle.

The EVREG model has been used for predicting the time at which a system or a component will no longer perform its intended function (machinery prognostic) for industrial application. Niu and Yang [41] used the EVREG model to construct time series, whose prediction results are validated using condition monitoring data of a methane compressor to predict the degradation trend. They compared the results of the EVREG model with six statistical indexes, resulting in a better performance of the EVREG model. Baraldi et al. [42] used the model to estimate the remaining useful life of the equipment. Their results have shown the effectiveness of the EVREG method for uncertainty treatment and its superiority over the Kernel Density Estimation and the Mean-Variance Estimation methods in terms of reliability and precision.

3. Data and Problem Description

We wanted to predict the health care cost of a patient in the future year based on their past medical records and health insurance billing information. The health records data was provided by the Tsuyama Chuo Hospital, from 2016 and 2017. These records are obtained from health insurance claims that the hospital must report to the Japanese Government. In these claims, each patient can be identified by a unique id and contains the patient's information of symptoms, treatments, procedures, and billing. The detailed documentation of this claim can be found at http://www.iryohoken.go.jp/shinryohoshu/file/spec/22bt1_1_kiroku.pdf. We used this data to obtain:

- **Demographics**: Patient gender and age.
- **Patients attributes**: General information about patients such as height, weight, body fat, and waist measurement.
- **Health checks**: Results from health check exams a patient had undergone. Japanese workers undergo these exams annually by law. A code indexes each exam, and the result is also included. Some examples are creatinine levels and blood pressure. There are 28 different types of exams, and the date when they were collected is also included.
- **Diagnosis**: Diagnosis for a patient illness registered by date and identified by their ICD-10 codes [43].
- **Billing information**: Each patient had a score registered for each visit or stay in the hospital. This score translates directly to the cost of a patient bill and this is the value we wanted to predict for the next year.

As shown by [6,16], it is challenging to predict patients' health care costs by only using clinical information. The best indicator for future health care costs are previous costs: the additional history of health care expenses is known to improve the prediction. Based on this fact, prediction of future health care costs is better done when patients' data is known for consecutive periods. At least a two years history is needed when trying to predict the costs for one year.

Our dataset had patients' monthly history for 2016 and 2017. However, there are many missing values because most patients had few claims each year. Therefore, we chose to group claims yearly so that we could have fewer missing values. This strategy did not work as expected since many patients had data only for 2016. We then filtered these patients out, and thus, the final set of patients are those who have clinical history for both 2016 and 2017. Table 1 shows the basic statistics of this patients.

Table 1. Statistics of patients' records.

Statistics	Value
Total number of patients	25,464
Mean score for costs	10,649
Mean age	47.09
% Male	48.59
% Female	51.41

The value we needed to predict is the score of each patient in the future year. This score translates directly to the money a patient paid for healthcare. Figure 1 shows the distribution of patients' scores. The chart shows that the score has the same distribution as described in [6], with a spike at 0 and a long right-hand tail as expected for health care cost.

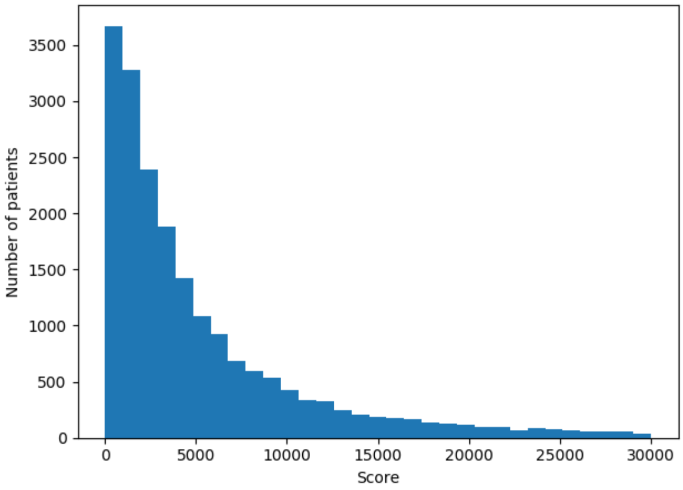

Figure 1. Patients score distribution.

It has been reported [14,16,17] that the use of clinical features yields the same performance as using only cost predictors. Despite clinical information seems not to affect prediction performance, we prefer to keep it, because having it in the model could increase the number of dimensions which may improve vector differentiation. Encoding a patient's history was done by using all sources available as features. The sources are demographics, health checkups result, ICD-10 diagnosis groups, previous score, and actual score. Table 2 shows a detailed description of the patient's vector. As input vector, we used all dimensions shown in Table 2 except for the actual score that was used us our target variable.

Table 2. Patients encoding.

Variable Number	Description
1–2	Demographics
3–30	Health checkup results
31–51	ICD-10 Diagnosis groups
52	Previous score
53	Actual score

4. Proposed Model

4.1. Model Implementation

In this work, we extended the evidence regression model (EVREG) proposed by Petit-Renaud and Denœux [7] to be applied to the prediction of health care cost; we called this method the Interpretable Evidence Regression (IEVREG) model.

To predict a patient p_i health care cost (y), we use a set of other patients as evidence. First, we compute a mass (m_i) of each patient in the evidence set; this mass represents the similarity of the evidence patients with p_i (the one for whom we want to predict the health care cost). Then the target variable y (health care cost) can be calculated as the expected value of the mass m_i and target value y_i of each evidence patient (4).

$$E[y] = \sum_{i=1}^{N} m_i * y_i \tag{4}$$

Formally, we define the training set as:

$$\mathcal{L} : \{p_i = (x_i, y_i)\}_{i=1}^{N} \tag{5}$$

where x_i is the input vector of patient p_i and y_i the actual score (health care cost) for this patient (our target variable). To compute this cost, we need to obtain the target value y_i from the training set $\mathcal{L}$. Each patient in $\mathcal{L}$ is a potential evidence to discover the value y_i, all patients in the training set have a mass (m_i) that represents their similarity with the patient which cost we are trying to predict [44]. The training set $\mathcal{L}$ has also an upper and lower bound for the y variables (max and min-cost), so besides each patient in the training set, the domain of the variable y is also considered another piece of evidence.

The similarity of the patient whom we are predicting the costs, with the ones in $\mathcal{L}$ is measured by a distance function **d**, $\mathbb{R}^n \rightarrow \mathbb{R}$ and a discount function $\phi(d(x_i, x_j)), \mathbb{R} \rightarrow [0, 1]$ which takes the input vectors of the patients. When the distance between the vectors is 0, the discount function is 1, and when the distance is infinite, the discount function is 0.

We defined the distance **d** as :

$$d(x_i, x_j) = \|(x_i - x_j) * w\| \tag{6}$$

where w is a vector of the same dimension as x_i and x_j, representing the weights for each dimension, i.e., the amount which a dimension (age, gender, the result of an exam) contributes to the distance between two patients, thus the weight is the importance of each feature. For this purpose, the values of the input vector should be normalized (e.g., all values must be between 0 and 1).

The discount function we used is defined as:

$$\phi(d) = e^{-\frac{d}{\gamma}^2} \tag{7}$$

where w and γ are values learned during the training phase, in our case we started with $\gamma = 1$ and w as a vector containing only ones. Then the mass of each patient in $\mathcal{L}$ is computed using Dempster rule of combination (2) obtaining:

$$m_j(x_i) = \frac{1}{K}\phi(d(x_i, x_j)) \prod_{h!=j} (1 - \phi(d(x_i, x_h))) \tag{8}$$

where,

$$K = \prod_{j=1}^{N} (1 - \phi(d(x_i, x_j))) + \sum_{j=1}^{N} \phi(d(x_i, x_j)) \prod_{h!=j} (1 - \phi(d(x_i, x_h))) \tag{9}$$

As we said, the domain of the target variable is also a piece of evidence, so the domain mass (m^*) is obtained using (1) and resulting in,

$$m^* = 1 - \sum_{j=1}^{N} m_j(x_i) \tag{10}$$

Finally, to obtain the predicted value of the target variable y_i, we need to transform our belief function into a probability function, satisfying certain axiomatic requirements. Smets et al. [45] showed that the Pignistic transformation could be used for this purpose. With this function, we can get the expected value of the predicted target variable $\widehat{y_i}$ as:

$$\widehat{y_i} = \sum_{j=1}^{N} m_j(x_i) \cdot y_j + \frac{m^* \cdot (\sup_{y \in \mathcal{L}} y + \inf_{y \in \mathcal{L}} y)}{2} \tag{11}$$

with upper and lower expectations:

$$\widehat{y_i}^* = \sum_{j=1}^{N} m_j(x_i) \cdot y_j + m^* \cdot \sup_{y \in \mathcal{L}} y \tag{12}$$

$$\widehat{y_i}_* = \sum_{j=1}^{N} m_j(x_i) \cdot y_j + m^* \cdot \inf_{y \in \mathcal{L}} y \tag{13}$$

4.2. Training Phase

For a prediction to adjust as close as possible to its real value, we need to find the optimal hyperparameters of our model. These parameters are γ in the discount function (7) and the weights of each dimension used in the distance function (6). To obtain these values, we opted for using gradient-based algorithms for the training phase because they have a clear mathematical justification for reaching optimum values.

The gradient descent algorithm is an iterative algorithm that optimizes variable values in order to minimize or maximize a target function. We used the Mean Absolute Error (MAE) as our target function, which is obtained as:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} ||y_i - \widehat{y_i}|| \tag{14}$$

where n is the number of patients, y_i is the true value of the future cost for patient i and $\widehat{y_i}$ is the predicted value i.

Given the loss function MAE that depends on the variable v_t, which can be γ or w, our goal is to minimize the value of the MAE function. The updated value of v_t called v_{t+1} is given by the following formula:

$$v_{t+1} = v_t - \alpha \frac{\partial \text{MAE}}{\partial v} \tag{15}$$

where α is called the learning rate, this algorithm gives us a sequence of values for $v_0, \ldots, v_k$ that minimizes the MAE, the initial value for v (i.e., v_0) is usually selected randomly. To apply gradient descent during the training phase, we try to predict the cost of each patient p_i in the training set $\mathcal{L}$ using all the others $N-1$ patients in $\mathcal{L}$ as evidence. An iteration computing predictions for all N patients is called an epoch, the method of the gradient descent converges by performing multiple epochs.

4.3. Computing Time Optimization

The computation time of a prediction grows linearly with the size of the training set. To compute the mass of a vector of dimension m in a dataset with a training set of size n, first we need the discounting function (7); this has a complexity of $\mathcal{O}(m)$. We can compute every discounting function of the input vector with the training set in $\mathcal{O}(mn)$. Then we can obtain K (9), each product sequence takes $\mathcal{O}(mn)$ and the summation also takes $\mathcal{O}(mn)$, so we can compute K in time $\mathcal{O}(mn)$. Finally to obtain the mass wee need the discounting function, K, and a product sequence, so we compute the masses of the input vector (8), maintaining $\mathcal{O}(mn)$. So then we can get the prediction using (11), with $\mathcal{O}(mn)$ complexity.

Ref. [7] has shown that it is possible to use a K-nearest Neighbors approach to speed up computation without a significant drop in performance. In particular, we used the implementation by Johnson et al. [46] for the exact nearest neighbor's search based on product Quantization. With this optimization, we create indexes for the K-nearest search during the training phase in time $\mathcal{O}(mn + Kn)$. Then a new prediction only needs the masses of the K-Nearest Neighbors, which will be computed in $\mathcal{O}(Km)$; the other masses are assumed to be null. Thus the complexity for a prediction is $\mathcal{O}(Km)$, once the model is trained.

4.4. Interpretability

The IEVREG is a transparent model. We can obtain the contribution (mass) of each piece of evidence in the training set $\mathcal{L}$ for every prediction we make. Thus we fully know how the predicted value is computed. This model can already be considered as an interpretable one, but for the model to be fully explanatory, we will create a set of rules for each prediction with the masses obtained from the training set and the weights of each dimension learned during the training phase. The goal is to estimate the amount each set of evidence contributes to the prediction. First, we create a set of rules for each one of the patients in the training set using the masses of the other $N - 1$ patients in the set and the weights of the dimension. These rules encode the ranges of the dimensions for each of the input features and their masses. Then, for making a prediction, the model finds the most similar patients in the training set and combine their rules to create a new set of rules for that prediction.

To illustrate how we obtain the rules with the IEVREG model we use a small health insurance dataset [47] with only five dimensions as input. The five input data (dimensions) and the predicted value for the care costs are shown in Table 3.

Table 3. Patient

Age	Gender	BMI	Children	Smoker	True Score	Predicted Score
41	Male	32.2	2	No	6836	7555

We predicted the score of this patient using only the 50 nearest neighbors. Then, we obtain the most important rules (higher weights) for the cost prediction; these are shown in Table 4. These rules are the ranges and values that the patient shares with the training set patients.

Table 4. Rules.

Rule	Weight
31.7640 < BMI < 32.9670	0.48
gender = 0.0	0.48
smoker = 0.0	0.48
children = 2.0	0.48
40.3107 < age < 41.3560	0.22

In Table 4 we can observe the interpretation of a patient's cost prediction (the one on Table 3). The IEVREG model assigns low weight to age, and high weight to the other ones. As a consequence, the algorithm tries to find similar patients in terms of BMI, gender, smoker status, and children, and not worry much about age.

5. Experiments and Results

To evaluate the performance of our model, we compare its results with two other methods reported by Morid et al. [6] and Duncan et al. [17] for the heath cost prediction problem; these works used GB and ANN methods respectively.

To measure the performance of each method, we may use the MAE (14), which computes the average absolute difference between the predicted cost and the real one.

However, the MAE is not useful to compare results with costs expressed in different currencies, so we will also use the Mean Absolute Percentage Error (MAPE) a modified absolute error where the MAE is divided by the mean cost and is computed as:

$$\text{MAPE} = \frac{\frac{1}{n}\sum_{i=1}^{n}||y_i - \widehat{y_i}||}{\bar{m}} \tag{16}$$

where $\widehat{y_i}$ is the predicted value for variable y_i and $\bar{m}$ is the mean of variable y defined as:

$$\bar{m} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{17}$$

We will also use another measure, the R^2 that is the Pearson correlation between the predicted and actual health care cost, and represents how close we are to the real cost curve. This value is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{m})^2} \tag{18}$$

For the GB model, we used the parameters used by Duncan et al. [17] with 5000 boosting stages, a maximum depth of individual regression estimator of 6 and a learning rate of 0.01. For the ANN we tried multiples configurations (number of layers, number of neurons in each layer and activation function) the best configuration had two hidden layers, the first with 52 neurons (same as the input dimensions) and the other with 30, the learning rate will be of 0.01, and with a Rectified Linear Unit ($f(x) = max(0, x)$) as an activation function. We let it iterate to a maximum of a 1000 epochs. For our model, we trained the fixed weights for each dimension using gradient descent for a maximum of 100 epochs, and for optimization, we used the mass functions of the closest 150 neighbors.

To test the models, we divided the data in 70% training and 30% for evaluation. The result obtained for each model is shown in Table 5. We can see that our IEVREG model outperformed the other models in every performance measure. The GB method obtains a better performance than the ANN, as reported by Morid et al. [6].

Table 5. Models performance (for MAE and MAPE lower is better, for R2 higher is better).

Model	MAE	MAPE	R^2
IEVREG	7638	0.77	0.44
GB	7966	0.80	0.40
ANN	8023	0.81	0.35

6. Discussion and Conclusions

We presented a new regression method that has the ability to easily show the reasons for making a particular prediction about possible health care costs, which is a desirable ability in the health domain. In order to test its predicting performance, we compared its results with the predictions made by two best models from the eleven analyzed and reported by Morid et al. [6]. This is the GB and the ANN. Comparing the results for the three models for the health care cost prediction problem we can conclude that our method obtains better performance, proving that it is possible to create a more transparent model for a problem like health care cost prediction, going against the common belief that sophisticated and black-box like methods are always the solution with the best performance for every problem that is presented.

We improved the Evidence regression model presented by Petit-Renaud and Denœux [7] to be used in the prediction of health care costs. Our results obtained using data of electronic health records from Tsuyama Chuo Hospital showed that our Evidence regression model with an $R^2 = 0.44$, a transparent and interpretable method, could outperformed the current state of the art supervised learning algorithms such as GB ($R^2 = 0.40$) and ANN ($R^2 = 0.35$).

Even though results are similar or better than other previous works, we believe our results are still improvable. One of the approaches to improve performance was classifying patients in cost buckets as recommended by various studies [6,14,17], this resulted in better performances but escaped the goal of this work, so for future work we can apply this classification process to obtain a patient risk class as first step to improve the performance of our IEVREG model, and to continue comparing our model to even more sophisticated methods we could try to solve the prediction of health care cost using deep learning methods but for this to be feasible we need a larger dataset. We also plan to apply this model to other regression problems in the health care domain, for example, predicting the hospital length of stay and predicting the days of readmission based on each patient's diagnosis and history, which are two classic prediction problems for this domain.

References

1. Organization, W.H. *Public Spending on Health: A Closer Look at Global Trends*; Technical report; World Health Organization: Geneva, Switzerland, 2018.
2. Yoo, I.; Alafaireet, P.; Marinov, M.; Pena-Hernandez, K.; Gopidi, R.; Chang, J.F.; Hua, L. Data mining in healthcare and biomedicine: A survey of the literature. *J. Med. Syst.* **2012**, *36*, 2431–2448.
3. Bilger, M.; Manning, W.G. Measuring overfitting in nonlinear models: A new method and an application to health expenditures. *Health Econ.* **2015**, *24*, 75–85.
4. Diehr, P.; Yanez, D.; Ash, A.; Hornbrook, M.; Lin, D. Methods for analyzing health care utilization and costs. *Ann. Rev. Public Health* **1999**, *20*, 125–144.
5. Kronick, R.; Gilmer, T.; Dreyfus, T.; Ganiats, T. *CDPS-Medicare: The Chronic Illness and Disability Payment System Modified to Predict Expenditures for Medicare Beneficiaries*; Final Report to CMS; 2002. Available online: http://www.hpm.umn.edu/ambul_db/db/pdflibrary/dbfile_91049.pdf (accessed on 24 June 2002)
6. Morid, M.A.; Kawamoto, K.; Ault, T.; Dorius, J.; Abdelrahman, S. Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation. *AMIA Ann. Symp. Proc.* **2017**, *2017*, 1312.
7. Petit-Renaud, S.; Denœux, T. Nonparametric regression analysis of uncertain and imprecise data using belief functions. *Int. J. Approx. Reason.* **2004**, *35*, 1–28.
8. Jones, A.M. *Models for Health Care*; University of York, Centre for Health Economics: York, UK, 2009.
9. Mihaylova, B.; Briggs, A.; O'Hagan, A.; Thompson, S.G. Review of statistical methods for analysing healthcare resources and costs. *Health Econ.* **2011**, *20*, 897–916.
10. Blough, D.K.; Madden, C.W.; Hornbrook, M.C. Modeling risk using generalized linear models. *J. Health Econ.* **1999**, *18*, 153–171.
11. Leung, S.F.; Yu, S. On the choice between sample selection and two-part models. *J. Econometr.* **1996**, *72*, 197–229.
12. Marshall, A.H.; Shaw, B.; McClean, S.I. Estimating the costs for a group of geriatric patients using the Coxian phase-type distribution. *Stat. Med.* **2007**, *26*, 2716–2729.
13. Lee, S.M.; Kang, J.O.; Suh, Y.M. Comparison of hospital charge prediction models for colorectal cancer patients: neural network vs. decision tree models. *J. Korean Med. Sci.* **2004**, *19*, 677–681.
14. Bertsimas, D.; Bjarnadóttir, M.V.; Kane, M.A.; Kryder, J.C.; Pandey, R.; Vempala, S.; Wang, G. Algorithmic prediction of health-care costs. *Oper. Res.* **2008**, *56*, 1382–1392.
15. Frees, E.W.; Jin, X.; Lin, X. Actuarial applications of multivariate two-part regression models. *Ann. Actuar. Sci.* **2013**, *7*, 258–287.
16. Sushmita, S.; Newman, S.; Marquardt, J.; Ram, P.; Prasad, V.; Cock, M.D.; Teredesai, A. Population cost prediction on public healthcare datasets. In Proceedings of the 5th International Conference on Digital Health 2015, Florence, Italy, 18–20 May 2015; ACM: New York, NY, USA, 2015; pp. 87–94.
17. Duncan, I.; Loginov, M.; Ludkovski, M. Testing alternative regression frameworks for predictive modeling of health care costs. *N. Am. Actuar. J.* **2016**, *20*, 65–87.
18. Elith, J.; Leathwick, J.R.; Hastie, T. A working guide to boosted regression trees. *J. Anim. Ecol.* **2008**, *77*, 802–813.
19. Sutton, C.D. Classification and regression trees, bagging, and boosting. *Handb. Stat.* **2005**, *24*, 303–329.
20. Zurada, J.M. *Introduction to Artificial Neural Systems*; West Publishing Company: St. Paul, MN, USA, 1992; Volume 8.
21. Breiman, L. *Classification and Regression Trees*; Routledge: London, UK, 2017.
22. Lipton, Z.C. The mythos of model interpretability. *arXiv* **2016**, arXiv:1606.03490.

23. Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–15 August 2015; ACM: New York, NY, USA, 2015; pp. 1721–1730.
24. Ustun, B.; Rudin, C. Supersparse linear integer models for optimized medical scoring systems. *Mach. Learn.* **2016**, *102*, 349–391.
25. Pattekari, S.A.; Parveen, A. Prediction system for heart disease using Naïve Bayes. *Int. J. Adv. Comput. Math. Sci.* **2012**, *3*, 290–294.
26. Montbriand, M.J. Decision tree model describing alternate health care choices made by oncology patients. *Cancer Nurs.* **1995**, *18*, 104–117.
27. Fonarow, G.C.; Adams, K.F.; Abraham, W.T.; Yancy, C.W.; Boscardin, W.J.; ADHERE Scientific Advisory Committee, Study Group, and Investigators. Risk stratification for in-hospital mortality in acutely decompensated heart failure: Classification and regression tree analysis. *Jama* **2005**, *293*, 572–580.
28. Maseleno, A.; Hasan, M.M. Skin diseases expert system using Dempster-Shafer theory. *Int. J. Int. Syst. Appl.* **2012**, *4*, 38–44.
29. Peñafiel, S.; Baloian, N.; Pino, J.A.; Quinteros, J.; Riquelme, Á.; Sanson, H.; Teoh, D. Associating risks of getting strokes with data from health checkup records using Dempster-Shafer Theory. In Proceedings of the 2018 IEEE 20th International Conference on Advanced Communication Technology (ICACT), Gang'weondo, Korea, 11–14 February 2018; pp. 239–246.
30. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning important features through propagating activation differences. *arXiv* **2017**, arXiv:1704.02685.
31. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin, Germany, 2014; pp. 818–833.
32. Ribeiro, M.T.; Singh, S.; Guestrin, C. Model-agnostic interpretability of machine learning. *arXiv* **2016**, arXiv:1606.05386.
33. Craven, M.; Shavlik, J.W. Extracting tree-structured representations of trained networks. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 27 November–2 December 1996; pp. 24–30.
34. Baehrens, D.; Schroeter, T.; Harmeling, S.; Kawanabe, M.; Hansen, K.; MÃžller, K.R. How to explain individual classification decisions. *J. Mach. Learn. Res.* **2010**, *11*, 1803–1831.
35. Kononenko, I. An efficient explanation of individual classifications using game theory. *J. Mach. Learn. Res.* **2010**, *11*, 1–18.
36. Krause, J.; Perer, A.; Ng, K. Interacting with predictions: Visual inspection of black-box machine learning models. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, 7–12 May 2016; ACM: New York, NY, USA, 2016; pp. 5686–5697.
37. Zhou, J.; Troyanskaya, O.G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* **2015**, *12*, 931.
38. 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 Mmining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 1135–1144.
39. Shafer, G. *A Mathematical Theory of Evidence*; Princeton University Press: Princeton, NJ, USA, 1976; Volume 1.
40. Shafer, G. Dempster's rule of combination. *Int. J. Approx. Reason.* **2016**, *79*, 26–40.
41. Niu, G.; Yang, B.S. Dempster–Shafer regression for multi-step-ahead time-series prediction towards data-driven machinery prognosis. *Mechan. Syst. Signal Process.* **2009**, *23*, 740–751.
42. Baraldi, P.; Di Maio, F.; Al-Dahidi, S.; Zio, E.; Mangili, F. Prediction of industrial equipment remaining useful life by fuzzy similarity and belief function theory. *Expert Syst. Appl.* **2017**, *83*, 226–241.
43. World Health Organization. *International Classification of Functioning, Disability and Hhealth: ICF*; World Health Organization: Geneva, Switzerland, 2001.
44. Denoeux, T. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. Syst. Man Cybern.* **1995**, *25*, 804–813.
45. Smets, P. What is Dempster-Shafer's model. In *Advances in the Dempster-Shafer Theory of Evidence*; John Wiley & Sons, Inc.: New York, NY, USA, 1994; pp. 5–34.

46. Johnson, J.; Douze, M.; Jégou, H. Billion-scale similarity search with GPUs. *IEEE Trans. Big Data* **2019**, doi:10.1109/TBDATA.2019.2921572.
47. Lantz, B. *Machine Learning with R*; Packt Publishing Ltd.: Birmingham, UK, 2013.