Computing (2024) 106:295–325
https://doi.org/10.1007/s00607-023-01224-3

REGULAR PAPER

Generalized hurdle count data models based on interpretable machine learning with an application to health care demand

Xin Xu[1] · Tao Ye[2] · Jieying Gao[1] · Dongxiao Chu[1]

Received: 18 July 2022 / Accepted: 4 September 2023 / Published online: 18 September 2023

Abstract
The zero-inflated count data model has long been viewed as an important research topic owing to its enormously different disciplines. As early classical statistical models of linear and logarithmic mean transformation are difficult to be consistent with reality, an enhanced hurdle model based on machine learning methods is proposed. The decision tree, random forest, support vector, and XGBoost methods are introduced in the two stages of the hurdle model. This framework allows to capture the decision-making behavior and predict the count more flexibly and accurately. The generalized hurdle model consists of traditional discrete distributions, which can fit under-dispersed, equi-dispersed, or over-dispersed count data. The extended hurdle models are utilized to fit health care data and compare their performance with traditional count models. The results show that the generalized hurdle model with random forest performs best. Variable importance, break-down plots, and partial plots provide better interpretability for the extended model, which makes the results more reliable and transparent. To the best of our knowledge, this is the first study to generalize the hurdle model with interpretable machine learning methods in count data.

Keywords Zero-inflated · Two-stage · Generalized hurdle · Interpretable machine learning

Mathematics Subject Classification 62P10

1 Introduction

In recent years, the birth rate is declining globally, while the population aging problem is becoming more and more serious. At the same time, in the wake of the aging of population and the enhancement of healthcare awareness, especially after the

Tao Ye, Jieying Gao and Dongxiao Chu have contributed equally to this work.

Extended author information available on the last page of the article

outbreak of COVID-19, people require more and more healthcare-related safeguard and pay more attention to their health problems. After the outbreak of COVID-19, healthcare expenditure has seen a sharp increase. For example, in 2020, the healthcare expenditure in the US reached 4.1 trillion yuan, an increase of 9.7% [1]. Excessive healthcare demand will not only put financial pressure on governments, but also continuously raise the cost of healthcare. Studies show that health spending, which accounts for more than 10 percent of global GDP, is growing rapidly [2]. With this in mind, accurate prediction of healthcare demand can rationalize the budget arrangement of the government and alleviate the financial burden to some extent.

The existing literature typically estimates and predicts health needs based on the number of visits or length of stay [3]. As research on health economics has intensified, count data modelling has been widely used in different fields in recent years. Overall, these studies are divided into two groups in terms of the model application: (1) single-stage parametric count models; (2) two-part semi-parametric count models. Among the single-stage count models, discrete distributions such as the Poisson and negative binomial distribution were usually adopted [4]. However, patients often decide whether to receive medical services according to their real health and financial status, which usually leads to abundant values of zero in health care demand data. When the proportion of zeros is greater than a certain value that standard count distribution could predict, the traditional count models, such as the Poisson and negative binomial models may underestimate the standard errors and overestimate the significance level of parameters. Then, the various two-stage models, such as zero-inflated models and hurdle models, have been developed to address the problem of extra zeros in recent years [5–8]. Those predominant models are mainly parametrically estimated by the method of maximum likelihood. Although the models perform well, there are two types of assumptions that need to be met [9]. Firstly, the response variable needs to satisfy the special formal distribution hypothesis. Secondly, the relationship between the response variable and the independent variables is assumed to be log-linear. However, the log-linear assumption might be violated because of complex infrastructure data [10]. The actual relationship may be nonlinear. For example, in studying health care utilization, the gender, income, or the occupation of patients might cause the heterogeneity, and the parametric models incorporating the log-linearity assumption $E[Y \mid X] = exp(X^{'}\beta)$ neither reflect the change signs of regression coefficients nor reveal heterogeneous patterns in the data [11]. Moreover, for the first part of zero-inflated and hurdle models, Poisson and the logit specifications have been used in previous studies. The excessive flooding of zeros of response variables could result in an inaccurate prediction of zero probability by using the standard zero-inflated and hurdle parametric models. Both the hurdle and zero-inflated models are proposed to explain excess zeros in the data. The hurdle model has two parts: the first part is a logistic regression to model the probability that a count is zero or a positive integer value, and the second part is a truncated-at-zero distribution to model the number of counts greater than zero (i.e., positive integer numbers) [12]. The zero-inflated model is considered a mixture of a logistic regression and either a Poisson or a negative binomial model [13]. The logistic model is used to separate zeros from positive counts and the Poisson or negative binomial model is used to model the positive counts. Therefore, the hurdle

model is distinct from zero-inflated model in the way of interpreting and analyzing zero counts. The hurdle model assume that all zeros are from one "sampling" source, while the zero-inflated model assume that zero observations have two different origins, namely "structural" and "sampling". That is zero-inflated count models are two-part mixtures consisting of a degenerate distribution at zero and an untruncated count distribution. The hurdle model is a two-part mixture model consisting of a point mass at zero followed by a zero-truncated count distribution for the positive observations. We can consider the two parts of hurdle model independently.

Although the two-stage model developed in the literature can handle the problem of zero inflation better than the single-stage model, it is difficult for the two-stage model to divide individuals' decision-making process into two stages because of the constraints of the parameter hypothesis and the logit assumption when the data face the problem of zero inflation and imbalance. The logistic regression model does not always accurately predict zero probability and explain decision-making behavior in the first stage. As effective and flexible methods, nonparametric models such as classification and regression trees (CART) [10], random forest (RF) [14], support vector machine (SVM) [15] and eXtreme Gradient Boosting (XGBoost) [16], can be considered and get rid of the constraints of function form and describe the decision-making behavior of individuals. To improve the prediction accuracy and flexibility of the two-stage model, this paper further extends the hurdle model based on machine learning methods.

Our study contributes primarily to three aspects. First, we propose a two-stage framework for count models with zero-inflated data based on machine learning. This distinguishes our work from those in the literature with only one-stage models and parametric zero-inflated or hurdle models.

Second, the proposed generalized hurdle model does not require knowledge of unobservable distribution, and the prediction of zero values uses machine learning classifiers rather than logistic regression, while the count values are estimated using machine learning regression. The extended hurdle model consists of traditional discrete distributions such as Poisson, negative binomial, zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), zero-truncated Poisson (ZTP), zero-truncated negative binomial (ZTNB) and so on, which can fit under-dispersed, equi-dispersed or over-dispersed count data. Third, interpretable machine learning methods are proposed to explore the differences between the factors of the decision-making behavior in the first stage and the final results in the second stage, which can overcome the weakness of the black box. At the same time, the reliability of the results is enhanced while generalizing the hurdle model. To our knowledge, there is no current work using these interpretable machine learning methods in the count data model. The remainder of this paper is organized as follows. Section 2 reviews existing literature. Section 3 introduces the methodologies for modeling count data and outlines the generalized hurdle models. Section 4 describes the health care datasets used. Section 5 presents the results of the proposed models and compares them with state-of-the-art methods used in count regression models. Meanwhile, we further explain the results using interpretable machine learning methods and outline the limitations of the proposed models. The final section draws conclusions and future research directions.

2 Literature review

Prior studies of single-stage count model and two-stage count model are reviewed here.

2.1 Single-stage count model

The benchmark single-stage regression model for count data is the standard Poisson distribution. Early work adopting the Poisson regression took place in actuarial science [17], economics [18], and health care [19]. A remarkable feature of the Poisson distribution is that the mean and variance are equal, but the actual count data does not satisfy the equi-dispersion property, which show either over-dispersion (variance greater than the mean) or under-dispersion (variance less than the mean). Therefore, a large number of count data models have been developed to settle the over-dispersion. It turns out that the negative binomial distribution as a Poisson–Gamma mixture model is appreciated and can be used when the data are over-dispersed. The problem of over-dispersion is more common in reality, especially in the field of actuarial [20] and health care [19]. In fact, there is no reason to restrict the over-dispersed shape to the gamma distribution. A Poisson-inverse Gaussian model can be constructed, which is also an ideal candidate. The Poisson-inverse Gaussian model is more suitable than the negative binomial model when the overdispersion of count data is due to highly right-skewness [21]. Poisson-lognormal models are adopted when the shape of overdispersion is lognormal and often used in biostatistical studies [22]. In fact, any distribution with support in the half-positive real line might be a candidate to simulate the shape of overdispersion under mixed Poisson distributions. However, the more complex the mixed model is, the more difficult it is to use. Because the Poisson-inverse Gaussian and Poisson-lognormal models have no closed forms, the two models are rare in the field of health care demand studies. Another method to deal with over-dispersed problems is the generalized Poisson distribution. There were several distributions proposed for modeling over-dispersion, such as the generalized Poisson [23], the double Poisson [24], the Conway–Maxwell–Poisson distribution [25], and so on. Nevertheless, some models are not able to capture the dispersion of the count data.

2.2 Two-part count model

In some circumstances, most of the count data hold an excessive number of zeros, which exceed the standards of the Poisson, negative binomial, or other standard discrete distribution. The presence of those extra zeros could result in over-dispersion. To overcome the limitation of these standard count models, a lot of models using some discrete distributions with zero-inflated structures have been proposed. The zero value of zero-inflated count data can be regarded as the sampling zeros from the standard count distribution, and the other part comes from the additional structural

zeros. Therefore, there are conditional approaches and unconditional approaches of zero-inflated modes widely adopted by researchers based on the sources of zero value.

In terms of the unconditional method, Lambert [13] first introduced the zero-inflated Poisson model to detect manufacturing defects that have large counts and zeros. In practice, over-dispersion and a heavy tail in positive count data still exist even after structural zeros are split. Greene [18] replaced a negative binomial distribution instead of the Poisson distribution in the ZIP models and made an extension of Lamber's work. Yip and Yau [26] adopted the approach to model the claim counts in general insurance, and made a comparative study of diverse zero-inflated models. Neelon et al. [27] used zero-inflated models to examine emergency department visits. Preisser et al. [28] and Liu et al. [29] introduced marginalized two–part models for count data to analyze healthcare utilization in health economics and health services research. Due to the excess zeros coming from heterogeneous sources, specific models using subgroup analysis of zero-inflated Poisson regression model are proposed by Chen et al. [30] with insurance application.

As a well-known conditional method for excess zeros count data, hurdle model, which was first discussed by Mullahy [12], is more appropriate when the zero response is independent of the under statistical process. An extended hurdle model with a generalized logistic in the first stage introduced by Gurmu [31] was used to deal with the problem of overdispersion and underdispersion in health care data. Ehsan Saffari et al. [32] presented a negative binomial hurdle model that performs better than single-stage model. Baetschmann et al. [33] explored a dynamic hurdle as a new approach for zero-inflated count data. Xu et al. [34] made a comparison of hurdle and zero-inflated models to depict hospitalization decisions and utilization for the elders.

2.3 Machine learning for count data

Some work using machine learning methods in count data. Sakthivel and Rajitha [35] used artificial neutral networks to forecast the claim counts in general insurance. Gao et al. [36] proposed boosting Poisson regression models to predict the claim frequency based on the behavior of driver data. Liu et al. [37] considered each count as a type rather than a number and use the support vector machine method to solve the problem of unbalanced data. Lee [38] presented a zero-inflated Poisson regression with boosting algorithm to handle a class imbalance in count data. Kong et al. [39] suggested a deep hurdle network for zero-inflated multi-target regression to estimate multiple species abundance. Zhang et al. [40] developed a multivariate zero-inflated hurdle model to analyze count data.

Despite a lot of literature on count data modeling with extra zeros, there are still mainly two limitations. One limitation is that the approaches based on statistical models are restricted to a linear form, which may be too rigid to be consistent with reality. Another is that the previous study using machine learning algorithms either established a single-stage model that cannot describe the decision-making behavior, or built a complex machine learning model that was difficult to explain the results. In this study, we

try to generalize the hurdle models with interpretable machine learning methods aiming to solve these problems.

3 Modeling framework

In this section, we give a brief introduction to the hurdle model, and then the focus is shifted to the proposed generalized hurdle based on the machine learning modeling framework.

3.1 Classical hurdle model

The hurdle model as a type of two-part stage model specifies the two processes of generating the zeros and nonzeros (positive) are not constrained to be the same. This leads to the hurdle model

$$\Pr[Y = j \mid X] = \begin{cases} f_1(0) & \text{if} \quad j = 0 \\ \frac{1-f_1(0)}{1-f_2(0)} \times f_2(j) & \text{if} \quad j > 0 \end{cases} \tag{1}$$

where density $f_2(\cdot)$ is usually a standard count density such as Poisson, negative binomial, generalized Poisson, and so on, whereas $f_1(\cdot)$ may also be a count data density, or a more simple probability. When $f_1(\cdot) = f_2(\cdot)$, Eq. (1) is equivalent to a single-stage count model. The basic idea of the hurdle model is that using a Bernoulli distribution governs the binary outcome of whether a response variable is zero or positive, and then a zero-truncated distribution fits the positive response variable. The expected positive count of response is

$$E[Y \mid X] = P_1[Y > 0 \mid X] \times E_2[Y \mid Y > 0, X] \tag{2}$$

3.2 Proposed modeling framework

The logistic regression with the binary outcome has been widely applied in the first stage as a benchmark model. In this paper, we provide a generalization of the hurdle model based on machine learning techniques that are used for classification in the first stage and regression in the second stage. The response in the first stage is usually limited to being considered as a binary flag in the classical hurdle model. However, we can extend this by making it a multi-class categorical variable, which may deal with the multiple making-behavior outcomes. Then the proposed modeling frameworks are following:

$$\Pr[Y \mid X] = \sum_{j=1}^{m} \Pr[Z = j \mid X] \cdot \Pr[Y \mid Z = j, X] \tag{3}$$

where $Y$is the count response, X denotes the vector of features, Z is a multi-class latent variable and $Z = j$ means that the outcome falls into class j. The Eq. (3)

indicates that the conditional density function of generalized hurdle model is given by

$$P[Y \mid X] = \varphi(X) \cdot P[Y \mid Z = 1, X] + [1 - \varphi(X)] \cdot P[Y \mid Z = 0, X] \tag{4}$$

where $Y$is the count response, Xdenotes the vector of features, Z is a latent with 0 or 1, and $\varphi(X) = Pr[Z = 1 \mid X]$, which implies that

$$E[Y \mid X] = \sum_{j=1}^{m} \Pr[Z = j \mid X] \cdot E\left[Y_j \mid Z = j, X\right] \tag{5}$$

Accordingly, if we define Z as a binary outcome, then the count number conditional on covariates can be computed as following:

$$E[Y \mid X] = \varphi(X) \cdot E[Y \mid Z = 1, X] + [1 - \varphi(X)] \cdot E[Y \mid Z = 0, X] \tag{6}$$

$E[Y \mid Z = 1, X]$and $E[Y \mid Z = 0, X]$ can be estimated from training data. As explained above, we can make use of a binary classifier to predict $\varphi(X)$. The algorithm of the generalized hurdle modelling framework is shown as:

Algorithm of the generalized hurdle

Step 1: Determine the training data set $Data_{train} = \left\{Z_j, Class_j\right\}, Class_j \in \{0, 1, \cdots, m\}$, where is the number of classification;

Step 2: Build a model $\hat{\varphi}_j$to estimate $\hat{\varphi}_j(X) = Pr\left[Z_j \in Class_j \mid X\right]$;

Step 3: Integrate the results to obtain a final model for prediction by

$E\left[\hat{Y} \mid X\right] = \sum_{j=1}^{m} \hat{\varphi}_j(X) \cdot E\left[\hat{Y}_j \mid Z \in Class_j, X\right]$;

Step 4: For a new instance with given X_{new}, the prediction is

$E\left[\hat{Y}_{new} \mid X_{new}\right] = \sum_{j=1}^{m} \hat{\varphi}_j(X_{new}) \cdot E\left[\hat{Y}_j^{new} \mid Z = j, X_{new}\right]$

The proposed conceptual framework is presented in Fig. 1. First, we divide the data into training and testing sets using K-fold cross-validation and then select the

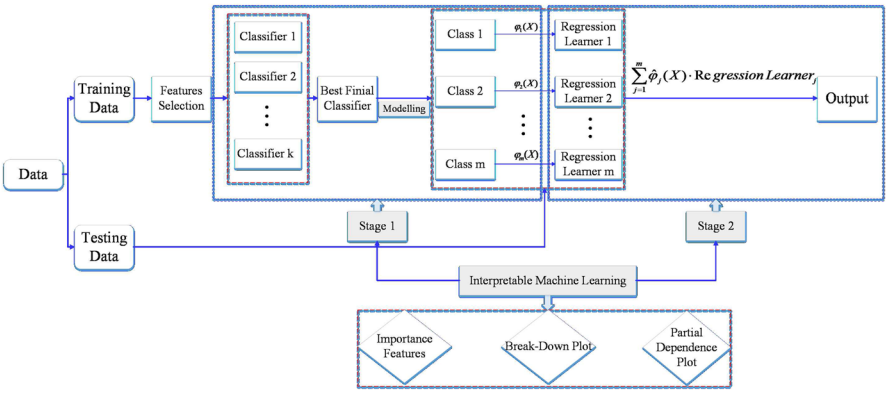

Fig. 1 Conceptual framework for generalized hurdle modelling

features. To classify the response in the first stage, we employ many classifiers and select the best one to classify the training data. Several regression learners are utilized for different response classes in the second stage. To further improve the performance of the proposed model in both stages, a feature selection strategy is introduced.

When Z is a binary outcome, if the logistic regression algorithm is chosen in the first stage, the count regression model is utilized in the second stage, and the extended hurdle model will become the traditional hurdle model.

4 Empirical design

For the sake of comparison, we design and implement the experiments on real data from the Deb and Trivedi [41] to show the effectiveness of our proposed approaches, which discuss how to model count data on medical care utilization by the elderly 66 years and over in the United States. We first describe the data used for the experimental study, and then the utilized classification and regression machine learning models are introduced briefly. Next, the experimental setup is given with the model performance measures.

4.1 Dataset description

The data set is a survey sample consisting of 4406 cases obtained from the National Medical Expenditure Survey conducted in 1987 and 1988, which is applied in the amount of literature to analyze how Americans use and pay for health care services. The data set is characterized by a lot of variables regarding measures for health service demand and variables of health-related and socio-economic determinants. In line with the previous literature, we select variables the same as the study of Deb and Trivedi (Table 1). Here, we focus on how to model the demand for health care services for the elders, and explore the important factors affecting the demand of the elders' behavior.

The number of office visits to physicians (OFP) is usually considered as one measure of individual health service utilization. Figure 2 shows many zeros, and the distribution of visit counts exhibits right-skewness and a heavy tail, which may indicate that the standard Poisson or Negative Binomial model could be unsuitable.

Some factors, such as gender, insurance plans, and limitations in activities, will affect the decision-making process of whether the elders visit an office physician. As can be seen from Fig. 3, the elders possessing unlimited insurance provided by Medicaid are more likely to choose outpatient. The female is more likely to see a doctor among individuals with limitations.

Figures 4 and 5 show that the individuals' education level, age, family income, and the number of chronic diseases are correlated with the visits to the office physician. With the increase of the number of chronic diseases, the visits to office physicians are increasing. The impact of family income on the visits seems to be relatively stable and slightly ascending (Fig. 5b).

Table 1 Variable definitions and summary statistics

Variable	Definition	Mean	Std	Min	Max	Frequency
OFP (numerical)	Number of physician office visits	5.77	6.76	0	89	—
OFP flag (categorical)	=1 physician office visit	0.84	0.36	0	1	0:683/1:3723
EXCLHLTH (categorical)	=1 self-health is excellent	0.08	0.27	0	1	0:4063/1:343
POORHLTH (categorical)	=1 self-health is poor	0.13	0.33	0	1	0:3852/1:554
NUMCHRON (numerical)	Number of chronic conditions	1.54	1.35	0	8	—
ADLDIFF (categorical)	=1 limits activities of daily	0.20	0.40	0	1	0:3507/1:899
NOREAST (categorical)	=1 living in northeastern	0.19	0.39	0	1	0:3569/1:837
MIDWEST (categorical)	=1 living in the midwestern	0.26	0.44	0	1	0:3249/1:1157
WEST (categorical)	=1 living in the western	0.18	0.39	0	1	0:3608/1:798
AGE (numerical)	Age in years divided by 10	7.40	0.63	6.60	10.90	—
BLACK (categorical)	=1 African American	0.12	0.32	0	1	0:3890/1:516
WEST (categorical)	=1 male	0.40	0.49	0	1	0:2628/1:1778
MARRIED (categorical)	=1 married	0.55	0.50	0	1	0:2000/1:2406
SCHOOL (numerical)	Number of years of education	10.29	0.50	0	18	—
FAMINC (numerical)	Family income in $10,000 s	2.53	2.92	−1.01	54.84	—
EMPLOYED (categorical)	=1 employed	0.10	0.30	0	1	0:3951/1:455
PRIVINS (categorical)	=1 covered by private health insurance	0.78	0.42	0	1	0:985/1:3421
MEDICAID (categorical)	=1 covered by Medicaid	0.09	0.29	0	1	0:4004/1:402

Categorical variable value is 0 or 1.Frequency represents the count number

4.2 Model selection

In this section, we briefly explain the classical count regression models and supervised machine learning models utilized in this paper.

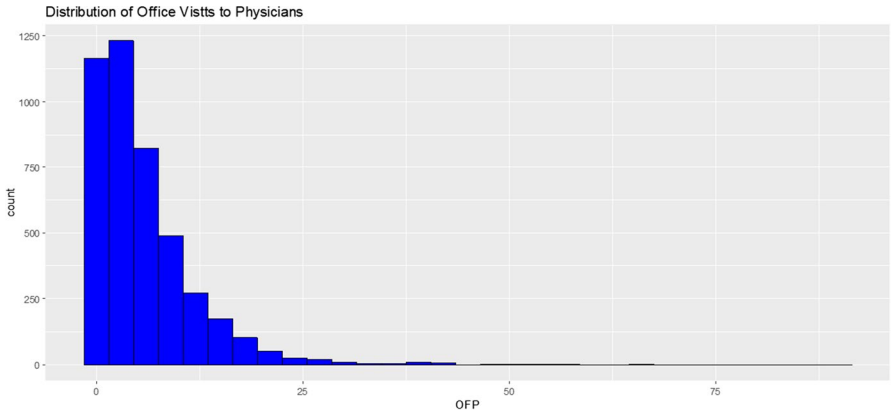

Fig. 2 Distributions of OFP

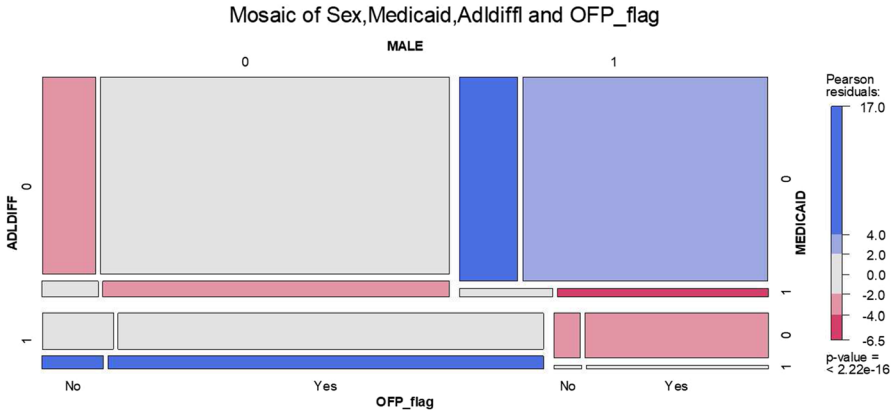

Fig. 3 Features on whether to visit office physician

4.2.1 Traditional count regression model

4.2.1.1 Standard discrete model The standard Poisson regression is a widely used count data model. This model is a special case of the generalized linear model by logarithm transformation of the response, and a benchmark count regression model. The standard Poisson regression model can be characterized by equi-dispersion, but many count data models do not satisfy the property. Instead, the data are usually over-dispersed in real cases. Negative binomial regression, as an over-dispersed model, is proposed to overcome the limitation. Therefore, modeling count data using a standard negative binomial model has become a foremost method of analyzing count response models with the properties of the right-skewed and heavy tail. In this study, we adopt the standard Poisson and negative binomial regression models as the benchmark analysis of the single-stage model.

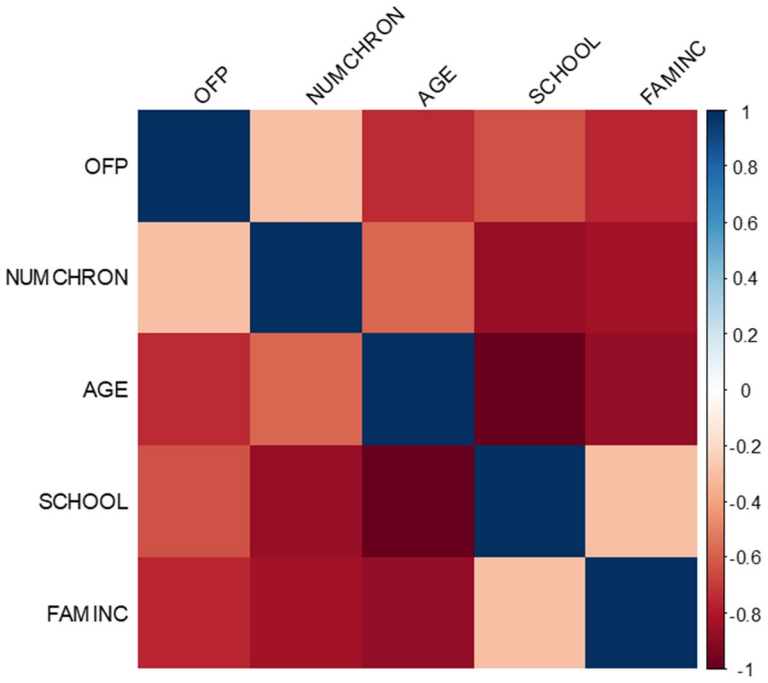

Fig. 4 Correlation among features

4.2.1.2 Zero-inflated model The actual count data are not only over-dispersed and right-skewed but zero-inflated as shown in Fig.2. The single-stage Poisson and negative binomial models fail to handle the problem, while the two-stage zero-inflated model works. As previously illustrated, zero-inflated Poisson, zero-inflated negative binomial, hurdle-Poisson, and hurdle-negative binomial models are selected here as another benchmark technique in our experimental study to explore the factors and demands affecting health care decision-making of the elders.

4.2.2 Machine learning model

Five machine learning methods are employed in this study: logistic, decision tree, random forest, support vector, and XGBoost. Logistic regression is a commonly used classification model that can resolve the binary outcome by logit mapping transformation. Unlike other machine learning algorithms, the logistic regression model, which is often compared to other techniques, can be combined with the classical statistical inference theory to make a hypothesis and predict the results. It is evident from previous a great deal of literature on traditional statistical models that logistic regression has been utilized to make a decision behavior analysis. The decision tree is another classification method proposed by Breiman et al. [10], using historical data to formulate so-called decision rules that can be drawn as a tree-like diagram. Because the decision tree does not make any assumptions towards the function form, and takes the information of responses into account at

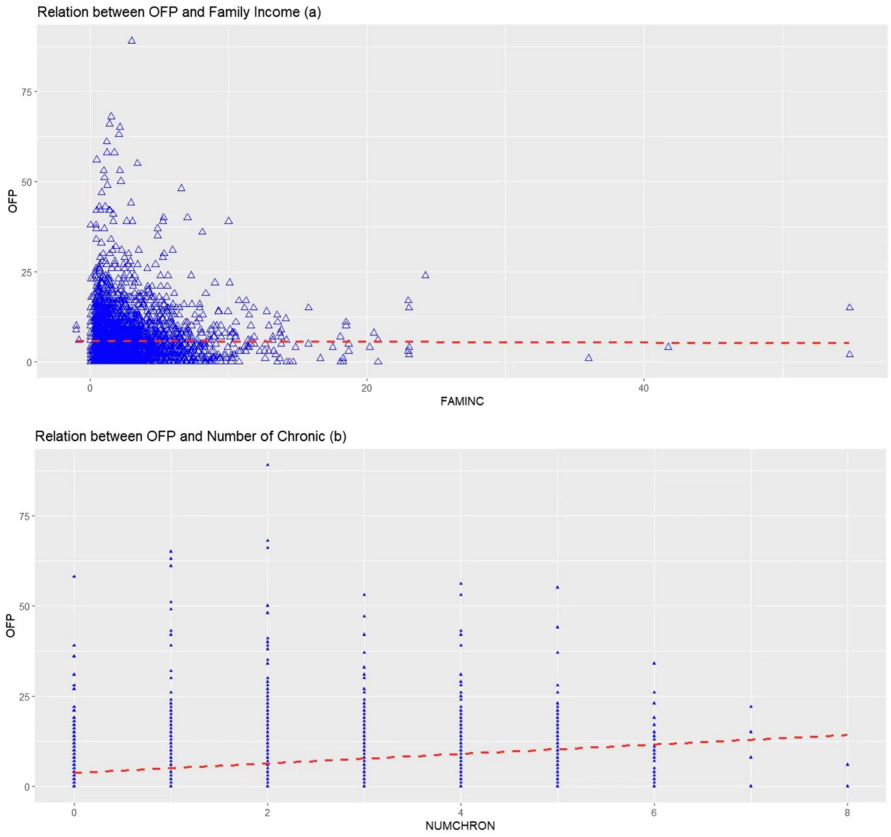

Fig. 5 Chronic and family income on visits to office physician

the splitting nodes, it will not be affected by noise. Hence, the results are relatively robust. Random forest, also proposed by Breiman [14], is an assembled algorithm model composed of multiple decision trees. Breiman [14] introduced a bootstrap approach to randomly select some variables as candidate splitting factors to reduce the correlation between trees based on the bagging method when each node of the decision tree is split. The random forest algorithm sacrifices a small amount of bias in exchange for a greater decline in variance, and thus achieves in reducing the mean square error. The support vector machine (SVM) introduced by Cortes and Vapnik [15] has been proven a useful technique in multiple areas. The idea of SVM is that the hyperplane composed of some support vectors divides different types of sample points. Whether the sample points are linearly separable or nonlinearly separable, the linear separability in kernel space is realized using the application of kernel functions. The final classifier formed by SVM only depends on some support vectors, therefore, it results in good robustness and avoids the "dimension disaster". Compared to other statistical models, the SVM algorithm has a more efficient performance in predicting. The extreme gradient boosting decision tree (XGBoost), as

an advanced implementation of gradient boosting, was proposed by Chen et al. [16]. It is an ensemble learning algorithm that is more efficient, accurate and scalable than RF and decision tree method. To achieve this goal, the XGBoost algorithm can customize the loss function and use the first and second derivatives of the loss function with the Taylor series to approximate the loss function. It is also able to control the complexity of the model, prevent overfitting, and enhance its generalization by adding a regular term to the loss function. In addition, XGBoost supports multiple types of basic classifiers, and provides a strategy similar to the random forest to sample to improve the efficiency and accuracy of the forecasting process. In this paper, we use logistical regression and classical zero-truncated Poisson regression as the benchmark of classification and regression models to analyze outpatient decision-making and forecast the health care demand.

4.3 Experimental setup

In our work, we first construct count regression models based on the characteristics of the elders' health care demand (as shown in Fig. 2) by using traditional statistical methods, and select the optimal model according to AIC, BIC, and other criteria. Next, we extend the two-stage hurdle models in line with approaches in Sect. 4.2 instead of the logistic regression algorithm in the first stage, and the classification learners are applied to predict the demand. Moreover, the regression learners are utilized to replace traditional count regression models. Finally, to account for the performance of those models, measures of the quality of the classification algorithm are built such as the sensitivity, specificity, accuracy, F1 scores, ROC curve, and so on, meanwhile, the mean squared error (MSE) and root mean squared error (RMSE) are employed to evaluate the performance of regression learners.

5 Results

In this section, the predictive performance of single-stage, two-stage, and generalized hurdle models about healthcare demand is presented and analyze.

5.1 Standard single-stage count model

As previously stated, Fig. 2 shows that the problem of overdispersion may exist in the distribution of health care demand, and Table 2 has proved such phenomenon.

Table 2 Test of over dispersion

Over-dispersion test	Obs.Var versus theory.Var	Statistic	*p*-value
Poisson	7.912013	34852.42	0

Note: In this paper, we use the "qcc" package in R software to test the excessive dispersion of Poisson model to determine the over dispersion problem of discrete data. Because the calculated p value overflows the set digits after retaining the Decimal separator, the p value is displayed as 0

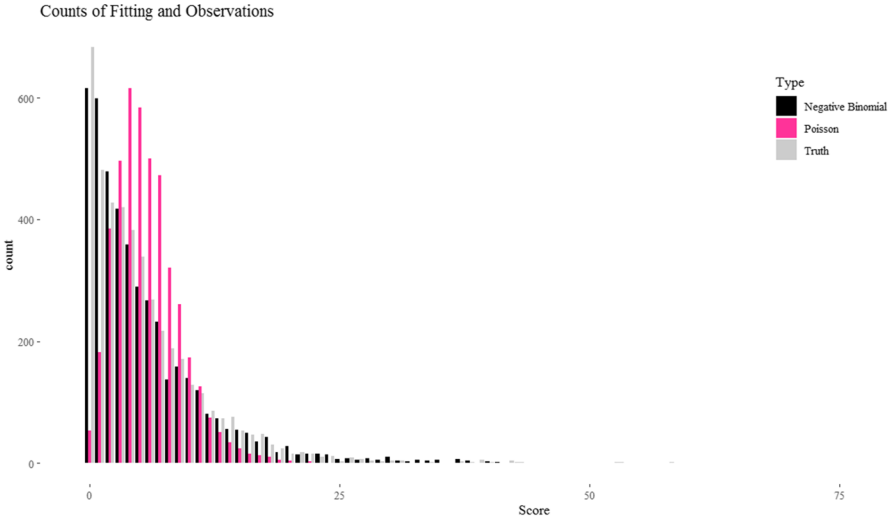

Fig. 6 Fitting of Poisson and NB versus observations

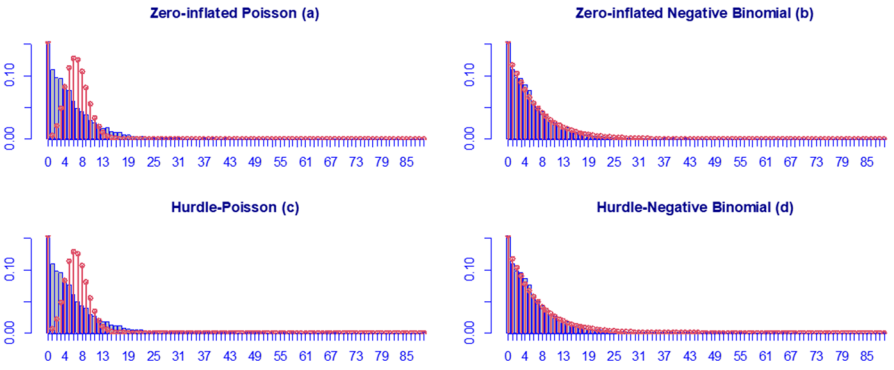

Fig. 7 Barplot and fitted ZIP, ZINB, Hurdle-Poisson, and Hurdle-NB for demand data

We also make use of the standard Poisson and negative binomial regression models to fit the data, and find their goodness-of-fit is poor as shown in Fig. 6. The negative binomial is better than the Poisson, but neither depicts the zero-inflated of the data.

5.2 Traditional two-stage count model

For the two-stage count models, Fig. 7 shows that the four models can fit the probability value of zero. The two negative binomial models perform better than the Poisson models as indicated by Fig. 7, especially in the long tail on the right

Table 3 Two-stage count models' comparisons

	Poisson	ZIP	NB	ZINB	Hurdle-P	Hurdle-NB
AIC	36303.13	32647.61	24475.13	24303.86	32647.63	24290.98
BIC	36411.78	32864.89	24583.77	24527.53	32864.91	24514.66
Vuong_test	−17.58050***		−13.56758***		−13.50872***	

*** represents the p value is very small (< 2.22e-16)

Table 4 Results of Hurdle-negative binomial regression

	Stage-one			Stage-two		
	Estimate	Std. Error	Pr(>|z|)	Estimate	Std. Error	Pr(>|z|)
Intercept	5.049379	(1.589618)	0.00149**	1.6650919	0.2156372	1.15e−14 ***
EXCLHLTH	0.346837	0.291470	0.23406	−0.3358186	0.0634997	1.23e−07 ***
POORHLTH	−0.097255	0.468260	0.83547	0.3044647	0.0475806	1.56e−10 ***
NUMCHRON	−1.181773	0.172340	7.02e−12***	0.1470184	0.0121903	< 2e-16 ***
ADLDIFF	0.124222	0.334932	0.71072	0.0976583	0.0409017	0.01696 *
NOREAST	−0.166406	0.276581	0.54740	0.1017580	0.0432577	0.01865 *
MIDWEST	−0.357921	0.288030	0.21400	−0.0180926	0.0399191	0.65038
WEST	−0.207668	0.312645	0.50654	0.1240965	0.0444021	0.00519 **
AGE	−0.662740	0.206271	0.00131**	−0.0772319	0.0266343	0.00374 **
BLACK	0.466188	0.262326	0.07555	−0.0225782	0.0544869	0.67860
MALE	0.892536	0.232030	0.00012***	−0.0095780	0.0352598	0.78590
MARRIED	−0.700049	0.245267	0.00431**	−0.0860287	0.0363792	0.01804 *
SCHOOL	−0.096237	0.029825	0.00125**	0.0209491	0.0046516	6.68e−06 ***
FAMINC	0.009383	0.045387	0.83622	−0.0008531	0.0054238	0.87501
EMPLOYED	−0.298044	0.342002	0.38350	0.0058322	0.0523444	0.91128
PRIVINS	1.084298	0.247462	1.18e−05***	0.2420713	0.0488197	7.10e−07 ***
MEDICAID	−0.694963	0.382680	0.06936	0.1991560	0.0653244	0.00230 **

Signif. codes: 0'***',0.001'**',0.01'*',0.05'.',0.1

side of the figure. However, we can't distinguish the better one between the negative binomial model.

Further, it can be inferred that ZIP is better than the Poisson model, the ZINB is superior to the negative binomial model, and the hurdle-negative binomial excels hurdle-Poisson model based on the Vuong test in Table 3. We conclude that the hurdle-negative binomial model is the best one because of its smallest AIC and BIC.

The results of hurdle-negative binomial regression are shown in Table 4. Private insurance plan, age, health status, and a number of chronic seem to be more important in predicting the health care demand, which is consistent with the literature [42].

5.3 Generalized hurdle model

5.3.1 Empirical results

We divide 4406 data into training data set, test data set in terms of the ratio of 80:20, and randomly select the samples with replacement. For the generalized hurdles, Table 5 presents the precision, accuracy, recall, F1 score and AUC in stage one, and shows the MSE and RMSE in stage two, which are from the test set. It is noted that there is little variation in the performance of indicators assessed in stage-one models, although XGBoost outperforms others on precision, decision tree model is superior to others on recall, and logistic model overmatches others on AUC. However, none of the models outperforms others in terms of all respects. The ROC curve can reflect the classification ability of the classification learner. As shown in Fig. 8 below, the classification effects of logical regression and random forest classifier are not significantly different, and they are significantly better than other classifiers. According to the residual boxplot of the classification results listed in Fig. 9, the median residual of the classification prediction of random forest algorithm is relatively smaller than that of logistic regression. Part of the reasons may be the right thick tail and extra zeros of demand data.

In the second stage, two zero-truncated count models (zero-truncated Poisson, ZTP; zero-truncated negative binomial, ZTNB) are used as benchmarks to predict the health care demand of the elders after their first physician office visit. The MSE and RMSE of models based on machine learning are significantly less than the benchmark models as shown in Table 5. Among the three regression learners, the MSE and RMSE of random forest are the smallest and the distribution of RF residual is approximately distributed to the normal distribution with the 0 mean (Figs. 10 and 11).

Table 5 Performance of generalized hurdle model

Stage-one					
	Logistic	Tree	RF	SVM	XGBoost
Precision	0.8422	0.8379	0.8462	0.8394	0.8596
Accuracy	0.8413	0.8379	0.8413	0.8379	0.8129
Recall	0.9973	1.0000	0.9905	0.9973	0.9283
F1 Score	0.9133	0.9118	0.9127	0.9116	0.8668
AUC	0.7581	0.5000	0.7408	0.6078	0.7282
Stage-two					
	ZTP	ZTNB	RF	SVM	XGBoost
MSE	63.8224	64.7970	36.6061	38.5715	41.3253
RMSE	7.9889	8.0497	6.0503	6.2106	6.4285

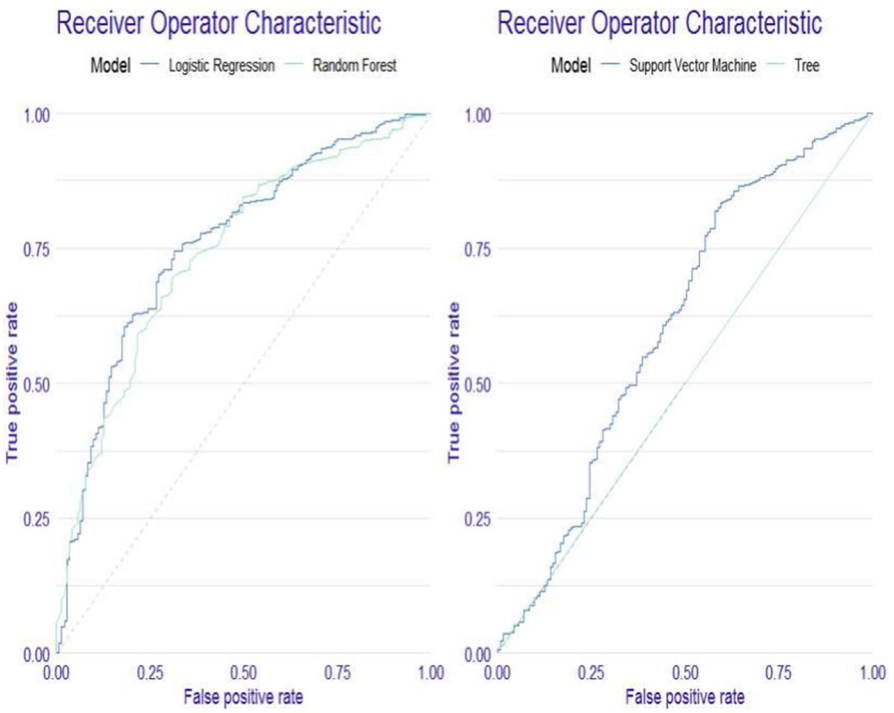

Fig. 8 Receiver operator characteristic

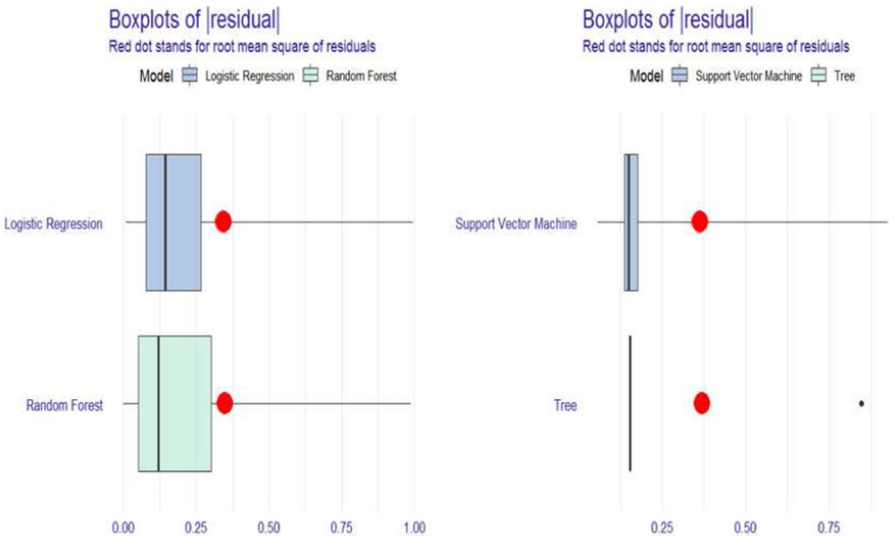

Fig. 9 Boxplot for residuals

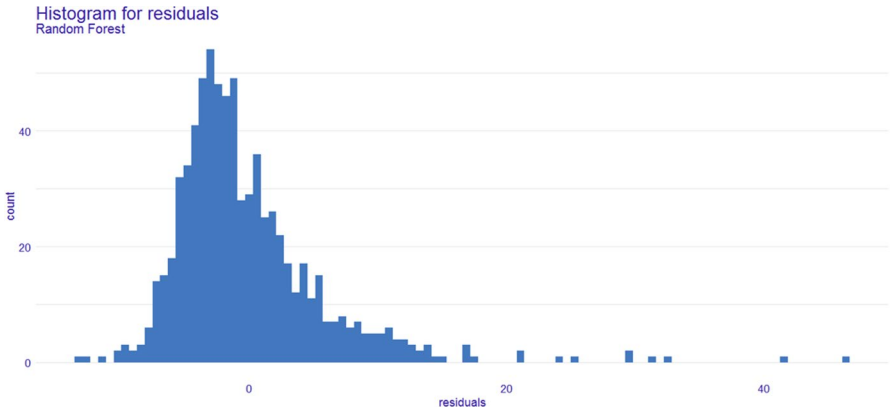

Fig. 10 Histogram of residuals for the RF model

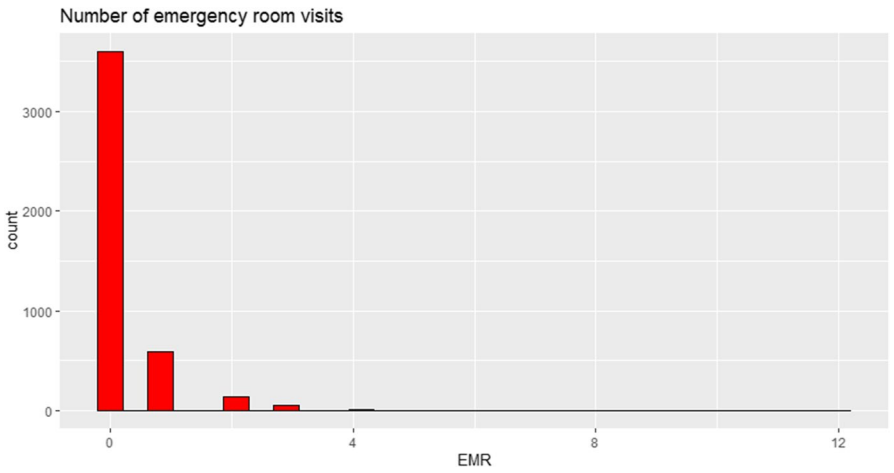

Fig. 11 Number of emergency room visits

Based on the previous analysis, we choose the RF as the best model of the two-stage generalized hurdle models to further discuss the health care demand of the elders in this study.

5.3.2 Robustness test

We selected the number of emergency room visits in Deb and Trivedi (1997) [43] as the dependent variable (EMR, number of emergency room visits). The independent variables remained unchanged. The distribution characteristics of the number of emergency room visits are as follows. It is clear that the number of emergency visits is characterized by zero inflation. It is difficult to fit the distribution of general counting models such as Poisson or negative binomial, as shown in Fig. 12.

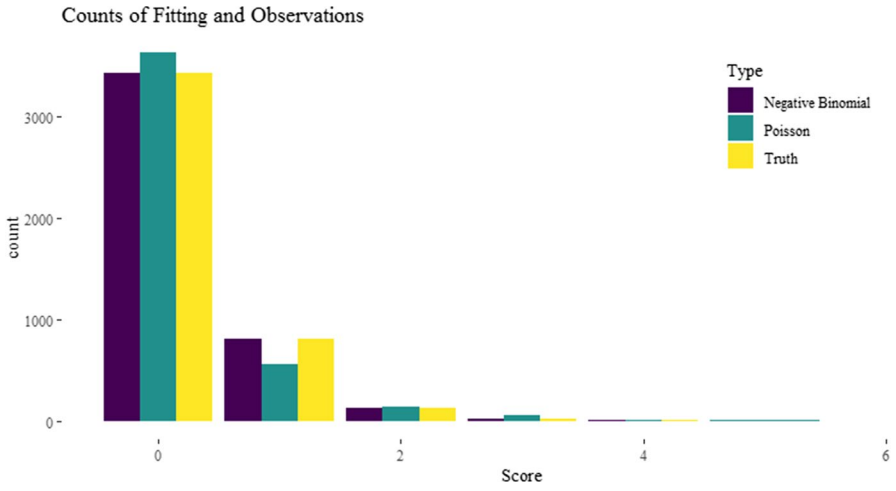

Fig. 12 Counts of fitting and observations

The zero expansion model and Hurdle counting model are more suitable for data that merges 0 out of the left hurdle, especially the predicted value coming out of zero, as shown in Fig. 13.

Due to the problem of zero expansion in the distribution of emergency number and the large proportion of zero visits, the ZINB calculation results were convergent. According to the Vuong test results in the table below, Hurdle-NB is more suitable for this type of data than the traditional Zero-inflated model (Table 6).

We use the extended Hurdle model to predict emergency treatment data and the performance of the model is as follows. Table 7 shows that the performance of the classification learner-based XGBoost model in the first stage is superior to that of the logistic model, and the stochastic forest-based model in the regressor-based model in the second stage is significantly better.

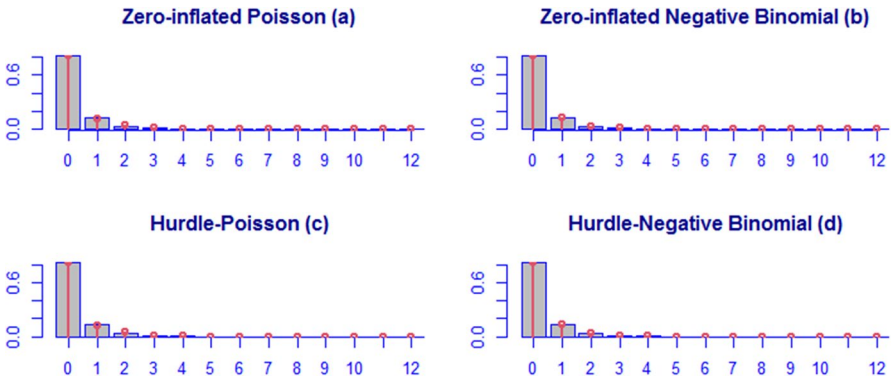

Fig. 13 Models of extra zeros

Table 6 Two-stage count models' comparisons

	Poisson	ZIP	NB	ZINB	Hurdle-P	Hurdle-NB
AIC	5655.371	5437.657	5378.569	-(misconvergence)	5435.827	5381.021
BIC	5764.013	5654.942	5487.212	-(misconvergence)	5653.111	5604.696
Vuong_test	−5.967776***		-		−2.300837***	

*** represents the p value is very small (< 0.01)

Table 7 Performance of generalized hurdle model

Stage-one			
	Logistic	RF	XGBoost
Precision	0.3333	0.3333	0.3333
Accuracy	0.7937	0.7891	0.8118
Recall	0.0055	0.0276	0.8642
F1 Score	0.0109	0.0510	0.8372
AUC	0.6439	0.6351	0.6645
Stage-two			
	ZTP	RF	XGBoost
MSE	4.6183	1.29252	1.4610
RMSE	2.1490	1.13689	1.2087

Accuracy represents the proportion of correct classification of the model; Precision represents the accuracy of the model; the higher the value is, the more accurate the prediction will be; Recall represents the recall of the model, and the number of values greater than zero in the sample will be correctly selected; AUC represents the comprehensive consideration of classification results; the closer the value is to 1, the better the classification effect will be. As you can see from Table 7 above, the XGBoost model performed best during the first phase of the classification process. In the second stage regression model, the smaller the mean square error and the root mean square error, the better the model effect. According to the results, the RF model performed best.

We also use auto insurance data as a robustness test. The extended Hurdle model in Table 8 and Fig. 14 evaluated the number of auto insurance claims in SAS Miner data. The Enterprise Miner data set is supplied by SAS. It consists of claims data on a class of auto insurance policies and includes policies on which there were no claims. A total of 10303 sample data, 33 explanatory variables. The number of claims with a one-year insurance term is selected as the explained variable, and its distribution is shown in Fig. 14.

Due to the possibility of data and the existence of multi-zero phenomenon as well as the non-balance problem, the classification results of the first stage have

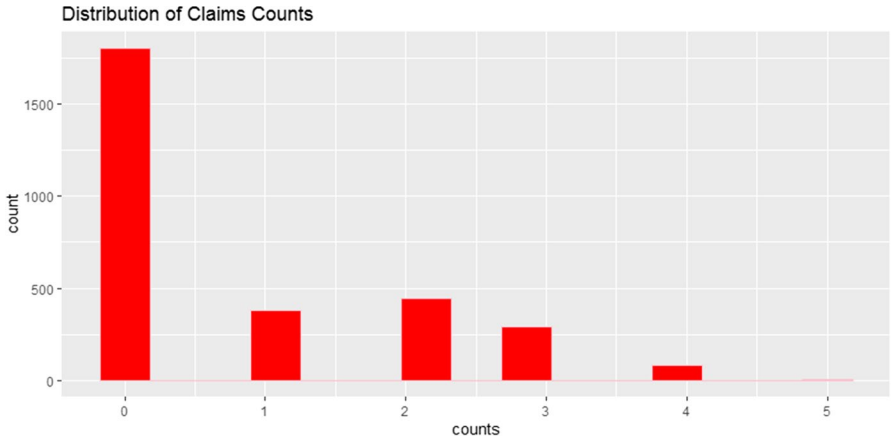

Fig. 14 Claims counts

Table 8 Performance of generalized hurdle model

Stage-one					
	Logistic	Tree	RF	SVM	XGBoost
Precision	0.5789	0.5705	0.5374	0.5823	0.6011
Accuracy	0.6644	0.6544	0.6461	0.6628	0.5576
Recall	0.4342	0.3728	0.5044	0.4035	0.4515
F1 Score	0.4962	0.4509	0.5204	0.4767	0.4990
AUC	0.6899	0.6724	0.6900	0.6916	0.5262
Stage-two					
	ZTP	ZTNB	RF	SVM	XGBoost
MSE	3.5788	3.5790	0.9534	0.9525	1.2138
RMSE	1.8918	1.8918	0.9764	0.9759	1.1017

little difference, but we can still see that SVM classification algorithm has the best effect. In the second stage, the prediction error of SVM is the smallest.

In conclusion, the model described in the paper remains significant when tested against other Hurdle data, so we consider the Hurdle model to be effective in solving the problem.

5.4 Further discussion

5.4.1 Variables importance

It is known that there are any methods of assessing explanatory variable's importance that depict particular elements of the structure of the model in classical

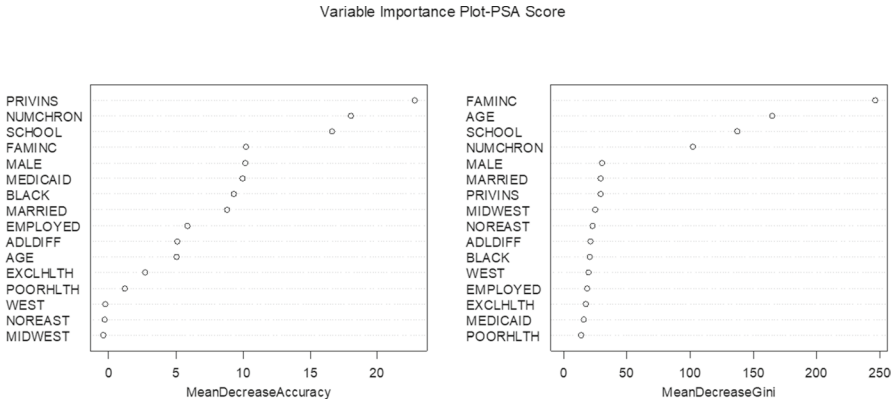

Fig. 15 Variable importance for RF in stage one

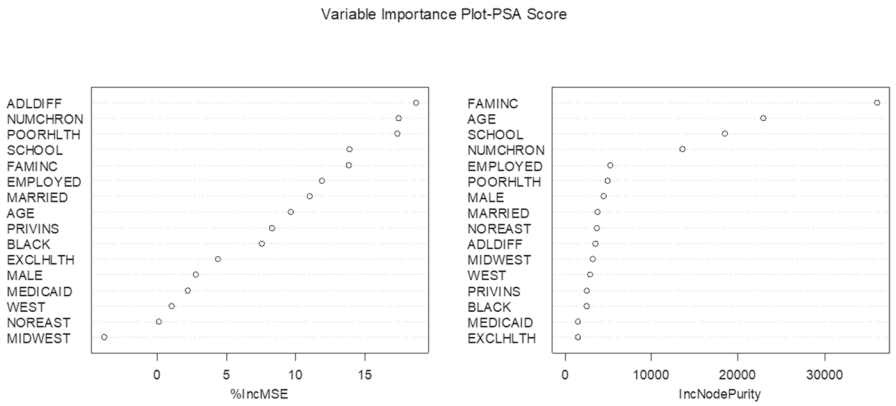

Fig. 16 Variable importance for RF in stage two

statistical theory. Generally speaking, they are model-specific methods, whose variable-importance measure rely on strict statistical assumptions that are hard to meet in reality. We make use of a model-agnostic method that has no requirements for the structure of the model. Therefore, it can be utilized to any predictive model. The method comes from the variable-importance measure proposed by Breiman [14] for random forest. The idea means that the model's performance will worsen if we permute the values of the variable, which implies that the variable is important. The greater the performance change of the model, the more important the variable is. In most cases, the plots of variable-importance measures are used to interpret the importance as shown in Figs. 15 and 16.

The important causal variables are in line with the RF model. The most prominent variables are consistent with the variables of Hurdle-NB from Table 4, which are NUMCHRON, SCHOOL, PRIVINS, MALE, and MARRIED in stage one. In stage two, the variables of NUMCHRON, FAMINC, SCHOOL, AGE, PORRHLTH,

MARRIED, and EMPLOYED are most important, which are basically consistent with the results of the second stage of the hurdle. We can find that the number of chronic diseases, education, and marriage of an elder affects his (or her) decision-making medical treatment, as well as the health care demand.

Although the plots of variable-importance measures offer several advantages, the main disadvantage of this approach is the importance of the variable often depends on the random nature of the permutations. As a result, we may attain different results because of different permutations.

5.4.2 Interpretability

The previous approach focuses on the important variables affecting the response from a global dataset level. Now, we expect to discover variables that contribute to the result for a model's prediction in a single instance. Some methods, such as LIME [44], Shapley values [45], SHAP [46] and so on, can be used to answer this question. Unfortunately, no single best approach exists. In this study, we introduce the break-down method [47] which is a mixture of ideas from PDDs and Shapely values. The break-down (BD) plots are recommended to present variable attributions by decomposing contributions of different explanatory variables in the model's prediction.

To evaluate the contribution of individual explanation to a particular single instance prediction, we choose a high-income female in stage one and select a low-income male in stage two as the representative to discuss the impact of features on health care demand. As indicated in Fig. 17, we can conclude that average predicted probability of visit to a physician office over all elder people is equal 0.861. The selected elders are more likely to make a visit than the average because her predictive probability is 0.993. Family income is a main cause that is above the average. The covered Medicaid and limited activities both increase the probability of her visit. In stage two, as illustrated in Fig. 18, the average frequency number of visits is 6.91. The predicted visits of selected elders are 4.636, which is below the average.

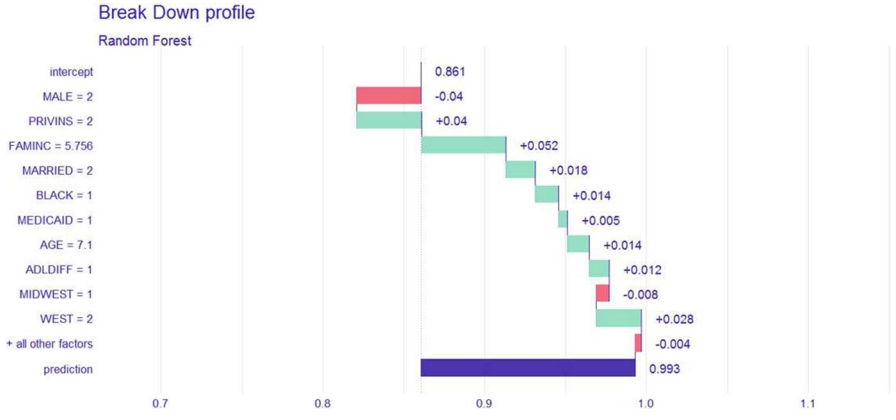

Fig. 17 Break-down plot for the RF in stage one

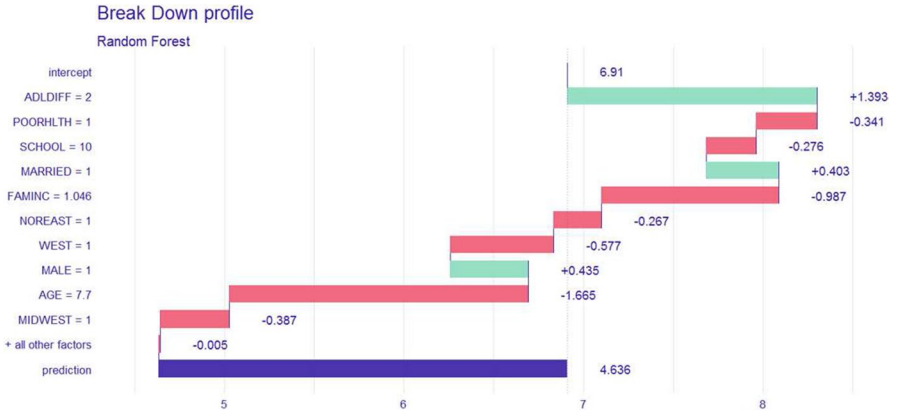

Fig. 18 Break-down plot for the RF in stage two

The preliminary reason perhaps is his age. It is seen from Fig. 18 that lower family income and education level can reduce his visits, instead, the covered Medicaid and unlimited activities enhance the frequency.

Generally speaking, the break-down approach is easy to understand. However, the drawback is that the method shows only the additive attributions, which indicates that the choice of the ordering of the explanatory variables used to calculate the variable importance is significant. Besides, BD plots may include many explanatory variables that contributes little to the instance prediction (shown "all other factors" in Figs. 17 and 18) when the model include a number of variables. Thus, it is impossible for us to intuitively find the degree of importance of some variables.

5.4.3 Partial dependence plot

The variable importance measure only ranks the importance of feature variables, and does not determine the marginal effect of feature variables on dependent variable. A partial dependency plot offers an approach to settle this problem [48]. PDP may exhibit the marginal effect of one or two features on the predictions of machine learning. It also relies on whether the relationship between the response and feature is nonlinear, monotonous, or more complex.

In this section, we investigate how the features affect the health care demand on men and women by PD plots in two-stage, which is shown in Figs. 19, 20, 21 and 22. From the figures, we can conclude that the male is more likely to make medical decisions. Figure 19 reveals that there is an inverse U-shaped relationship between the probability of visit and the number of chronic diseases, which means that extreme numbers of chronic diseases do not affect the making-behavior of visits. Figures 20 and 21 depict that as the education level is improved, the awareness of health care demand is increasing., when the age increases to 90, the possibility of seeing a doctor decreases. In the second stage, the number of visits and age of the elders showed a U-shaped relationship. Part of the reason is that after the age of 90, health problems of the elders mainly are caused by natural aging. The impact of

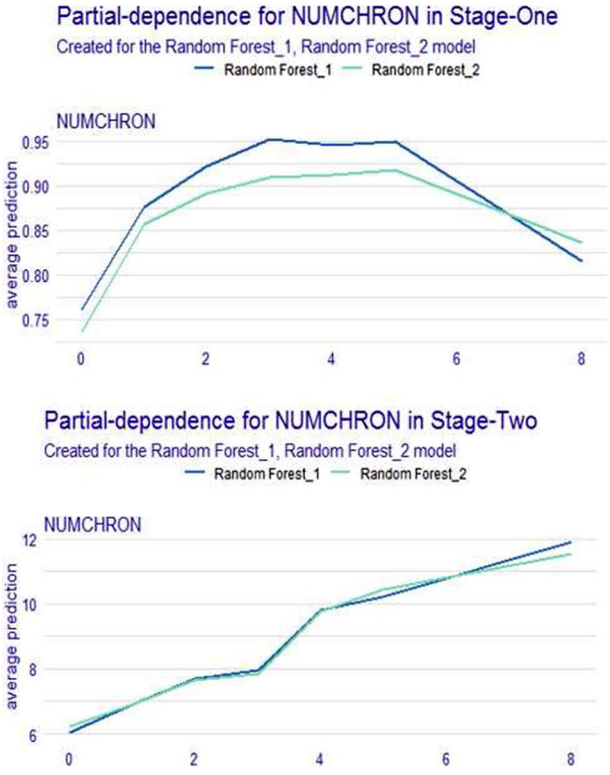

Fig. 19 Partial-dependence for NUMCHRON

family income on medical decision behavior also shows an inversed U relationship. Nevertheless, Fig. 22 shows that as their family income reaches a certain level, the marginal effect of family income decreases towards the decision-making process.

If one feature of PDP are not related to the others, the PDP can perfectly represent the impact of the features on prediction. Yet, in practice, it is often difficult to attach the completely independent, resulting in being hard to discover heterogeneity, which is a major limitation of PDP.

Note: The blue line represents female (Random Forest_1) and the green line represents male(Random Forest_2) in Figs. 19, 20, 21 and 22.

6 Conclusion and future work

Excess zeros often appear in count data, which causes underestimation of the standard errors and the overestimation of the significance of parameters without zero inflation. In this study, we proposed generalized hurdle models based on machine learning to solve the zero-inflated problem in health care demand data for the elders. Decision tree, RF, SVM, and XGBoost learners, instead of the logistical model, are

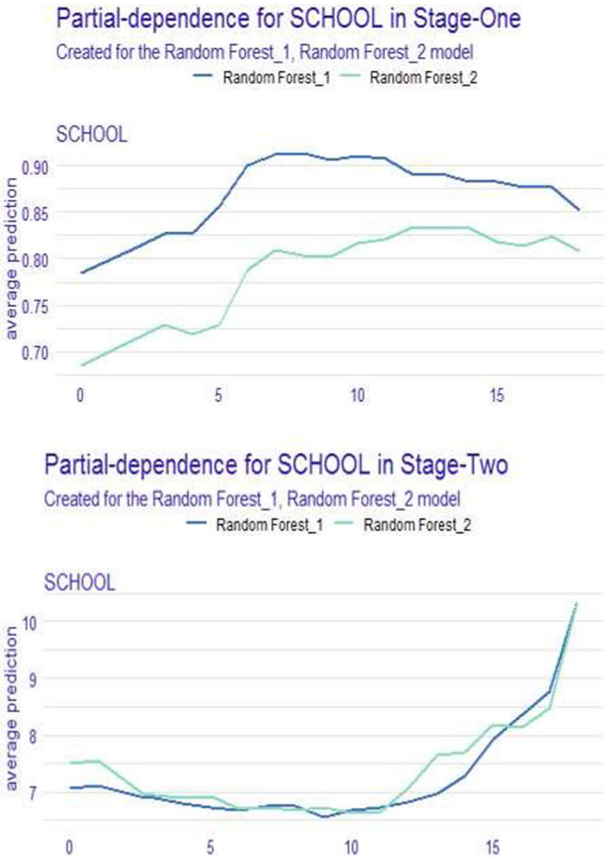

Fig. 20 Partial-dependence for SCHOOL

introduced to predict the elders' decision-making behavior in stage one. In order to predict the elders' health care demand in the second stage, we utilize the RF, SVM, and XGBoost regression learners, and compare them with the zero-truncated count models. It is shown that the RF is the optimal model in both stages based on the performance criteria. To measure the importance of the variable, we adopt the variable-importance method and find that the most important variables found via RF are basically consistent with those found by using traditional hurdle negative binomial regression in both stages, which indicates that our generalized hurdle model is effective and attractive. Besides, to explore the marginal effect of continuous explanatory variables on the response variable, we apply the PDP approach. As previously stated, the break-down method is applied to analyze the health care demand of an individual, and discuss the impact of feature factors on the individual. To the best of our knowledge, it is the first time to use these techniques in health care.

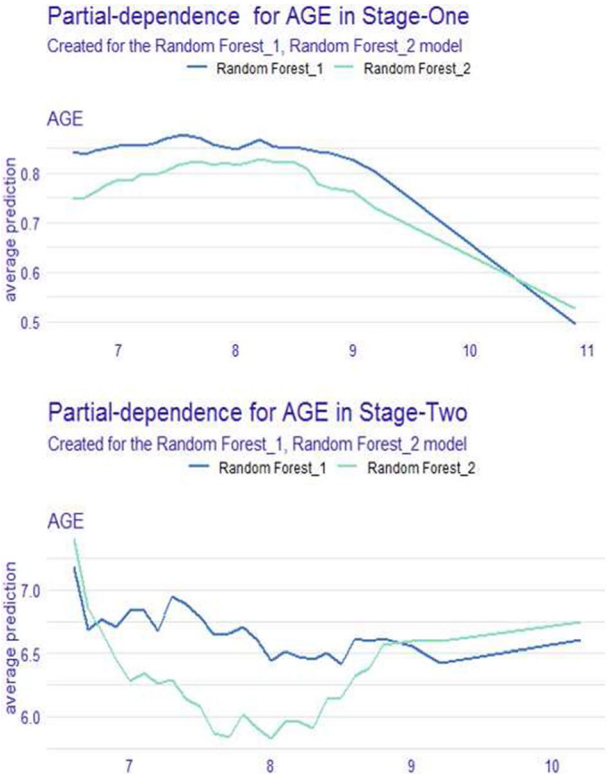

Fig. 21 Partial-dependence for AGE

The main novel contribution of this empirical study to the literature of the hurdle model is three-fold. First and foremost, we generalized the hurdle models based on machine learning and propose a new framework, which is a state-of-the art technique used to deal with the problem of two-stage multi-outcomes. The two-stage hurdle model is only a special case with binary outcomes. Second, the generalized hurdle model frame can include the traditional count regression and construct many combined models of count and machine learning models, which gives us a lot of choices. Third, we implement model interpretation by using the latest approaches including variable importance, break-down plot, and partial dependence plot.

One limitation of our study is that as there is no other available experimental data in the literature, it is difficult to analyze the two-stage multi-outcomes model. Another is that as the datasets used for training are randomly selected, it could make the results not robust.

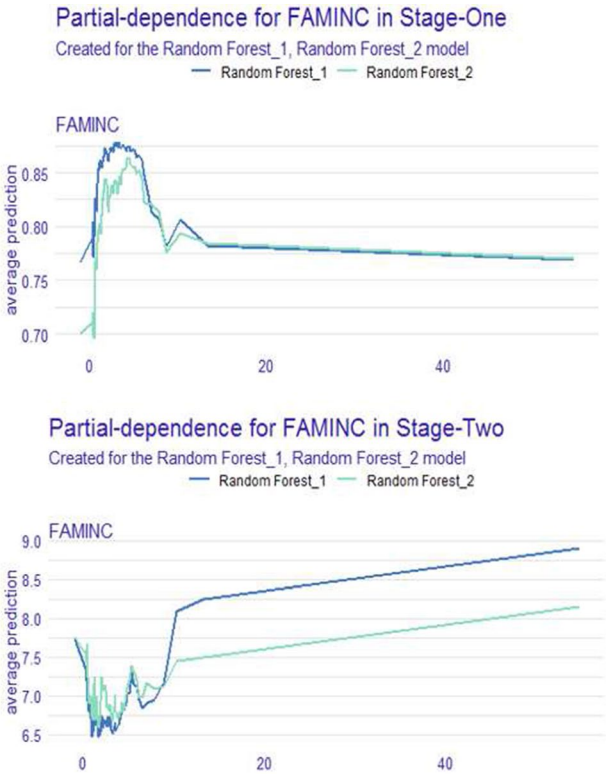

Fig. 22 Partial-dependence for FAMINC

In addition, we plan to develop other learning techniques such as the use of assembling learning, deep learning, and so on. We also hope to explore novel modeling frameworks to predict two-stage traditional statistical regression models.

Acknowledgements All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by XX, TY, JG and DC. The first draft of the manuscript was written by XX and DC, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Xin Xu acknowledges financial support from the National Social Science Foundation of China (22BTJ016).

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Hartman M, Martin AB, Washington B, Catlin A (2022) National health expenditure accounts team: national health care spending in 2020: growth driven by federal spending in response to the COVID-19 pandemic: national health expenditures study examines US health care spending in 2020. Health Aff 41(1):13–25
2. Rana RH, Alam K, Gow J (2021) Financial development and health expenditure nexus: a global perspective. Int J Financ Econ 26(1):1050–1063
3. Chen T, Zhang H, Zhang B (2019) A semiparametric marginalized zero-inflated model for analyzing healthcare utilization panel data with missingness. J Appl Stat 46(16):2862–2883
4. Cameron AC, Trivedi PK (1986) Econometric models based on count data: comparisons and applications of some estimators and tests. J Appl Econ 1(1):29–53
5. Abiodun GJ, Makinde OS, Adeola AM, Njabo KY, Witbooi PJ, Djidjou-Demasse R, Botai, JO (2000) A dynamical and zero-inflated negative binomial regression modelling of malaria incidence in Limpopo Province, South Africa. Int J Env Res Pub He 16(11)
6. Neelon B, O'Malley AJ, Smith VA (2016) Modeling zero-modified count and semicontinuous data in health services research part 1: background and overview. Stat Med 35(27):5070–5093
7. Rose CE, Martin SW, Wannemuehler KA, Plikaytis BD (2006) On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. J Biopharm Stat 16(4):463–481
8. Xu X, Ye T, Chu D (2021) Generalized zero-adjusted models to predict medical expenditures. Comput Intell Neurosci
9. Xu X, Chu D (2021) Modeling hospitalization decision and utilization for the elderly in China. Discrete Dyn Nat Soc 1–13
10. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Routledge, New York
11. Frölich M (2006) Non-parametric regression for binary dependent variables. Econ J 9(3):511–540
12. Mullahy J (1986) Specification and testing of some modified count data models. J Econ 33(3):341–365
13. Lambert D (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34(1):1–14
14. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
15. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
16. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K (2015) Xgboost: extreme gradient boosting 1(4), 1–4. R package version 0.4-2
17. Samson D, Thomas H (1987) Linear models as aids in insurance decision making: the estimation of automobile insurance claims. J Bus Res 15(3):247–256
18. Greene WH (1994) Accounting for excess zeros and sample selection in Poisson and negative binomial regression models
19. Cameron AC, Trivedi PK, Milne F, Piggott J (1988) A microeconometric model of the demand for health care and health insurance in Australia. Rev Econ Stud 55(1):85–106
20. Dionne G, Vanasse C (1989) A generalization of automobile insurance rating models: the negative binomial distribution with a regression component. ASTIN Bull J IAA 19(2):199–212
21. Willmot GE (1987) The Poisson-inverse Gaussian distribution as an alternative to the negative binomial. Scand Actuar J 1987(3–4):113–127
22. Bulmer MG (1974) On fitting the Poisson lognormal distribution to species-abundance data. Biometrics, 101–110
23. Consul PC (1989) Generalized Poisson distributions: properties and applications
24. Zou Y, Geedipally SR, Lord D (2013) Evaluating the double Poisson generalized linear model. Accid Anal Prev 59:497–505
25. Sellers KF, Shmueli G (2010) A flexible regression model for count data. Ann Appl Stat 943–961
26. Yip KC, Yau KK (2005) On modeling claim frequency data in general insurance with extra zeros. Insur Math Econ 36(2):153–163
27. Neelon BH, O'Malley AJ, Normand SLT (2010) A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. Stat Modell 10(4):421–439
28. Preisser JS, Das K, Long DL, Divaris K (2016) Marginalized zero-inflated negative binomial regression with application to dental caries. Stat Med 35(10):1722–1735

29. Liu X, Zhang B, Tang L, Zhang Z, Zhang N, Allison JJ, Srivastava DK, Zhang H (2018) Are marginalized two-part models superior to non-marginalized two-part models for count data with excess zeroes? estimation of marginal effects, model misspecification, and model selection. Health Serv Outcomes Res Method 18(3):175–214
30. Chen K, Huang R, Chan NH, Yau CY (2019) Subgroup analysis of zero-inflated Poisson regression model with applications to insurance data. Insur Math Econ 86:8–18
31. Gurmu S (1998) Generalized hurdle count data regression models. Econ Lett 58(3):263–268
32. Ehsan Saffari S, Adnan R, Greene W (2012) Hurdle negative binomial regression model with right Cencored count data. Sort (Barc) 36(2):181–194
33. Baetschmann G, Winkelmann R (2014) A dynamic hurdle model for zero-inflated count data: with an application to health care utilization. Commun Stat Theory Methods (151)
34. Xu X, Chu D (2021) Modeling hospitalization decision and utilization for the elderly in China. Discrete Dyn Nat Soc
35. Sakthivel KM, Rajitha CS (2017) Artificial intelligence for estimation of future claim frequency in non-life insurance. Glob J Pure Appl Math 13(6):1701–1710
36. Gao G, Wang H, Wüthrich MV (2022) Boosting Poisson regression models with telematics car driving data. Mach Learn 111(1):243–272
37. Liu Y, Wang BJ, Lv SG (2014) Using multi-class adaboost tree for prediction frequency of auto insurance. J Bank Financ 4(5):45
38. Lee SCK (2021) Addressing imbalanced insurance data through zero-inflated Poisson regression with boosting. ASTIN Bull J IAA 51(1):27–55
39. Kong S, Bai J, Lee JH, Chen D, Allyn A, Stuart M, Pinsky M, Mills K, Gomes CP (2020) Deep hurdle networks for zero-inflated multi-target regression: application to multiple species abundance estimation. arXiv preprint arXiv:2010.16040
40. Zhang P, Pitt D, Wu X (2022) A new multivariate zero-inflated hurdle model with applications in automobile insurance. ASTIN Bull J IAA 52(2):393–416
41. Deb P, Trivedi PK (1997) Demand for medical care by the elderly: a finite mixture approach. J Appl Econ 12(3):313–336
42. Gurmu S (1997) Semi-parametric estimation of hurdle regression models with an application to medicaid utilization. J Appl Econ (Chichester Engl) 12(3):225–242
43. Deb P, Trivedi PK (1997) Demand for medical care by the elderly: a finite mixture approach. J Appl Econom 12(3):313–336
44. Ribeiro MT, Singh S, Guestrin C (2016) "Why should i trust you?" Explaining the predictions of any classifier. arXiv-1602
45. Shapley LS (1997) A value for n-person games. Classics in game theory 69
46. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30
47. Staniak M, Biecek P (2018) Explanations of model predictions with live and breakDown packages. arXiv preprint arXiv:1804.01955
48. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 5:1189–1232

Authors and Affiliations

Xin Xu[1] · Tao Ye[2] · Jieying Gao[1] · Dongxiao Chu[1]

✉ Dongxiao Chu
chudongxiao@cueb.edu.cn

Xin Xu
xuxin@cueb.edu.cn

Tao Ye
yetao_uibe@163.com

Jieying Gao
gaojieying@cueb.edu.cn

[1] School of Finance, Capital University of Economics and Business, Beijing 100070, China

[2] School of Banking and Finance, University of International Business and Economics, Beijing 100029, China