Computing
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

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