APPLYING MACHINE LEARNING APPROACH TO STUDY THE WILLINGNESS TO PAY FOR IRRIGATION WATER

Hoa-Thi-Thu Bui[1]

Abstract: *In order to study the water user behaviors, most of the previous studies have been using the traditional* statistical *approach, however, the development of model analytical techniques like machine learning approach which will help analysts to get more results. In this article, both the traditional statistical analysis and machine learning (ML) approach is used to study the behavior of irrigation water consumption with a logistic regression model. 235 households in Nam Dinh, Thai Nguyen, and Phu Tho province were surveyed. The research results show that the regression coefficients between the two approaches are quite similar. However, by the machine learning approach, more specific research results are analyzed such as assessing the probability of water users for increasing the water charge, and the accuracy of the models. In other words, the machine learning approach offers a prognosis better compared to the traditional statistical approach.*

Keywords: Willingness to pay, machine learning.

1. INTRODUCTION

Normally, statistical analysis is applied to find out the inference relationships between variables based on the assumptions to the population. In recent years, the development of computer science and information technology with big databases, machine learning (ML) is considered as one of the approaches which are used popularly in many activities such as healthcare, education, economic, etc. (Yadav, 2015). In this paper, a machine learning approach is applied to study clearly the behavior of water users through their willingness to pay for irrigation water fees and compare these results with the traditional statistical analysis approach. In order to understand the behavior of irrigation water users as well as finding out the factors that influence their willingness to pay, the logistic regression model is analyzed in both traditional statistical analysis and machine learning approaches. The sample of 235 households are selected in typical irrigation areas in Thai Nguyen, Phu Tho, and Nam Dinh district to find out the probability of water users to respond with water charges.

2. LITERATURE REVIEW AND METHODOLOGY

Econometric analysis is one the most common tools applied in economic analysis, aiming to find out the relationship between variables based on the given hypothesis. Nowadays technology development and big database will help the researchers to analyse and reflect problems for more detail and reality and the machine learning approach is considered to be one of the most commonly applied approaches which will be added-in economic analysis (Crane-Droesch, 2017) and gradually replacing the traditional econometric analysis methods. The terms machine learning or artificial intelligence (AI) and deep learning (DP) are often used interchangeably. ML is part of artificial intelligence, with the aim of learning from data and using statistical methods (Goodfellow et al., 2016). The ML approach seeks to find the suitability of the model by separating the existing data set into a data set for

[1]*Division of E-commerce, Thuyloi University, Vietnam*

training, validation and testing (Hastie et al, 2009). The training data set is used to estimate the model and the validation set is used to track off-sample predictive errors. The model with the lowest level of non-sample predictive error in the validation set is selected. Test set is used to evaluate the predictive error outside the sample of the selected model.

In this study, machine learning is used to analyze the extra willingness to pay (WTP) for irrigation water services with a logistic regression model. The water user's willingness is found based on the current irrigation water fees by using both traditional econometric analysis and machine learning approach with a logistic regression model. For the traditional econometric approach, the logistic model is built on survey data to find out the probability of agreeing/disagreeing to increase the current irrigation water fees as well as to find out the factors influencing their willingness to pay. Jonse Bane (2005) has used a binary model to study willingness to pay for irrigation water services in Ethiopia, using 260 randomly selected households. Research shows that the main factors influencing people's willingness to pay for this service are income, age, gender, family size, irrigation water management, quantity of irrigation water consumption, etc. Tang et al. (2013) studied the factors affecting the WTP of farmers in China on irrigation water such as age, education level, cultivated land size, size family, family income, family expenditure, etc. The logistic model has also been used by Latinopoulos (2005) to determine the factors influencing willingness to pay for irrigation water services in Greece.

For several years, new developments in machine -learning and artificial intelligence techniques hold great potential for irriga. Li & Xuewei Chao (2020) reviewed the application of machine learning approaches such as artificial neural network for classification and yield forecasting, analysis of agricultural research surveys related to the farmer behaviors could be included. Kamilaris et al. (2018) applied deep learning- one of the methods of machine learning approach to study various agricultural and food production challenges. By comparison between machine learning and other popular techniques, in respect to differences in classification or regression performance, they found that deep learning provides higher accuracy results. The willingness to pay for urban water supply was estimated under machine learning approach and traditional econometrics (multiple regression) (Malik et al. 1999). The results show that the forecasting error of the machine learning model was less than the traditional regression and found out the accuracy of the model as well.

In this study, both traditional econometric and machine learning are applied to analyze the results to find out the difference between these approaches. 235 households in 3 provinces of Nam Dinh, Thai Nguyen, and Phu Tho are surveyed randomly in typical irrigation areas such as gravity and pumping irrigation system.This study focuses on two typical irrigation systems which are gravity system located in Thai Nguyen and Phu Tho province, and the other is pumping irrigation system in Nam Dinh province.These irrigation systems are a part of the sub-systems in the Red - Thai Binh River Basin. The main reason to choose and conduct an investigation in such three provinces can be explained based on topographic characteristics. Most of the irrigation systems in Thai Nguyen province are mainly gravity systems with active in-field fee collection systems; the systems are typical pumping irrigation areas with long-term farming experience in Nam Dinh province and the other systems in Phu Tho province are combined.

The research sample was selected by random sampling method. Simple random sampling is one of the probability-based methods, and the smallest sample size must be representative of the population. Based on the opinions of experts and

experienced people in local water management authorities, a survey of 235 households was conducted to represent the populations of about 6000 households. The methodology to estimate the sample was followed by random sample selection of Smith (1986).

235 households are selected randomly in these areas to understand their WTP and awareness of irrigation water management. 85 households out of 235 households are selected from Thai Nguyen province, 101 householders from Phu Tho province, 49 households from Nam Dinh province. The sampling method is a stratified method, from the provincial level, classified according to the district and commune levels. The difference in the surveyed number of households between three provinces due to actual irrigation conditions. Irrigation systems combine gravity and pumping irrigation characteristics. Therefore, the number of surveyed households is more diverse than that of Thai Nguyen and Nam Dinh. The surveys were conducted from 2016 to 2017.

The logistic regression model is established to examine the effects of the factors on the probability of responding to the willingness to pay for irrigation water. The basic model for binary analysis is based on the utility randomization theory proposed by Hanemann (1984). Based on respondent's "Yes" or "No" responses, the probability statement about their willingness to pay for irrigation water can be estimated. Model logistic regression as below: $log(\frac{p}{1-p}) = \alpha + \beta x + \varepsilon$ or log (odd) = $\alpha + \beta x + \varepsilon$, where: p is probability of statement and $\log\left(\frac{p}{1-p}\right)$ is logit(p) (logistic) with α, β is coefficients of model, ϵ is residual with normal distribution, x is the factor effect to their response. The odd ratio can be calculated as: $odd = \frac{p}{1-p} = e^{\alpha+\beta x+\varepsilon}$

In general, we have the logistic regression model with k rish factors $x_1, x_2, \ldots x_k$ can be formulated as : $f(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$, where $z = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k$

Most of the irrigation systems in Viet Nam are funded and built by the government, farmers are mainly charged only for on-farm irrigation fees which depend on the size of their land and difference levels between provinces. These fees are established by the local community for pumping and other services in this area and calculated per area per crop. From investigation, the average on-farm irrigation fee in selected areas in Phu Tho province was about 22,000 VND/sao/crop, Thai Nguyen 17,000 VND/sao/crop and Nam Dinh was 20,000 VND/sao/crop). In order to find out the probability of willingness to pay or farmer reaction with the irrigation fee, as well as the factors affecting to WTP, the average on-farm irrigation fee was estimated (20,000 VND/sao/crop) as the baseline water fee to study the water user behaviors.

Based on the current irrigation fee payment of respondents about 20,000 VND per crop per *sao*, the respondents are asked to study their willingness to pay for increasing or decreasing compared to the current fee level. The logistic model can be estimated as

$$Ln\left[\frac{P(Y=1)}{P(Y=0)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_2 + \cdots + \beta_n X_n$$

where Y = 1 (WTP> VND 20,000 per *sao per* crop) and Y = 0 (WTP <VND 20,000 per *sao per* crop, *sao is the area unit in Northern Vietnam, 1 sao ~ 360 m^2*); Xj is the factor affecting to their responses.

This article also develops the model by using a machine learning approach to find out more results for predicting their responses compared to the results from the traditional statistical analysis. On a standard machine learning approach, the input data is typically split into training data and testing data. Normally, about 70% to 80% splitted input data is training data and the rest is test data. The first split of data is the training data , which is

the initial reserve of data used to develop the model. After model is developed based on patterns extracted from the training data with the accuracy of its predictions, this model can be tested which based on the remaining data, known as the testing data (Judith.H, et al.2018). From the survey database, input data was split into two groups as testing data (69 respondents) and training data (166 respondents) (Figure 1). R language programming software is used in this model. The results are tested with testing data to find out the accuracy and suitable model.

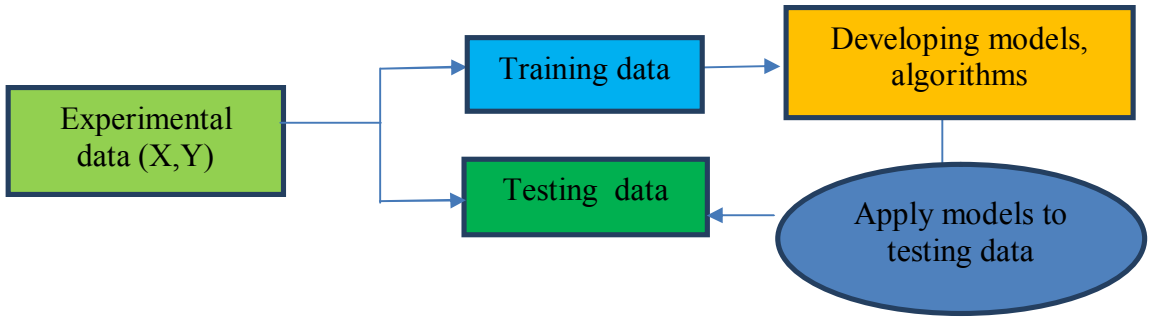

Figure 1. The analyzing procedure of machine learning approach

3. RESULTS

Most of the respondents interviewed were females which account for 57.9% (136 persons) out of 235 households, while the other (42.1%) were males. The fact that the women respondents interviewed were basically representing their husbands who were out for work during the study period. The average age of respondents was 50 years old.The occupation of these respondents is mainly farmers. Most of the land is used for paddy cultivation during two main seasons: Winter – Spring and Summer-Autumn. During winter, maize and vegetables are planted in certain areas.

The average cultivated area is about 5 *sao* per household. In order to support the farmers, Vietnam's government has decided to exempt irrigation water fees for them. In 2007, Decree No.154/2007/ND-CP, issued by the Vietnamese government to exempt irrigation fees was enacted from January 1st, 2008 nationwide. It is noted that farmers have not had to pay irrigation fees for the limited land area used for agriculture. Although the irrigation water fee exemption policy has been applied, the farmers still have to pay an on-farm irrigation fee for pumping water and other services. This fee may vary by region. However, to ensure a more sufficient and timely water supply, people are still willing to contribute to improving the service and quality of water for cultivation. The irrigation water fee policy has been changed several times since 2007. Although the government has adjusted the price calculation method in order to reduce the cost burden for farmers, the policy of irrigation water fee exemption has caused a lot of controversy from many different perspectives. In order to manage water resources efficiently, pricing for irrigation water is proposed in the new Law on Irrigation Passed on June 19, 2017. Although water pricing is considered as one of the demand approaches by many countries in the world, it has not been easy to implement in Vietnam at this moment. However, understanding the farmer's willingness to pay is really important to provide more information for the policymaker to have road maps to achieve the proposed objectives in managing irrigation efficiently. In this paper, the willingness to pay is studied based on the understanding of their probability to agree or disagree to pay more extra fees for irrigation water.

Spring- Winter and summer-autumn crops are

the two main crops in the farming year in the Northern Delta. However, due to weather conditions having rain in the Summer-Autumn crop, the amount of irrigation water is quite small compared to the Spring- Winter crop. Therefore, the criterias in the winter-spring crop are selected to reflect water- related behaviors of farmers clearly.

In order to find out in more detail the factors affecting user's WTP such as age, sex, area for paddy cultivation, yield, etc, regression models are also applied, using R language for testing the defects of the models. The testing results show that the ***AREA_AGR*** and ***YIELD_SW*** variable are the main factors affect to the probability of WTP wit the estimated coefficient of the model is statistically significant. Based on the investigated data and using the traditional analysis, the binary logistic regression equation is established as below:

$$LnY = -5.17 + 0.1414\,AREA\_AGR + 0.025 YIELD\_SW$$

where Y is the dummy variable, Y= 1 (answering WTP > 20,000 VND/*sao*/crop) and Y = 0 (answering WTP <20,000 VND/*sao*/crop); ***AREA_AGR*** is the area for paddy cultivation and ***YIELD_SW*** is yield of Spring- Winter crop.

In this study, two independent variables are the area for paddy cultivation (***AREA_AGR***) and the yield of Spring- Winter crop (***YIELD_SW***) as the main impact on the probability answering yes/no for increasing irrigation water fees. According to the traditional econometric approach, the results show that the statistical values are significant with the p_values less than 0.05. The logit regression equation represents the odds of WTP will increase by 0.1414 times if the area increases by one unit in case the Spring- Winter yield (***YIELD_SW***) remains constant. The odds of WTP will increase 0.025 times if ***AREA_AGR*** increases by one unit in case of ***YIELD_SW*** does not change. The area for paddy cultivation (***AREA_AGR***) is factor impact to WTP clearer than ***YIELD_SW***.

From the Machine Learning (ML) approach, the collected data is splitted into training data which accounts for 70% of the data (166 respondents) and 30% of the data (69 respondents) used for testing data. The binary logistic regression equation based on the training data set is also estimated by using R language. Logistic regression results from the training data show that the independent variables ***AREA_AGR*** and ***YIELD_SW*** are statistically significant. The coefficients of this model have slightly changed compared to the model above. Specifically, the odds of WTP will increase 0.14458 times if the planted area increases by one unit in case the Spring- Winter yield (***YIELD_SW***) does not change. Odds of WTP will increase 0.0222 times if ***YIELD_SW*** increases by one in case the planted area remains constant. It can be concluded initially that the relationships between the variables as well as the values of the coefficients in both traditional econometric and the machine learning approach are mainly similar. However, there is more information from the model that can be analyzed by the machine learning approach, especially the predictability of the model which the traditional analysis has not mentioned yet. It can be explained by checking the accuracy of the model through testing procedures with the testing data set as mentioned above. According to the machine learning approach, results from training models are tested with a testing data set to find out the accuracy of the model and forecasting other results. The prediction results that determine the probability of answering Yes/No for each respondent are shown as below.

Figure 2 shows the probability of agreeing and disagreeing with an increase in irrigation water fee for each respondent. For example, the probability of the 4th respondent (coded by ID4) agreeing to increase irrigation water fee is 79.04% and disagree is 20.96%. Similar to other households, the model shows a prediction probability to answer 'Yes' or 'No' to pay more irrigation water fees compared to the current.

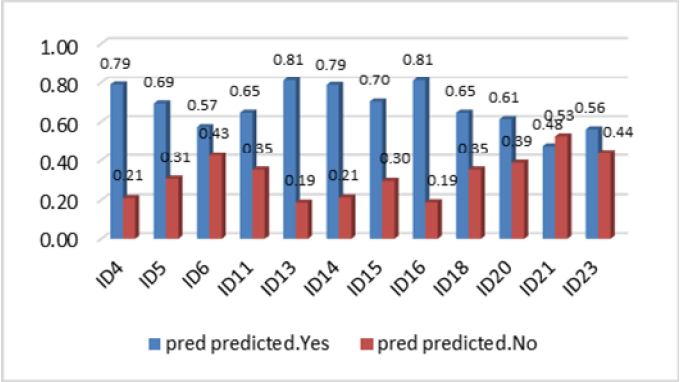

Figure 2. Predicting the probability of respondent answer 'Yes/ No' to improve irrigation water fee

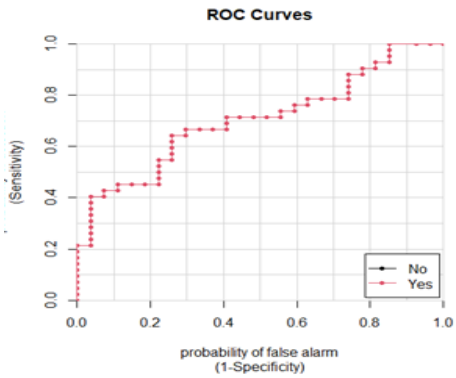

Figure 3. The relationship between sensitivity and specificity (ROC curve)

To evaluate the accuracy of the model, the AUC index is determined based on the results of the predictive model and testing data. The predictive model's accuracy is assessed based on sensitivity and specificity, as well as the AUC index. The sensitivity index of the model is 0.786 (78.6%), which represents the probability that respondents are willing to pay more than the current irrigation fee in both the prognostic and actual model. The specificity index represents the probability of disagreements in both the prognostic model and the actual data, and this index is estimated to be 0.33 (33%). These two indicators show the degree of compatibility between the prognostic model and the actual model, the majority of respondents are willing to pay more irrigation water fees compared to the current charges. The accuracy of the model is calculated at about 60.87% and the AUC index is 0.746 (74.6%). That means the probability of agreeableness to increase irrigation water fee is higher than the disagreement level of 74.6%. The higher the AUC level is, the higher accuracy of the model is. In addition, the ROC (Receiver Operating Characteristic) diagram also clearly shows the relationship between sensitivity and specificity as the figure below (Figure 3).

The application of machine - learning approaches in understanding the consensus to pay more for irrigation water fees has shown more detailed results compared to the traditional logistic regression model approach. Besides the factors affecting willingness to pay, the model also shows the probability of responding to agree/disagree for each respondent with the accuracy of the model as well.

From the results of the model, it can be seen that the machine learning has more advantages and different objectives compared to the traditional analytical approach. Traditional econometrics approaches are usually interested in obtaining reliable estimates of marginal effects, especially to find out or estimates of the coefficients. In contrast, machine learning approaches are intended for prediction tasks with the aim to obtain accurate predictions. The predictive ability of machine learning in complex and high-dimensional settings can also be used to improve causal estimates. Compared to the traditional analytics approach, the results from the machine learning show clearly the accuracy of the model, the prediction of probability for each respondent to respond with the current water fee. These results are useful for the analysers to have multi-dimension and objective views for prediction and forecasting.

4. CONCLUSION

Nowadays high technology development, applying the machine learning approach in economic analysis is widely applied with the high accuracy level and multidimensional analysis. In this study, two analytical approaches are the

traditional statistics and machine learning in order to find out the willingness to pay farmers for the irrigation water fee. This study is surveyed in three provinces of Thai Nguyen, Phu Tho, and Nam Dinh with a sample number of 235 households. The results show that the area for paddy cultivation (*AREA_AGR*) and yield of Spring- Winter crop (*YIELD_SW*) are the main factors influencing the willingness to pay irrigation water fees. The logistic regression results are similar in both approaches. The machine learning approach also offers logistic regression results similar to traditional logistic regression results. However, the model's accuracy and the probability of agreeing/disagreeing for each individual respondent are shown clearly with the machine learning approach. The establishment of the model or choosing approach for analyzing the willingness to pay for water fees is considered useful in providing insights to build up pricing mechanisms in water resources. Therefore, the results will provide more information to policy makers and water managers to improve the planning and management efficiency of water resources in Vietnam and design the suitable and acceptable water price for users.

REFERENCES

Crane-Droesch, A. (2017). *Technology diffusion, outcome variability, and social learning: evidence from a field experiment in Kenya.* American Journal of Agricultural Economics 100: 955–974.

Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y. (2016). *Deep Learning*. Cambridge: MIT press

Hanemann, M.W. (1984), *'Welfare Evaluations in Contingent Valuation Experiments with Discrete Response Data'.* American Journal of Agricultural Economics 66(3):332-341

Hastie. T, Tibshirani.R & Friedman.J (2009), *The elements of statistical learing: Data mining,* Inference and Prediction, 2rd. ed, Springer.

Jonse Bane (2005): *Valuing Non-Agricultural use of Irrigation Water*, Evidence from the Abbay River Basin of Amhara Regional State, MSc thesis AAU, Ethiopia.

Judith Hurwitz and Daniel Kirsch (2018). *Machine learning for dummies.* John Wiley &Sons, Inc.

Kamilar Andres & Francesc X. Prenafeta-Boldu (2018). *Deep learning in argriculture: A survey. Computers and Electronics in Agriculture*. Volume 147. Pages 70-90.

Latinopoulos P. (2005): *Valuation and Pricing of Irrigation Water: An Analysis in Greek Agricultural Areas,* Global NEST Journal, Vol 7, No 3, pp 323-335, 2005, Greece

Malik Ranasinghe , Goh Bee-Hua & T. Barathithasan (1999). *Estimating willingness to pay for urban water supply: a comparison of artificial neural networks and multiple regression analysis.* Impact Assessment and Project Appraisal.SSN: 1461-5517 (Print) 1471-5465.

Smith, M. J. (1986). *Contemporary communication research methods.* Belmout, CA: Wadsworth Publishing, pg 223- 225.

Tang.Z, Nan.Z,Liu.J (2013) *"The willingness to pay for irrigation water: a case study in Northwest China*. Global NEST Journal, Vol 15, No 1 , pp 76-84.

Yadav, S. K. (2015). *Sentiment analysis and classification: A survey.*

Yadav, S. K. (2015). *Sentiment analysis and classification: A survey.* International Journal of Advance Research in Computer Science and Management Studies, 3(3), 113–121.

Yang Li & Xuewei Chao (2020). *ANN-Based Continual Classification in Agriculture.* Journal Agriculture 2020, 10, 178; doi:10.3390/agriculture10050178.

Tóm tắt:

ỨNG DỤNG CÁCH TIẾP CẬN HỌC MÁY ĐỂ NGHIÊN CỨU Ý MUỐN THANH TOÁN ĐỐI VỚI DỊCH VỤ CUNG CẤP NƯỚC TƯỚI

Để tìm hiểu hành vi người tiêu dùng nước, những nghiên cứu trước đây thường sử dụng cách phân tích thống kê truyền thống. Tuy nhiên với sự phát triển các kỹ thuật phân tích hiện đại ngày nay như máy học sẽ giúp các nhà phân tích có được nhiều kết quả đáng mong đợi hơn.Trong bài viết này tác giả sử dụng cả hai tiếp cận phân tích thống kê truyền thống và tiếp cận học máy (Machine Learning-ML) để nghiên cứu hành vi tiêu dùng nước tưới thông qua phản ứng của người dân khi phí nước tưới thay đổi thông qua mô hình hồi quy logistic. Nghiên cứu được tiến hành điều tra 235 hộ tại 3 tỉnh Nam Định, Thái Nguyên, Phú Thọ. Kết quả nghiên cứu cho thấy các hệ số hồi quy giữa hai cách tiếp cận khá tương đồng. Tuy nhiên với cách tiếp cận máy học sẽ đưa ra nhiều kết quả nghiên cứu cụ thể hơn như đánh giá xác suất đồng ý tăng phí nước, tính chính xác của mô hình, hay nói cách khác tiếp cận máy học đưa ra khả năng tiên đoán hơn so với tiếp cận thống kê truyền thống.

Từ khóa: Ý muốn thanh toán, học máy

Ngày nhận bài: 24/7/2021

Ngày chấp nhận đăng: 29/8/2021