Prediction of Customer Behavior using Machine Learning: A Case Study*

Tran Duc Quynh** and Hoang Thi Thuy Dung

International School-Vietnam National University, Hanoi, Vietnam
ducquynh@vnu.edu.vn
17071347@isvnu.vn

Abstract. Understanding what customers want and need- and, ideally, anticipating their needs - is a constant challenge for marketers. With the advent of machine learning, researchers have successfully it for some problems in customer behavior analysis. For a specific problem, we need to study machine learning models and data preprocessing techniques to get the highest accurate solution. In this paper, we consider the problem and the dataset given in Wang et al. (2017). The task is to predict the decision of customers with the question of whether they accept a restaurant coupon recommended by an in vehicle system, given a set of input about customer's driving context. Most of attributes being categorical and missing values makes the problem more difficult. We investigated methods to transform categorical attributes to numerical attributes, handle the missing values and then applied various classification models. The result showed that the proposed approach overperforms the previous methods given in Wang et al. (2017). Besides, we also obtain some interesting findings about the used methods and the impact of variables on the customer decision.

Keywords: Customer Behavior, Classification Models, Bagging

1 Introduction

Customer behavior analysis is very important for an enterprise. It helps companies understand what the customer wants and needs. Hence, the company can improve the service or offer a suitable product to the customer. Thanks to customer behavior analysis, the company can increase their sales and be more successful in business. Nowadays, digital transformation affects all activities of enterprises. Data of products and customers can be collected via information systems. These data may be used to understand more deeply about the customer's intention by using tools in data science.

** Corresponding author.

In recent years, the use of machine learning methods to study problems in customer behavior analysis has become more and more attractive. The problems can be considered as supervised or unsupervised models in machine learning. The used models may be clustering models, regression models or classification models, such as decision tree, random forest, support vector machine, neural network, logistic regression,…Safia et al. (2015) used decision trees to find a number of important controllable characteristics: interactions, playtime, location and understand what influences a customer's purchase choice. Larivière, B., and Van den Poel, D. (2005) proposed a method based on random forest for solving the customer retention prediction problem. In 2020, V. Shrirame, J. Sabade, H. Soneta and M. Vijayalakshmi (2020) employs data visualization, natural language processing and machine learning to explore the demographic of an organization. Dou, X (2020) used an ensemble learning to predict online purchase behavior of customers. In 2017, Wang et al. proposed a Bayesian framework for learning rule sets to solve a classification problem and application for predicting the customer intention in in-vehicle recommendation systems. S. Cao et al. (2019) proposed a deep learning model to analyze customer churn.

Although machine learning is promising for solving customer behavior prediction problems, the number of research in this field is limited. Besides, we need to preprocess data before applying a machine learning model. Moreover, there does not exist a good method for every dataset. Hence, the study of preprocessing methods and machine learning models to improve the results for a specific dataset is necessary. In this paper, we consider the problem of customer intention prediction and the data given in Wang et al. (2017). Our approach is to tackle directly the original dataset instead of dividing it in 5 folders. We also proposed two methods to handle missing values and then used some machine learning models for the resulting dataset. The comparison based on AUC showed that our approach is better than the methods given in Wang et al. (2017). The performance of machine learning models was explored to find the best models and then detected important features that have strong effects on the customer decision.

The paper is structured as follows. Section 2 introduces the problem and the dataset used in this research. Section 3 is reserved to present the methodology while results are reported in Section 4. Conclusion is given in Section 5.

2 Problem Statement and Data

2.1 Problem Statement

The context of enhancing the prediction of customer behavior has been broadened in many aspects and is going further with the help of advanced technology. However, there is still less research about specific solutions for business when they consider which to choose in order to solve their problems among a wide world of machine learning. This study is a new approach that covers a variety of prediction models and improves their accuracy. Hence, the marketers will have a large and clear vision on how machine learning can help them and have more selection on applying machine learning models in developing their businesses. We investigate the data collected from a system which is put on the car and makes recommendations on the coupon from local businesses. As the output data is binary of whether the driver accepts the coupon,

classification models would be chosen to make analyses and predictions like decision tree, random forest, support vector machine, etc. The data is also experimented through several models of preprocessing to improve performance as well as predicting with calculated accuracy. The results then are compared to generate the best model that would be helpful in predicting the customer behavior. Moreover, the importance of each feature could be indicated to take a deeper understanding aiming to advance the overall system.

2.2 Data

In-vehicle coupon recommendation dataset which contains 23 attributes with 12.684 records was gathered using an Amazon Mechanical Turk poll published at UCI website in 2017 https://archive.ics.uci.edu/ml/datasets/in-vehicle+coupon+recommendation. The survey goes over various driving scenarios, such as the destination, present time, weather, passenger, and so on, before asking the person if he will accept the voucher if he is the driver.

The goal of the prediction problem is to forecast whether a client will accept a coupon for a specific venue based on demographic and contextual factors. Replies that the user will drive there "right away" or "later before the coupon expires" are labeled "Y = 1", whereas answers of rejecting the coupons are labeled "Y = 0". We are looking into five different categories of coupons: pubs, takeaway food restaurants, coffee shops, low-cost restaurants (under $20 per person), and high-cost restaurants (between $20 and $50 per person). The difficulty of this problem come from the attributes. Most of attributes (19/ 23 attributes) are categorical features and there are five attributes containing missing values.

3 Methodology

3.1 Preprocessing data

From having an overview over the dataset, we notice that there are a number of missing values and the data is quite balanced with the acceptance percentage being approximately 50%. Furthermore, to experiment with the models, the data must be numerical type. Hence, we propose several methods to handle those problems.

To transfer variables from categorical to numerical type, we use the combination of two methods-integer encoding (mapping) and one-hot encoding (get_dummies function). Mapping is used for ordinal attributes such as age, time, expiration, income, etc. while get_dummmies is used for nominal attributes. Dealing with missing values, we use 2 methods which are imputing by mode and imputing by random forest. Using mode imputing, the missing values are replaced by the mode value in the attribute range while using random forest imputing, the missing values are replaced by prediction values using random forest classification model. Besides, we also add scale to the range of value to see if it can enhance the model performance. The method we use is min-max scaling, which consists in rescaling the range of features to scale the range in [0,

1]. With the mix of above mentioned methods, we create 4 sub-datasets described in table 1.

Table 1. Building 4 sub-datasets.

Sub-datasets	Mode imputing	Random forest imputing
No scale	d1	d3
Scale	d2	d4

3.2 Machine Learning Method

In this research, we use classification methods of machine learning. The task of approximating the mapping function from input variables to discrete output variables is classified predictive modeling. The basic goal is to figure out which category or class the new data belongs to. For the experiments, we adopted several different classification approaches that were selected due to their extensive use, well-understood behavior, and promising results in a range of categorization tasks. Our goal was to examine classifiers that differ in terms of the functional forms of classification boundaries they may learn, as well as classifiers that are based on distinct assumptions about the relationship between distinct features. We studied the performance of a set of classification models including decision tree, random forest, support vector machine (SVM), feedforward neural networks (MLP), logistics regression, Bagging, AdaBoost, XGBoost.

3.3 Estimating Model Performance Measurement

We use the k-fold cross-validation approach with 3 measurements: accuracy, f1 score, and AUC (the Area Under The ROC Curve). Accuracy is the simplest intuitive performance metric. It is just the ratio of properly predicted observations to all observations. F1 score is the weighted average of precision and recall. As a result, this score considers both false positives and false negatives. Area Under Curve (AUC) score represents the degree or measure of separability.

4 Results and Evaluation

All accuracy results are shown in Table 2 below:

Table 2. Performance results.

		Dataset			
Model	**Measure**	**d1**	**d2**	**d3**	**d4**
Decision	acc	0.69	0.69	0.69	0.69

tree	f1	0.73	0.73	0.73	0.73
	auc	0.68	0.68	0.68	0.68
Random forest	acc	**0.76**	**0.76**	0.75	0.75
	f1	0.79	0.79	0.79	0.79
	auc	**0.83**	**0.83**	**0.83**	**0.83**
Logistic regression	acc	0.69	0.69	0.69	0.68
	f1	0.74	0.73	0.74	0.73
	auc	0.74	0.74	0.74	0.74
SVC	acc	0.69	0.69	0.69	0.69
	f1	0.74	0.74	0.74	0.74
	auc	0.74	0.74	0.74	0.74
MLP	acc	0.74	0.74	0.74	0.75
	f1	0.78	0.78	0.78	0.78
	auc	0.81	0.81	0.81	0.81
Bagging	acc	**0.76**	**0.76**	**0.76**	**0.76**
	f1	**0.80**	**0.80**	**0.80**	**0.80**
	auc	**0.83**	**0.83**	**0.83**	**0.83**
Adaboost	acc	0.68	0.68	0.68	0.68
	f1	0.73	0.73	0.73	0.73
	auc	0.74	0.74	0.74	0.74
XGBoost	acc	0.72	0.72	0.72	0.72
	f1	0.77	0.77	0.77	0.77
	auc	0.79	0.79	0.79	0.79

The accuracy results range is from 68% to 76%. They are quite equal among 4 datasets. Therefore, the type of imputing missing value and the scale has less effect on accuracy of predicting values.

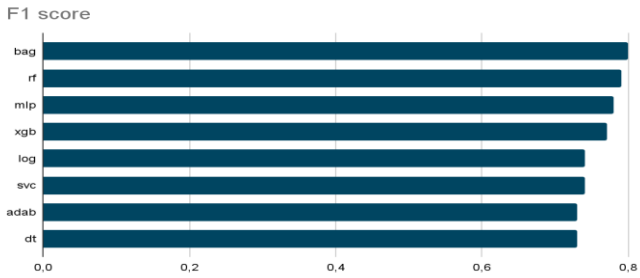

Fig. 1. F1 score of models.

From figure 1, we can see that the f1 score of 8 models is quite close to each other with the values of greater than 70%. The bagging model has the highest accuracy with 80%. The second highest rank model is random forest. Then that is the demonstration for the help of bagging in enhancing the performance of random forest classification.

Putting in comparison with the most related work done with the same dataset of Wang Tong and partners in Wang et al.(2017), we found that we have reached a better result with 83% of the bagging model while their result is approximately 73% with Bayesian rule sets in using the same accuracy measurement - AUC. So, this study can be considered as an enhancement in the accuracy of prediction for marketers. As results, we suggest using ensemble learning methods as Bagging for the dataset and for customer's purchasing intention prediction problem in gerenal.

For deeper understanding about the data and the scenario, we use the best model - bagging with random forest based to indicate the importance of each feature (see Fig. 2).

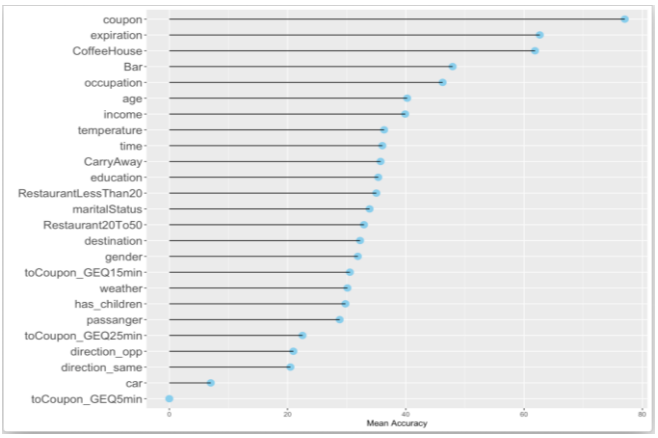

Fig. 2. Feature importance.

Since *coupon* is the type of coupon that drivers were recommended by the system, then it is the most important feature. Two other important features can be mentioned are *expiration*, and *CoffeeHouse*. *car* and *toCoupon_GEQ5* are two features having the least importance so they should be removed from the model. We notice that with the five types of coupons (bar, coffee house, take away, expensive restaurant, and less expensive one), the coffee house coupon has the most attribution to the results that means customers tend to accept that kind of coupon. The others in descending order are bar, take away, less expensive restaurant, and then the expensive one.

5 Conclusion

All above results bring us to a conclusion about the best model among 8 ones that fits the given dataset of in-vehicle coupon recommendation systems. Before going to implement models for analyzing and predicting values, the dataset should be preprocessed by filling missing values by mode or random forest (as suggested), for addition, it is not necessary to be scaled. The model should be used to give the best results performance is bagging classification with random forest the based estimator.

The challenge of forecasting consumer purchasing decisions using easily quantifiable aspects of the purchasing context was investigated in this research. The findings and discussion showed that the proposed methods are effective for the given dataset (and promising for the problem of customer's intention prediction in general).

This research still has its limitations of only covering several models in a numerous and diverse world of machine learning. However, it can be a source to bring the idea for further creative research along with the upward ever-changing trend in the demand and coordination of technology with other industries.

References

1. Asghar, N. 2016. Yelp dataset challenge: Review rating prediction. arXiv preprint arXiv:1605.05362.
2. Buckinx, W.; Verstraeten, G.; and Van den Poel, D. 2007. Predicting customer loyalty using the internal transactional database. Expert Systems with Applications 32(1):125–134.
3. Ding, Y., DeSarbo, W. S., Hanssens, D. M., Jedidi, K., Lynch Jr., J. G., & Lehmann, D. R. (2020). The past, present, and future of measurements and methods in marketing analysis. Marketing Letters, https://doi.org/10.1007/s11002-020-09527-7.
4. Dou, X. (2020). Online purchase behavior prediction and analysis using ensemble learning. In 2020 IEEE 5th International conference on cloud computing and big data analytics, ICCCBDA 2020 (pp. 532–536). https://doi.org/10.1109/icccbda49378.2020.9095554.
5. Kaefer, F.; Heilman, C. M.; and Ramenofsky, S. D. 2005. A neural network application to consumer classification to improve the timing of direct marketing activities. Computers & Operations Research 32(10):2595–2615.
6. Ladas, A.; Garibaldi, J.; Scarpel, R.; and Aickelin, U. 2014. Augmented neural networks for modelling consumer indebtedness (sic). In Proc. IEEE International Joint Conference on Neural Networks 3086–3093.
7. Larivière, B., and Van den Poel, D. 2005. Predicting customer retention and profitability by using random forests and regression forests techniques. Expert Systems with Applications 29(2):472–484.

8. S. Cao, W. Liu, Y. Chen and X. Zhu, "Deep Learning Based Customer Churn Analysis," 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP), 2019, pp. 1-6, doi: 10.1109/WCSP.2019.8927877.
9. Sifa, R.; Hadiji, F.; Runge, J.; Drachen, A.; Kersting, K.; and Bauckhage, C. 2015. Predicting purchase decisions in mobile free-to-play games. In Proc. Conference on Artificial Intelligence and Interactive Digital Entertainment.
10. V. Shrirame, J. Sabade, H. Soneta and M. Vijayalakshmi, "Consumer Behavior Analytics using Machine Learning Algorithms," 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 2020, pp. 1-6, doi: 10.1109/CONECCT50063.2020.9198562.
11. Wang, Tong, Cynthia Rudin, Finale Doshi-Velez, Yimin Liu, Erica Klampfl, and Perry MacNeille. 'A bayesian framework for learning rule sets for interpretable classification.' The Journal of Machine Learning Research 18, no. 1 (2017): 2357-2393.
12. Xie, Y.; Li, X.; Ngai, E.; and Ying, W. 2009. Customer churn prediction using improved balanced random forests, Expert Systems with Applications 36(3):5445–5449.