Journal of Actuarial, Finance and Risk Management (JAFRM)
Vol. 1(1), June, 2022.

Online Shopping Website Analysis for Marketing Strategy Using Clickstream Data and Extra Trees Classifier Algorithm

Diah Prastiwi[1*]

[1] *Study Program of Informatics Technology, Gunadarma University, 16310, Indonesia*
**Corresponding author: diahp@staff.gunadarma.ac.id*

Abstract— On an online shopping website, the platform may provide a service to the shop owners by suggesting which items to promote. One possible consideration is price. If an item is priced more expensively than the average price of other items in the same category, then the item should be advertised more intensely, or repriced. Due to the quickly growing number of products and categories, calculating the average price in real time can be difficult or slow. Alternatively, one may employ machine learning algorithms. In this study, we use Extra Trees Classifier on clickstream data, which is user activity report. We demonstrate the algorithm on the clickstream data of a an online shopping website for pregnant women, retrieved from UCI Machine Learning Repository Dataset. The data has 14 attributes and 165474 entries. The model is trained on 75% of the data, and tested on the remaining 25%, with an observed accuracy of 99 %.

Keywords— Clickstream Data; Extra Trees Classifier; Online Shopping Website.

I. INTRODUCTION

After transitioning into a globally interconnected network, the internet has emerged as a marketing tool for both domestic and international trade. Online shopping is a recent phenomenon that is rapidly developing and still has a huge market potential. The convenience of online shopping is one factor causing its quick adoption among the consumers, especially Gen Y and its successors. [10]

Advertising correctly is important to ensure that sales goal is achieved, especially for the least popular items. Price is a significant factor for many customers. The more expensive products should be repriced, or else should be advertised more intensely. To decide when a product is "expensive", one may look at the average price for other products on the same category. However, a product may belong to many categories, and it may be difficult to obtain a complete price data for the numerous other products. Therefore, it is desirable to have a procedure to predict whether a product's price is higher than average, by simply looking at the product's attributes. In this study we discuss one way to achieve that, by using a machine learning algorithm on user activity data from the online shopping website.

Clickstream analysis is the process of data gathering, analysis, and reporting on user activities in a website. Clickstream data (user activity report) is usually stored on a web server as an access log file including IP address, reference page, and visiting time [15]. In this study, the clickstream data was retrieved from UCI Machine Learning Repository Dataset, already in csv format, of an online shopping website for pregnant women's clothing [5]. The data has 14 attributes and 165474 entries. Among the data attributes, the independent variables are: year, month, day, order, country, season ID, page 1 (main product categories: trousers, skirts, blouses, sale), page 2 (product codes for each of the 217 products), color, photo location (where the product's photo is located on the page as the screen is divided into six parts), model photography (1-en face, 2-profile), price, and page (the page number within the online shopping website / from 1 to 5). The independent variable is price 2, a variable informing whether the price of a certain product is higher than the average price of the product's categories, either 1-yes or 2-no.

The data undergoes pre-processing to search for missing values [12], followed by exploratory data analysis (EDA) [4] including visualization and several things of concern such as which countries have many users visiting the website and purchasing a product, and which location and placement of a product's photo on the website are clicked more often. The main topic of analysis is the prediction of which product is priced higher than average, using a classification algorithm known as Extra Trees Classifier [6]. We will also consider which of the independent variables have more impact on the dependent variable. The data will be split into training data (75%) and testing data (25%) [13]. The training data is used to train the machine learning model, while the testing data is used to test the performance and accuracy of the model [3]. Finally, the model's performance will be evaluated [1]. The main algorithm in this study, Extra Trees Classifier, is an ensemble learning method that is fundamentally based on decision tree [8].

II. METHOD

The Extra-Trees algorithm builds an ensemble of unpruned decision or regression trees according to the classical top-down procedure. Its two main differences with other tree-based ensemble methods are that it splits nodes by choosing cut-points fully at random and that it uses the whole learning sample (rather than a bootstrap replica) to grow the trees [6].

TABLE 1
EXTRA-TREES SPLITTING ALGORITHM (FOR NUMERICAL ATTRIBUTES)

Split_a_node(S)
Input : the local learning subset S corresponding to the node we want to split
Output : a split$[a < a_c]$ or nothing
- If **Stop_split**(S) is TRUE then return nothing.
- Otherwise select K attributes $\{a_1, \dots, a_K\}$ among all non constant (in S) candidate attributes;
- Draw K splits $\{s_1, \dots, s_K\}$, where $s_i =$ **Pick_a_random_split**$(S, a_i), \forall i = 1, \dots, K$;
- Return a split s_* such that score $(s_*, S) = \max_{i=1,\dots,K} score(s_i, S)$

Pick_a_random_split (S, a)
Inputs : a subset S and an attribute a
Outputs : a split
- Let a^S_{max} and a^S_{min} denote the maximal and minimal value of a in S;
- Draw a random cut-point a_c uniformly in $\left[a^S_{min},\ a^S_{max}\right]$;
- Return the split $[a < a_c]$

Stop_split(S)
Input : a subset S
Output : a boolean
- If $|S| < n_{\min,}$ then return TRUE ;
- If all attributes are constant in S, then return TRUE;
- If the output is constant in S, then return TRUE;
- Otherwise, return FALSE

The Extra-Trees splitting procedure for numerical attributes is given in Table 1. It has two parameters: K, the number of attributes randomly selected at each node and $n_{min,}$, the minimum sample size for splitting a node. It is used several times with the (full) original learning sample to generate an ensemble model (we denote by M the number of trees of this ensemble). The predictions of the trees are aggregated to yield the final prediction, by majority vote in classification problems and arithmetic average in regression problems. From the bias-variance point of view, the rationale behind the Extra-Trees method is that the explicit randomization of the cut-point and attribute combined with ensemble averaging should be able to reduce variance more strongly than the weaker randomization schemes used by other methods. The usage of the full original learning sample rather than bootstrap replicas is motivated in order to minimize bias. From the computational point of view, the complexity of the tree growing procedure is, assuming balanced trees, on the order of N log N with respect to learning sample size, like most other tree growing procedures. However, given the simplicity of the node splitting procedure we expect the constant factor to be much smaller than in other ensemble based methods which locally optimize cut-points. The parameters K, $n_{min,}$ and M have different effects: K determines the strength of the attribute selection process, $n_{min,}$ the strength of averaging output noise, and M the strength of the variance reduction of the ensemble model aggregation. These parameters could be adapted to the problem specifics in a manual or an automatic way (e.g. by cross-validation) [6]. However, we prefer to use default settings for them in order to maximize the computational advantages and autonomy of the method [6].

III. APPLICATION AND RESULTS

A. Dataset Loading

The dataset was retrieved from [5]. In this study, we use Python (Google Colab). First, standard libraries such as pandas, numpy, seaborn, matplotlib, and sklearn are called. Then the dataset is loaded. The dataframe has 14 attributes (columns) and 165474 entries (rows). We assign all attributes except PRICE 2 as the independent variables, and we assign PRICE 2 as the dependent variable.

B. Data Preprocessing

At this stage, first we look for missing values on the data. It turns out that there are no missing values. So we look for the presence of categorical data, which cannot be directly processed by the machine learning algorithm and must be converted into numerical data. There is an attribute (page 2 - clothing model) that is categorical and we encode it as numerical data. Figure 1 shows the data types, and Figure 2 shows that there are no missing values.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 165474 entries, 0 to 165473
Data columns (total 14 columns):
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   year                     165474 non-null  int64
 1   month                    165474 non-null  int64
 2   day                      165474 non-null  int64
 3   order                    165474 non-null  int64
 4   country                  165474 non-null  int64
 5   session ID               165474 non-null  int64
 6   page 1 (main category)   165474 non-null  int64
 7   page 2 (clothing model)  165474 non-null  object
 8   colour                   165474 non-null  int64
 9   location                 165474 non-null  int64
 10  model photography        165474 non-null  int64
 11  price                    165474 non-null  int64
 12  price 2                  165474 non-null  int64
 13  page                     165474 non-null  int64
dtypes: int64(13), object(1)
memory usage: 17.7+ MB
```

Figure. 1 Data types of all the attributes in the dataset.

```
year                     0
month                    0
day                      0
order                    0
country                  0
session ID               0
page 1 (main category)   0
page 2 (clothing model)  0
colour                   0
location                 0
model photography        0
price                    0
price 2                  0
page                     0
dtype: int64
```

Figure. 2 Checking for missing values.

C. Exploratory Data Analysis (EDA)

From the data of the dependent variable (“price 2”), 84695 items are priced higher than the average of all items (across categories), while 80779 items are priced lower than the average of all items. The country with the greatest number of buyer is Poland. The location of placement of the product’s photo on the website that is clicked most often is “top left”. By plotting the data of “page 2 (main category)” against click count, we observe that trousers are clicked most often (see Figure 3). By plotting “page 2 (main category)” against “price”, we observe that skirts has the highest price followed by trousers.

Figure 4 shows the heat map visualization, where the correlation between each pair of variables is shown. We observe that “price” has the strongest influence on “price 2”, which is clear. Some independent variables have weak correlation with “price 2”, such as “location”, “color”, and “page 2 (main category)”. Because the attribute “price” has a strong correlation with “price 2”, the outlies in “price” are removed.

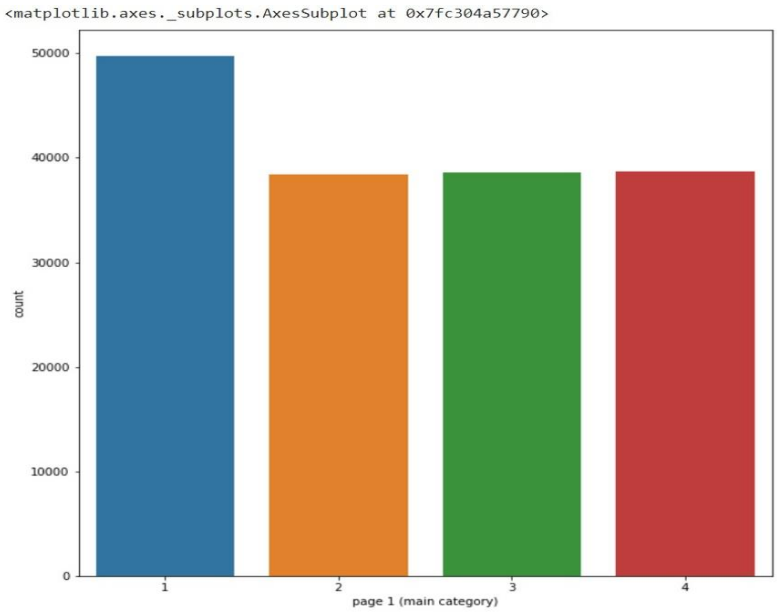

Figure. 3 Visualizing “page 1 (main category)” vs click count.

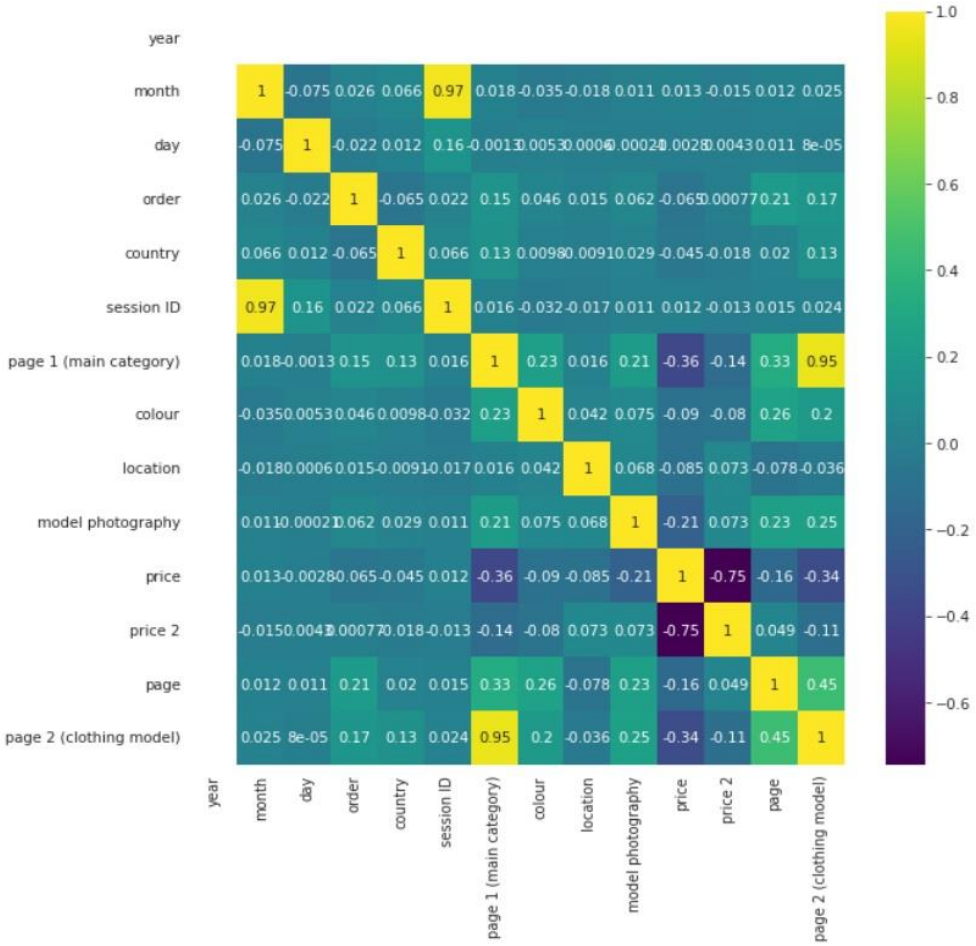

Figure. 4 Heat map visualization, correlation between every pair of attributes.

From these observations, we may give some recommendations: the more expensive items in the “trousers” category should be promoted/advertised or repriced because there is a high interest already. Some promotion such as year-end or other special discounts should be given more to Polish customers, as Poland is the biggest customer base. As for the other countries with low visitor count, such as UK and Germany, the shop owners or even the platform managers should invest in promoting their products on other platforms or social media accounts that are

visited more often by people from those countries, through integrated digital marketing such as product placement. Items that are currently priced lower than average should be leveraged by placing the items on the page location with the highest number of clicks (“top-left”)

D. Classifier Building dan Model Evaluation

After the data is cleaned and ready for classification, we separate the attributes into X (independent variables) and Y (dependent variable) [9]. As we mentioned earlier, only “price 2” is considered as the dependent variable. To ensure that no single attribute dominates the other attributes, we perform a feature scaling so that all numerical data on the dataset have the same interval (scale) [11]. Next we split the data into training data (75%) and testing data (25%). The split is performed randomly. The training data is used to create a machine learning model by the Extra Trees Classifier algorithm. To evaluate the model, we compute the Confusion Matrix (or Error Matrix) by counting TP (true positives), TN (true negatives), FP (false positives), and FN (false negatives) [14]. Then we compute some parameters of the Confusion Matrix: accuracy (the ratio of correct predictions over all predictions), precision (the ratio of true positives over all positive predictions), recall (the ratio of true positives over all true data), and F1-score (the harmonic mean of precision and recall) [7]. From the model, we obtain TP=62165 , TN = 6e+04 , FP = 1327 , FN = 377 . The precision is 98% for Yes, and 99% for No. The recall is 99% for Yes, and 98% for No. The F1-score is 99% for Yes and No. See Figure 5.

Classification Report for Extra Trees Classifier :

	precision	recall	f1-score	support
1	0.98	0.99	0.99	62542
2	0.99	0.98	0.99	61564
accuracy			0.99	124106
macro avg	0.99	0.99	0.99	124106
weighted avg	0.99	0.99	0.99	124106

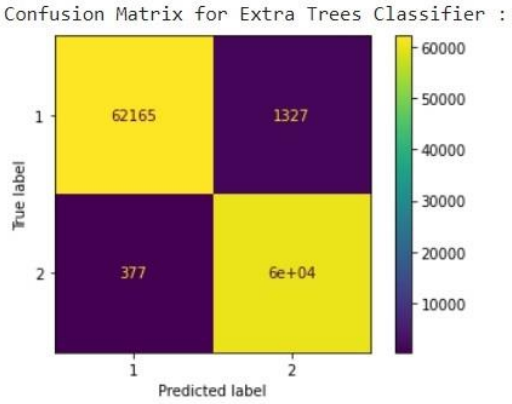

Figure. 5 The Confusion Matrix and related parameters from the Extra Trees Classifier

IV. CONCLUDING REMARKS

The main goal of this study, to create a machine learning model to predict when a product’s price is above the average, has been achieved by the Extra Trees Classifier. The resulting model has a high rate of accuracy, namely 99%, therefore can be concluded to be a very good model. Future research may consider factors other than price, or other data such as economic, financial, or actuarial data.

REFERENCES

[1] D. Ali, M. B. Hayat, L. Alagha, & O. K. Molatlhegi, " An evaluation of machine learning and artificial intelligence models for predicting the flotation behavior of fine high-ash coal", *Advanced Powder Technology*, vol. 29, no. 12, pp. 3493-3506, 2018, DOI: https://doi.org/10.1016/j.apt.2018.09.032

[2] R. Atanassov, P. Bose, M. Couture, A. Maheshwari, P. Morin, M. Paquette, M. Smid, & S. Wuhrer, "Algorithms for optimal outlier removal", *Journal of Discrete Algorithms*, vol. 7, no. 2, pp. 239-248, 2009, DOI: https://doi.org/10.1016/j.jda.2008.12.002

[3] M. C. Belavagi and B. Muniyal, " Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection", *Procedia Computer Science*, vol. 89, pp. 117-123, 2016, DOI: https://doi.org/10.1016/j.procs.2016.06.016

[4] C. Chatfield, "Exploratory data analysis", *European Journal of Operational Research*, vol. 23, no. 1, pp. 5-13, Jan 1986, DOI: https://doi.org/10.1016/0377-2217(86)90209-2

[5] Clickstream Data For Online Shopping : https://archive.ics.uci.edu/ml/datasets/clickstream+data+for+online+shopping

[6] P. Geurts, D. Ernst, & L. Wehenkel," Extremely randomized trees", *Machine Learning*, vol. 63, pp. 3-42, 2006, DOI: https://doi.org/10.1007/s10994-006-6226-1

[7] D. J. Hand, P. Christen, & N. Kirielle, "F*: an interpretable transformation of the F-measure", *Mach Learn*, vol. 110, pp. 451–456, 2021, https://doi.org/10.1007/s10994-021-05964-1 .

[8] I. Jenhani, N. B. Amor, & Z. Elouedi, " Decision trees as possibilistic classifiers", *International Journal of Approximate Reasoning*, vol. 48, no. 3, pp. 784-807, 2008, DOI: https://doi.org/10.1016/j.ijar.2007.12.002

[9] T. F. Laura, J. F. Kevin, & R. B. J. Katherine, "Independent, Dependent and Other Variables in Healthcare and Chaplaincy Research". *Journal of Health Care Chaplaincy*, pp. 161-170, 2014, DOI : http://dx.doi.org/10.1080/08854726.2014.959374

[10] Y. J. Lim, A. Osman, S. N. Salahuddin, A. R. Romle, & S. Abdullah, "Factors Influencing Online Shopping Behavior: The Mediating Role of Purchase Intention," *Procedia Economics and Finance*, vol. 35, pp. 401-410, 2016, DOI: https://doi.org/10.1016/S2212-5671(16)00050-2

[11] S. Manochandar and Punniyamoorthy, M., " Scaling feature selection method for enhancing the classification performance of Support Vector Machines in text mining", *Computers & Industrial Engineering*, vol. 124, pp. 139-156, 2018, DOI: https://doi.org/10.1016/j.cie.2018.07.008

[12] A. Rogier, T. Donders, Geert J. M. G. van der Heijden, T. Stijnen, & K. G. M. Moons, "Review: A gentle introduction to imputation of missing values," *Journal of Clinical Epidemiology*, vol. 59, no. 10, pp. 1087-1091, Oct 2006, DOI: https://doi.org/10.1016/j.jclinepi.2006.01.014

[13] J. J. Salazar, L. Garlan, J. Ochoa, & M. J. Pyrcz," Fair train-test split in machine learning: Mitigating spatial autocorrelation for improved prediction accuracy", *Journal of Petroleum Science and Engineering*, 2021, DOI: https://doi.org/10.1016/j.petrol.2021.109885

[14] K. Singh, M. Elhoseny, A. Singh, & A. Elngar, *Machine Learning and the Internet of Medical Things in Healthcare*, Academic Press, 2021, pp. 89-111, https://doi.org/10.1016/B978-0-12-821229-5.00003-3.

[15] S. Supriyadi, Y. Nurhadryani, & A. I. Suroso, "Website Content Analysis Using Clickstream Data and Apriori Algorithm." *TELKOMNIKA (Telecommunication, Computing, Electronics and Control)*, vol. 16, no. 5, pp. 2118-2126, Oct. 2018, DOI: http://dx.doi.org/10.12928/telkomnika.v16i5.7573