Integrated Machine Learning Approaches for E-commerce Customer Behavior Prediction

Yuran Dong[1, *, †], Junyi Tang[2, †], and Zhixi Zhang[3, †]

[1]*Rensselaer Polytechnic Institute, Troy, New York, 12180 USA*
[2]*School of Business, Macau University of Science and Technology, Macau, 999078 China*
[3]*University of California, Irvine, CA 92697 USA*
[*]*Corresponding author. Email: dongy9@rpi.edu*
[†]*These authors contributed equally.*

ABSTRACT

How to predict the customers' behavior is always a crucial problem for enterprises in E-commerce. In this paper, a data set containing the behavior data for 2019 October and November from a large multi-category online store has been used as well as diverse Machine Learning algorithms are used in Python to precisely predict the behaviors of customers. By extracting 5 datasets containing 10,000 observations out of one billion observations and applying the concepts of Label Encoder, this paper was able to build the models and hence analyze this paper's data. As a result, this paper found that Pipeline and Random Forest works the best that both of them perform a prediction accuracy of 96% which is significantly greater than other algorithms. In addition, the feature of user id and user session present the greatest importance among all the features. On the customers' side, they would focus more on the price-performance ratio, which is price, because it would help customers with making purchasing decisions. This paper were able to recommend individually customized products for each single person based on their personal preference and emphasize the features of data, user id and user session, that sellers should be focus on.

Keywords: *E-commerce, Classification, Machine Learning, Customer behavior.*

1. INTRODUCTION

E-commerce behavior prediction refers to the prediction of consumers' possible consumption demand and habits according to their behavior on E-commerce platform [1]. For example, a user recently browsed the travel bag, tent and other goods behavior. From such behavior, it can be known that the user may have travel plans soon and push some related products, such as sunscreen and raincoat.

Due to the intense impact of Covid-19, the significance of online shopping is far greater than that of offline shopping; the reason behind is simple: online shopping is more efficient, convenient and cost-effective, with more categories comparing with offline shopping. Therefore, there is no need to say more about the fierce competition between different firms of E-commerce. Since the products and services that provided by different E-commerce are basically the same, or identical in other word, what directs users to make decisions between various E-commerce become a sophisticated problem in real life, and this is the motivation that this paper was trying to solve this real-life, business-related problem.

In order to interpret customers' behavior deeper, this paper need to associate with the data that contains the features of customers' behavior. It is not possible for a human being to every record in the table of data word by word and then summarize the characteristics of customers' behavior since that is too time-costly and inefficient; therefore, this paper needs to utilize the algorithm in Machine Learning. The learning algorithms work based on the strategies, events, prediction that succeeded or worked well before, these algorithms have high possibility that it would continue to work well on the desire target since they ideally share some features and characteristics in most cases. Machine learning models show, on average, approximately 10% more accuracy in relation to traditional models based on the comparison between machine learning models and traditional statistics techniques such as discriminant analysis and logistic regression [2]. Methods of random forest and neural network have lots of advantages over

traditional models and would work well with the dataset.

Traditional prediction models are not fit for the enormous dataset in the area of E-commerce and cannot work out with accurate prediction [3], under-fitting would generally happen since most of the traditional prediction models are unable to capture the correct relationship between input and output. It occurs mostly when the models are too simple to predict the correct output therefore high bias and low variance on output might happens and hence resulting error or poor performance of prediction. On the other hand, data mining and machine learning methods are born to be studying these big data problems.

Thus, instead of the traditional statistical analysis, Machine Learning algorithm shows a deep understanding on model, especially the complex multiple interactive or non-linear relationships. In addition, traditional model always does the prediction based on the assumption that might not be true, or even indistinguishable whether is true or false, just like black-boxes. Machine Learning, on the other hand, doesn't apply any assumption; it leaves the data as unknown and only take the input and desired output to build its model so that it could produce outputs that focus on precision and accuracy over interpretability [4]. As a result, this paper would be able to attract more customers and gain more profits in the area of E-commerce.

As a result, this paper decided to apply relevant machine learning algorithms to the E-commerce dataset this paper selected from Kaggle website to see if there are discovery that cannot derive from traditional statistical analysis. Not only that, this paper managed to explore further to find out which are the influencing factors that affect the E-commerce customers' behaviors. On the basis of the feature importance analysis, this research result would certainly contribute to the improvement of E-commerce companies marketing strategy and help them better predict their clients' behaviors.

2. DATA

In this section, this paper would introduce the basic information of the dataset chosen and the related features.

2.1 Dataset description

This paper used a dataset of E-commerce behavior data from multi category store between October 2019 and April 2020. This dataset has been selected in Sep 18th, 2021 from the website of Kaggle[5]. Each column in the dataset represents a unique feature that events could possess. Each row in the dataset represents an event. All events are related to the characteristics of customers and products. Finally, this paper's dataset are 2 comma-separated values files containing approximate 110 million users' events in total. Table 1 below shows the related elements of the 9 features in the dataset, which reflect the events related to products and users.

Table 1. Description of the dataset's features.

Feature Name	Feature Type	Description of data and Possible Values (for Nominal Features)
Event-time	Numerical	The time of the event happens (in UTC)
Event-type	Nominal	View, Adding to cart, Remove from cart, Buy
Product-id	Numerical	The ID of a specific product for classification
Category-id	Numerical	The ID of a specific product's category for classification
Category-code	Nominal	The taxonomy of a specific product's category to make sure whether it is possible to produce it
Brand	Nominal	Down-cased brand name
Price	Numerical	Price of the product
User-id	Nominal	Permanent ID for a user for recognizing the specific user
User-session	Nominal	Temporary session ID for a user for recognizing the specific user's session

2.2 Descriptive statistics

As described before, the dataset has 4 numerical data in the dataset, which are Event-time, Product-id, Category-id, and Price. This paper analyzed and visualized several basic statistical attributes for those data: counts, mean, standard deviation(std), and min, accuracy of 25%, 50%, 75%, max. Counts represents the total number non-empty or null variables. Mean represents the mean of the variables. Standard derivation represents the standard derivation of the variables. Min represents the minimum value in the variables. Max represents the maximum value in the variables. Accuracy of 25%, 50%, 75% represents the corresponding percentage of variables are below the value following the percentage.

Table 2. Descriptive statistics of numerical features.

	Product id	Category id	Price	User id
count	4835.00	4835.00	4835.00	4835.00
mean	6.00e+06	2.06e+18	354.21	5.37e+08
std	1.10e+07	1.68e+16	375.24	2.11e+07
min	1.00e+06	2.05e+18	0.87	4.09e+08
25%	1.01e+06	2.05e+18	110.66	5.16e+08
50%	1.80e+06	2.05e+18	218.94	5.33e+08
75%	5.05e+06	2.05e+18	460.50	5.60e+08
max	1.00e+08	2.17e+18	2574.04	5.80e+08

As above figure shown, these descriptive statistics doesn't work well on product id, category id and user id features. The reason is that all of these three features are designed to distinguish the unique product, category and user. However, based on the Table 2, this paper could analyze some features about price. There are 4835 products here with their mean value equal to 354.21 which means that for every person who purchase a product on E-commerce, it would usually cost him or she 354.21 to buy this product on average. And the highest price that a person purchases a product is 2574.04 and lowest price of a person purchasing a product is 0.87. There is a great price gap between different products. The standard derivation of price is 375.24 which represents that the amount of variation of the price is staying in a normal range that people seldom purchase a product that not only at a very high price, but also at a very low price. The general range of price for a product is about 0 to 600.

Among these four numerical features, there are two features that is available to be compared with: Price and Event-time, because this paper could analyze what is the peak time for customers to make some events and what is the price segments are favored by customers. According to the initial data, this paper can explain several prominent characteristics of data based on the distribution of data.

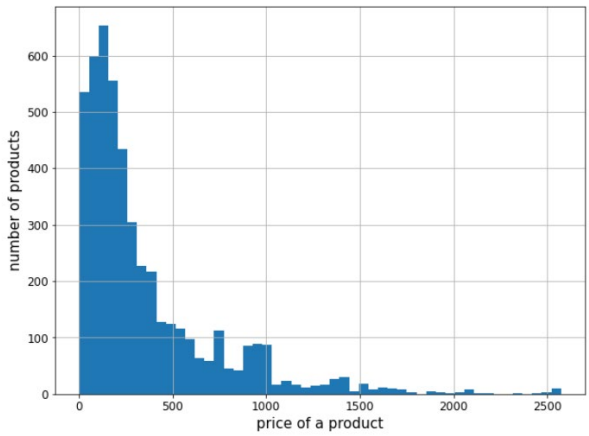

Figure 1 Price distribution.

As figure 1 shown, even though products that sell at a price at 2500 would bring 5 times the profits more than the products that sell at a price at 500, the proportion of products that sell at a price below 500 far exceeds the proportion of products that sell at a price above 2000. Thus, it is easily to conclude that the E-commerce might need to focus more on the customers who target price of products are below 500, which are generally considered as the middle class.

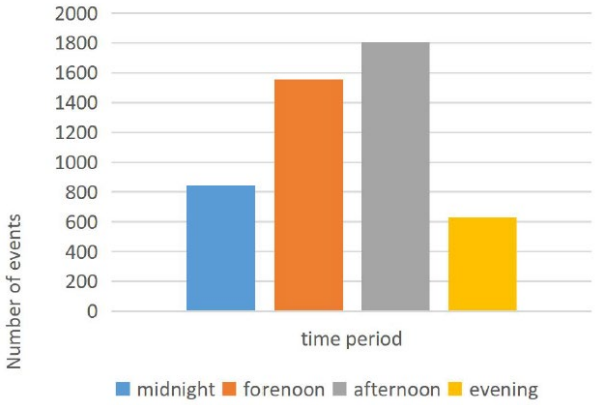

Figure 2 Number of events versus time period.

There are 24 hours in a day, and this paper manually separate them into four time period: Midnight, Forenoon, Afternoon, and Evening. Each of these time period contains 6 hours; that is Midnight represents the time period of 00:00 a.m. to 6:00 a.m. Forenoon represents the time period of 6:00 a.m. to 12:00 p.m. Afternoon represents the time period of 12:00 p.m. to 18:00 p.m. Evening represents the time period of 18:00 p.m. to 00:00 a.m. The reason is that this paper want to see in which time period customers access to the E-commerce the most. As this paper could see from figure 2, this paper could know that most people access to the E-commerce in the time period of 12:00p.m. to 18:00p.m. and the number of events is about 1750.

In addition to nominal features, this paper does consider the Event-type as an important element because this represents what do customers indeed do for each event, whether they are trying to view the products, or they want to put the products that they like into their personal cart, or they indeed purchase the products and bring profits for the E-commerce. In graph 3, this paper observed that about 93.86% of the events are customers viewing the products and 4.28% of events are customers putting the products into their personal cart and finally only 1.86% of events are customers indeed purchase a product. Hence, this paper was able to see that most people don't purchase the products, not evening putting the products into their personal shopping carts. And this is the problem that this paper was trying to solve, how to increase the percentage of customers who indeed purchase some products and hence bring the companies more profits.

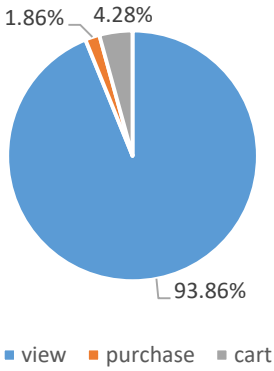

Figure 3 Behavior percentage of customers.

3. METHODOLOGY

In this paper's project, this paper use KNN, Neural Network, Random Forest, AdaBoost, Pipeline to measure the data.

3.1 Methodology and algorithms

3.1.1 KNN

The KNN is a non-parametric classification method first developed by Fix and Hodges in 1951 [6], and later expanded by Cover in 1968 [7]. The core of KNN is the K neighbors closest to it in each sample to represent data. A sample is most similar to k samples in the dataset, and if most of k samples belong to a category, that sample also belongs to that category, an example of KNN is shown in figure 4. KNN can be used for both regression and classification. The distance from the unknown object to the nearest neighbor is measured in different ways. The most common is the Euclidean distance. The distance is from the unknown (green circle) to the nearest neighbors.

$$Euclidean\ distance = d(p,q) = d(q,p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (1)$$

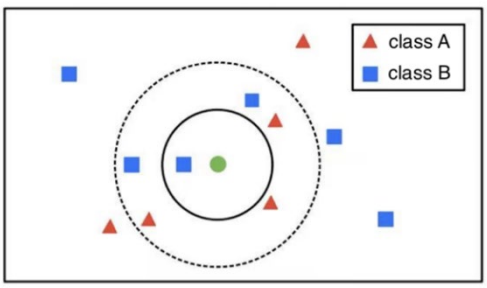

Figure 4 Basic example of KNN.

3.1.2 Neural Network

Neural network algorithms are like human neuron principles, building models based on many neurons, each neuron treated as a unit of learning. There are no links between neurons in the same layer or across layers. Neural networks are good at forecasting and classification.

3.1.3 Random Forest

The first algorithm for random decision forests was created in 1995 by Tin Kam Ho [8]. An example of random forest is shown in figure 5, there are many split trees in the random forest, and when this paper classified a sample, classify it in each tree, and derive the result from each tree, each number will produce a result, the random forest will randomly extract and get different classification results from these classifiers, and produces the most common output as its classification. It is also used for regression or prediction problems.

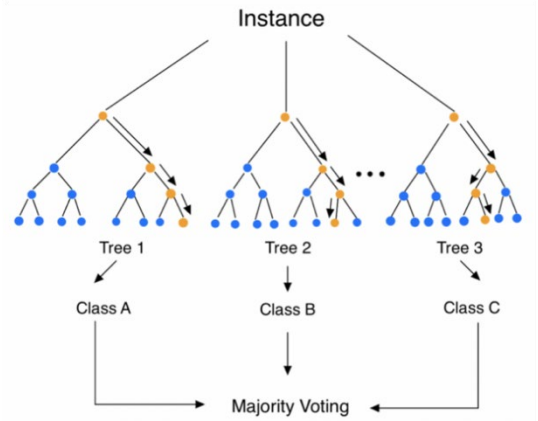

Figure 5 Flowchart showing random forest. *Notes:* The blue dot is the optimal solution for random feature extraction and the yellow dots are the ordinary points of random feature extraction. the dots except the dots on the final of each branch are split nodes and the dots of final of each branch are leaf nodes.

3.1.4 AdaBoost

AdaBoost is a statistical classification meta-algorithm formulated by Yoav Freund and Robert Schapire, who won the 2003 Gödel Prize for their work [9]. It is a training set to train different weak classifiers, and these weak classifiers form a strong classifier.

3.1.5 Pipeline

Converters are often combined with classifiers, regressors, or other estimators to build composite estimators. The most common tool is Pipeline. Pipeline is often used in conjunction with Feature Union, which connects the output of the converter to a composite feature space. Transformed Target Regressor processing transformation goal (i.e., logarithmic transform y). In contrast, only observed data (X) is converted. A Pipeline can be used to link multiple estimators into one. This is useful because there is usually a fixed sequence of steps in processing data, such as feature selection,

normalization, and classification [10].

3.1.6 Methods of evaluate the accuracy of the model

Confusion matrix is a measure used when solving binary classification problems. This research used confusion matrix to evaluate the accuracy of a classification which is shown in Table 3. True Positive (TP) is the situation when Actual class is positive, the predicted class is positive. False Positive (FP) is the situation when Actual class is negative, the predicted class is positive. False Positive (FN) is the situation when Actual class is positive, the predicted class is negative. True Negative (TN) is the situation when Actual class is negative, the predicted class is negative [11].

Table 3. Confusion matrix.

	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Notes: The left column of the matrix is actual class and the top row is predicted class.

3.1.7 Accuracy

Accuracy is the most intuitive performance measure, and it is a ratio of correctly predicted observation to the total observations.

$$Accuracy = \frac{(TP+FN)}{(TP+TN+FP+FN)} \times 100\% \quad (2)$$

3.1.8 Precision

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

3.1.9 Recall

Recall is the ratio of correctly predicted positive observations to all actual classroom observations.

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

3.1.10 F1 Score

F1 Score is the weighted average of Precision and Recall.

$$F1 = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (5)$$

4. EXPERIMENTS AND RESULTS

4.1 Preprocessing

In the first step of data preprocessing, there appeared several typical problems concerned with data type and processing capacity. Firstly, the size of dataset, which contains over one billion observations, is too enormous for us to carry through the further analysis, either for this paper's processing capacity and time-consuming consideration. Secondly, for the type of dataset problems, there are needs to transform the data in time format and nominal type to numerical data in order to provide convenience for this paper's further modeling work. Finally, the missing data in the dataset is required to be processed.

To deal with these problems, this paper develops relevant preprocessing methods. For the first step, this paper randomly extract five datasets within 10,000 observations from the whole dataset. Applying the algorithms to the datasets can ensure that the result of the small datasets can represent the whole one.

For the second step, a method of Label Encoder is applied to the nominal variables of the dataset (event time, category id and category code). Label Encoder is a practical way of transforming nominal data into numerical data, creating favorable conditions while doing data preprocessing. And the method worked well with the variables, resulting with a coding dictionary that shows the corresponding value and variables. By the way, for the treatment the variable 'event time', this paper first distributed the time-form values into four groups, which is forenoon, afternoon, evening and midnight, and then put Label Encoder into use. As a result, this paper develops a coding dictionary that transforms forenoon, afternoon, evening and midnight into zero to three correspondingly, converting the event time feature to an easily-analyzed form. And this paper applies the same methods and develop coding dictionaries for the feature of category id and category code.

After all, this research succeeded in solving all the problems met in preprocessing. And for the further model developing work, this research separates the dataset into 70 percent of training set and 30 percent of test set.

4.2 Results and prediction accuracy comparison

The results of the algorithms are demonstrated in Table 4.

Table 4. Results of algorithms.

Methods	Accuracy Rate (%)	Precision rate	Recall rate	F-measure
Random Forest	96.5	0.90	0.91	0.91
KNN	94.3	0.90	0.94	0.92

Ada boost	94.1	0.89	0.94	0.91
Pipeline	96.0	0.96	0.96	0.96
Neural Network	92.0	0.93	0.93	0.93

The accuracy rate of the methods represents for the proportion of the test set that can be explained by the model built in the training set, in other words, can be defined as the count of correctly classified instances. Among all the five data learning methods, it's quite distinct that the Random Forest algorithm and the method of Pipeline work the best with accuracy rate of 96.47% and 96.00% correspondingly.

Obviously, the method of Pipeline has the largest value for all of the three measurements, revealing its outstanding fitness within the dataset prediction.

In summary, there are two methods in Table 5 reveals a magnificent result: Pipeline and Random Forest. Random Forest tends to have higher correct rate and precision than other algorithms and also performs well in the area of Recall and F-measure. Pipeline performs the same with the accuracy rate of 96% and surpasses the other methods in the area of precision rate, recall rate and F-measure.

4.3 Feature significance analysis

In this part, there would be further analysis on the feature significance to discover more. In the Random Forest algorithm, the importance of features can be easily obtained using the methods of Gini Importance, which is also called Mean Decrease in Impurity (MDI), and this method is applicable for tree-based model. Each node in the model split the data from its parent node on the feature that gives the greatest improvement in Gini impurity.

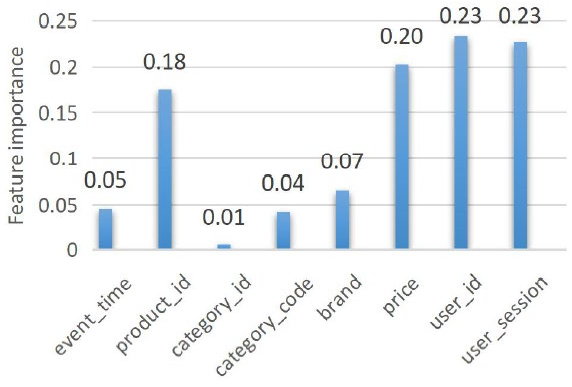

Figure 6. Feature importance using MDI

The results of feature importance using Mean Decrease in Impurity (MDI) is demonstrated in the following figure 6.

As a result, the feature of user id and user session is showing the greatest importance among all of the features, showing us the significance of segmenting customers and studying their behaviors. This can be explained by todays' big data environment and the prevalent product customization. Once firms have an understanding of who is online, they need to focus on how consumers behave online. Models of consumer behavior attempt to predict or "explain" what consumers purchase and where, when, how much, and why they buy. The expectation is that if the consumer decision-making process can be understood, firms will have a much better idea how to market and sell their products [12]. Consequently, in the period of big data, most successful online stores transferred their focus on building big data recommendation system, which can help analyze the preference of customers in line with their view history and purchase history and then recommend the befitting commodities to those customers who are likely to buy them.

By the way, the cost performance of merchandises can be crucial as the features of product id and price also plays important roles in the prediction of customers' purchase action. The reason behind is quite simple – among the most important reasons are price and the availability of free shipping. That the seller is someone whom the purchaser trusts is also a very important factor. The ability to make a purchase without paying tax and the availability of an online coupon are also significant factors [12]. In conclusion, when customers have already decided the target merchandise they are up to, comparing the performance and cost of each similar merchandise would be the foremost influencing factor.

5. CONCLUSION

This paper manages to study how to apply machine learning approaches to the area of E-commerce customer behavior prediction and this paper's goal is to discover the machine learning model that shows the best fitness of the problem and can help us predict the action of customers.

On the purpose of the article, this research chooses the target dataset from the Kaggle website. It includes the data of a large multi-category online store from October 2019 to April 2020 and is made up of features like event type, event time, price etc. and there are attempts to predict how online customers would react under different circumstances. In the beginning, this paper briefly analysis the character of each feature itself. This research investigates on the data distribution for numerical data and count the quantities of each group of the feature for nominal data. Facing the difficulties of nominal data, missing data, time series data and the enormous size of dataset, this research removes all of the missing rows and then use the methods of data extracting and Label encoding. The dataset after preprocessing no more contains nominal features, which

is much easier for the further model development as a result.

Subsequently, this research develops and apply five algorithms on the extracted dataset and then evaluate each model by measuring their accuracy rate, recall rate, precision rate and F-measure value. The measurements demonstrates that Pipeline method and Random Forest tends to have the outstanding fitness towards the prediction.

Based on the Random Forest model, this research carries out the further analysis of feature importance that studies which of the features in the model have the greatest influence towards the purchase action of online customers. The analytical methods are based on Mean Decrease in Impurity (MDI) method, which is designed for tree-based model, and this research discovers that the features of customer information (customer id and customer session) and product information (product id and price) take vital role in affecting customer behaviors.

Firstly, the methods of Pipeline and Random Forest perform well and surpass the other algorithms in the area of predict accuracy and other assessed values. Their accuracy rates are all above 96 percent, which is 96.47% and 96.00% correspondingly.

Secondly, the features of user id and user session have the greatest impact on predicting customer behavior among all of the features, showing us the growing significance of studying on customer information and developing specific recommendation system and customization service.

Thirdly, online customers would evaluate the cost performance of merchandises as the features of product id and price plays important roles in the prediction of customers' purchase action. After clients have decided their target goods, the user performance of the product and the cost of acquisition, including product price, transportation expense, discount rate and etc., turn to be the most important factors that influence the decision of customers.

Studying the E-commerce customer behavior has a far-reaching significance. On the one hand, this paper's predict models can help us understand the relationships between E-commerce behaviors and other relevant features like price, brand or type of merchandise. On the other hand, E-commerce companies using the predict models can learn from this problem and develop further recommendation systems. By grouping and tagging the customers and analyzing the product that they are interested in, companies can form a more excellent recommendation system with higher succeed rate in promoting merchandise.

However, there is still shortage exist in this paper's research. Because of limited computer capability, this research cannot develop the data mining models on the whole dataset and this research use data extracting and analysis the extracted dataset of 10,000 observations. There could be more discoveries when accessing a larger dataset or the whole dataset.

REFERENCES

[1] N. Shaw. Ecommerce demand forecasting: get it right & leapfrog your competition. Available: https://www.bigcommerce.com/blog/ecommerce-demand-forecasting/

[2] F. Barboza, H. Kimura, and E. Altman, "Machine learning models and bankruptcy prediction," *Expert Systems with Applications*, vol. 83, pp. 405-417, October 2017.

[3] M. Piwoni. (May 2017). eCommerce: How to use statistics to increase sales?. Available: https://towardsdatascience.com/ecommerce-how-to-use-statistics-to-increase-sales-1830cf7eb684

[4] M. Stewart. (May 2019). The Actual Difference Between Statistics and Machine Learning. Available: https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3

[5] M.Kechinov. eCommerce behavior data from multi category store. Available: https://www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi-category-store

[6] Fix, E., & Hodges, J. L. (1989), "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties," *International Statistical Review / Revue Internationale de Statistique*, 57(3), pp.238–247. Available: https://doi.org/10.2307/1403797

[7] Altman, N. S. "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression." *The American Statistician*, vol. 46, no. 3, [American Statistical Association, Taylor & Francis, Ltd.], 1992, pp. 175–85. Available: https://doi.org/10.2307/2685209.

[8] Ho, T. Kam (1995), "Proceedings of the 3rd International Conference on Document Analysis and Recognition," *Montreal*, QC, 14–16 August 1995. pp. 278–282.

[9] Wikipedia contributors. "AdaBoost." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 18 Sep. 2021. Web. 27 Oct. 2021.

[10] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

[11] R. Joshi. (September 2016). Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures - Exsilio Blog. Exsilio Blog,, Available: http://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/.

[12] C. Laudon and G. Traver, "E-commerce 2020-2021, Business," *Technology and Society*, 16th Global Edition, ch. 6, pp. 386-388.