2021 6th International Conference for Convergence in Technology (I2CT)
Pune, India. Apr 02-04, 2021

A Study on Impact of Feature Selection on Product Valuation

Sankalp Jain
M. Tech Scholar
Information Technology
NIT Raipur
sjain.mtech2019.it@ntirr.ac.in

Naresh Kumar Nagwani
Associate Professor
Computer Science & Technology
NIT Raipur
nknagwani.cs@nitrr.ac.in

***Abstract*—E-commerce is emerging as the most favored retail destination. As more and more people are connecting to internet, the sales and revenue of online markets are increasing rapidly. In order to encourage more people to shop online e-commerce platforms set attracting offers on their products, offers which a common brick and mortar owner can never think off. To sell products in such low rates and also earn profits, online markets use machine learning algorithms to analyze their sales transactions and build better algorithms to predict prices and offers. In this paper transactional data is analyzed to discover features which influence offer value of products and then use different regression algorithms to predict better offer value for products.**

Keywords— Pricing, E-commerce, Feature importance, offers, consumer, online retail, brick and mortar, transaction, Linear Regression, Decision Tree, Random Forest Regressor.

I. INTRODUCTION

Electronic commerce is one of the business models which allow small and medium sized businesses and consumers to purchase and sell products online. There are four market segments in which e-commerce functions business to business (B2B), business to consumers (B2C), consumer to business (C2B) and consumer to consumer (C2C) [1]. Companies like Amazon and Flipkart works on B2B and B2C model. This paper focuses only on Business to Consumer part of e-commerce. The trend of purchasing online is growing rapidly. In 2019 consumers spent $601.75 billion purchasing online, which resulted in 14.9% growth compared to 2018 in United States. Sales of Brick and mortar stores were increased by 3.8%, which is far less compared to the online market. In the other hand India's e-commerce retail market touched $30 billion mark, and is estimated to cross $200 billion by 2026 [2]. Rise in the electronic commerce requires machines to develop capabilities to evaluate products better than humans. Algorithms which can evaluate product with some metric and label them with suitable offers are required. Predicting offers for product has become a vast research area, as offer value has become the most important feature which customer sees before purchasing a product. Ex: to evaluate an automobile, a automobile system requires information of vehicle, to produce dynamic offers to get better consumer reach [3]. Developing such algorithms does not require huge computational powers but requires machines and humans with better understanding of data.

II. RELATED WORK

As internet is expanding more and more people are connected to each other. The current count for active users of internet is in billions. To convert these billion peoples into active customers businesses are expanding online. People are switching from traditional brick and mortar stores to online stores. There are multiple reasons for this economical shift.

First, the variety and range of products offered by e-commerce is big advantage towards the traditional physical stores [4].

Second, the amount of freedom which consumers get while purchasing online motivates them to prefer online purchase [5].

Third, electronic commerce is a highly transparent market, which makes it easier for consumers to search for products online and compare them with multiple other products to get better product value [6].

Increase in research in the field of machine learning has changed the way how organizations price their products, thus organization are focusing more on the aspect of dynamic pricing of products than traditional pricing strategies [7]. The three major factors which cause this change are: (1) easy availability of consumers purchasing data, (2) flexibility in pricing due to enhancement in technology, and (3) favorable pricing by using different decision support systems for better product evaluation [8].The trend of utilizing consumer behavior and analyzing factors affecting sales of products based on user generated data is gaining importance across the world [9]. Analyzing transactional data give different insights of consumers purchasing behavior, which can be used to segment consumers on the basis of their choice of product and buying patterns [10]. Consumer purchasing pattern can help companies to show choice specific products to targeted audience. Also proper analysis of data can determine future trends and save millions for organizations from manufacturing products which will have negligible impact on market. A research on consumer behavior revealed that price optimization depends on the company's degree of understanding of its consumer base. In this the major price optimization factor can be consumer's ability to pay [11].

Prices and offers of products are directly influenced by features like product type, brand name, country of manufacturing, product self date, purchase date etc. Adding features like country of parts, origin and manufacturing increases product sales and influences consumers buying preferences [12]. Organizations also admit that when consumers are willing to process and analyze information for highly rated products, then the effect of local identity and origin are diminished [13]. Brand popularity impacts on how offers are set on product. Brand's credibility decreases sensitivity towards price. Type of product plays a major role

2021 6th International Conference for Convergence in Technology (I2CT) | 978-1-7281-8876-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/I2CT51068.2021.9418054

on price sensitivity. Choices of consumer and price sensitivity vary across different products [14].

A survey conducted by Consumer Products group of Deloitte revealed that about two-thirds of costumers go through costumer-written product reviews on the internet before purchasing it. Among those customers who read reviews and rating, 82 percent say their purchase decisions have been directly influenced by reviews, among those 82 percent 69 percent share these reviews with friends, family and colleagues, thus further amplifying its impact [15]. Duan In 'The Dynamics of product sales- An Empirical investigation of movie industry' says that the count of customer rating influences the box office sales of movies [16]. Also according to the research by Clemons, in the beer industry, reveals that the change of rating and positive product reviews have a significant impact on the growth and production of craft beers [17]. There are circumstances when early rating may mislead future buyers, this happens when companies showcase their product as best in the market, but later use of product results in lack of functionality as promised earlier [18]. Another pricing strategy is Inventory driven pricing. In 1995, Hewlett Packard noticed a mismatch between demand and supply which led to excess inventory, this played as the main driver for lowered PC cost in the year [19]. A Servqual model of service quality was used to analyze customer preferences, results indicated that consumers valued the safety and assurance feature the most while purchasing online [20].

There are several machine learning algorithms that can be applied on transactional data to get valuable insights. Which algorithms to be applied depend on the type of data and what outputs we are expecting. In this paper, the model outputs offer for a product by evaluating several price influencing factors, so this is a regression problem. Different regression algorithms like Linear Regression, Decision Tree, and Random Forest Regressor will be used to analyze data and get required output.

III. Methodology

A. Data Understanding

Synthetic dataset is generated for elaborating the presented model in this paper. The dataset has ten different product categories from summer clothing to winter clothing. These ten categories of products are available under eleven different brands. The meaning of attributes is presented in Table I.

TABLE I. Attributes in Generated Dataset

S No.	Attribute Name	Attribute Information
1.	P_Id	Product Id
2.	P_Type :	Type of product.
3.	B_Id :	Brand Id
4.	COM :	Country of Manufacturing
5.	D_code :	Dress Code
6.	NOC :	Number of Clicks
7.	Rating :	Customer rating of product
8.	PS_year :	Product self year, means when product was listed on companies website
9.	PS_month :	Product self month, means on which month product was added on companies website
10.	Markdown :	Are of two types temporary and permanent
11.	P_Qty :	Quantity of items purchased
12.	P_month :	Purchase month
13.	P_day :	Day of purchase
14.	P_cost :	Selling cost of product
15.	Offer :	Discount applied on product
16.	Season :	In which season was product purchased

1) Product sale over months:

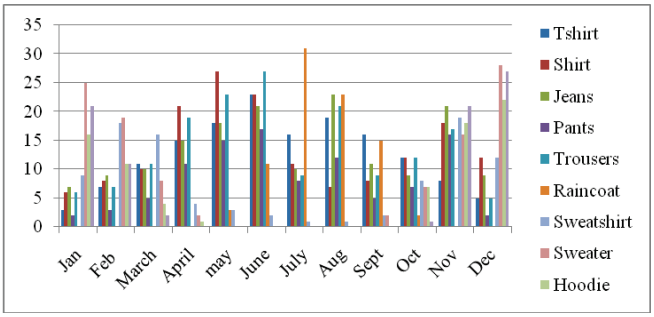

Fig. 1. Product sales over months

The above Fig. 1 shows individual product sales over the span of 12 months. The figure clearly shows the difference in the sale of products over different months.

2) Number of products sold per month:

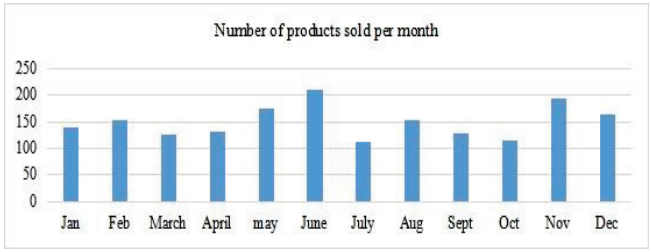

Fig. 2. Number of products sold per month

Fig. 2 shows number of products sold per month. From the figure it is clear that the months like May, June, November and December are the peak selling months. Also Fig. 1 shows that peak sale for winter products is in November and December and for summer products it is May and June.

3) Average offer value over months:

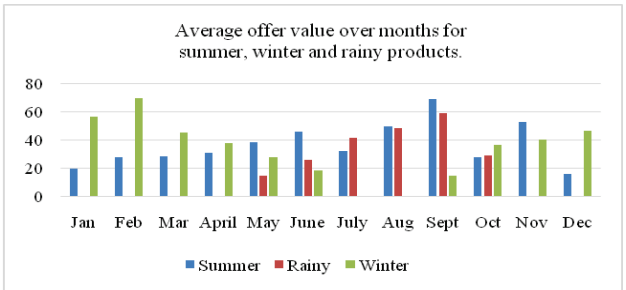

Fig. 3. Average offer value of products over months

Fig. 3 depicts average offer values for summer, winter and rainy products. One of the main driving feature for offer value is season. In the dataset season is categorized in eight categories as:

TABLE II. SEASON TABLE

No.	Season	Time Period
1.	Pre-winter	Late September to October
2.	Winter	November to mid December
3.	Mid-winter	Mid December to Late January
4.	End of winter	Late January to February
5.	Pre-summer	March to April
6.	Summer	May to June
7.	Mid-summer	July to mid August
8.	End of Summer	Mid August to September

During the Pre-winter and Pre-summer, due to low demand of winter products offers values for winter section is low. But in the same time offer value for summer products is high due to end of summer season. This means that in the beginning of any season offer values are low and at the end they are kept high to clear the stock. It is important for organizations to clear their stocks on right time, because fashion changes very fast and no one likes to purchase old fashioned goods.

B. Data Processing

The transactional data consist of integer and float values. These values vary a lot from each other. To train a model well these values should be in a particular range. To bring all the values in different columns in common range normalization is used. The goal of applying normalization is to change the values of numeric columns to a common scale, without affecting differences in the range or losing information. Here, min-max normalization is used to bring the values in common scale.

1) Min-Max Normalization: Min-Max normalization scales every value between 0 and 1. Here, zero is the minimum value and 1 is the maximum value. The Min-Max normalization for the range between 0 and 1 is shown in Eq. (1).

$$New\_Value = \frac{Old\_Value - \cdot}{max - min} \quad (1)$$

C. Data Analysis

In the Table. III attributes like Markdown, P_Type, B_Id and season show good positive correlation, where as attributes like NOC, Rating, D_code and COM show high negative correlation with Offer. Certain attributes like P_Qty and P_cost show low correlation with Offer so these attributes are dropped from the dataset before applying the model. Now we need to create a model which can predict offer value for a product by evaluating all the attributes given above. But before building model we need to split the dataset into training and testing set.

TABLE III. ATTRIBUTE CORRELATION TABLE

Parameter	Value
Offer	1.000000
Markdown	0.376516
P_Type	0.214312
B_Id	0.211510
P_Id	0.211428
season	0.136719
P_day	0.105341
P_month	0.084922
P_Qty	-0.024207
P_cost	-0.032212
PS_year	-0.058964
PS_month	-0.101229
COM	-0.115681
D_code	-0.117175
Rating	-0.272492
NOC	-0.273822

D. Train-Test Splitting

Train Test splitting is done in 80:20 ratio i.e. 80 percent of dataset is used for training the model and the remaining 20 percent is used for testing. While doing train-test splitting the data should be fairly distributed in both the parts. Ex: If we use shuffle function to split the data, then the data gets distributed randomly in 80:20 ratio. But in data there is a binary variable (consisting of 1's and 0's) called Markdown which may not have proper distribution of these values due to shuffle split. Markdown consist 1198 entries of zeros and 70 entries of ones. Due to shuffle split there may be chances that majority of the 1's fall on the test dataset and very few remain in train dataset. This may generate a model which is not trained very well. So, train-test splitting needs to be done on the basis of Markdown.

As the problem discussed in this paper is a regression problem, so various regression models like Liner Regression, Decision Tree, Random Forest Regressor etc. will be used to predict offer value. And whichever model will fit best in the dataset will be used for further analyses. For analyzing the model a evaluation metric is used. Here we will use Mean Squared Error (MSE) as our loss function and r2_score to evaluate the model. R-scored or r2_score is a statistical measure to evaluate how close the input data are fitted to the regression line.

E. Loss Function Used to evaluate model

In order to evaluate the presented work, Mean Squared Error (MSE) is used. MSE is calculated as given in the Eq. (2)

$$\mathrm{MSE} = \frac{1}{\mathrm{n}} \sum (y_i - \tilde{y}_i)^2 \quad (2)$$

[Using equation editor]

Where, n = Number of data points, y_i = Observed values and $\tilde{y}_i$ = Predicted values

After analyzing all the models, the model with the least MSE value will be selected and based on this model feature importance will be evaluated.

IV. EXPERIMENTS, RESULTS AND DISCUSSION

Experiments are conducted on dataset by applying different machine learning models. Models are first applied on training dataset, and if a particular model fits well in training dataset then it is further applied on testing dataset. First model to be applied is Linear Regression. It is one of

the basic regression algorithms which uses a linear approach in order to model a relationship between a scalar response and at least one explanatory variable [21]. It is used to predict value for a variable based on values of some other variables. Applying Linear Regression on train data gives Mean Squared Error equal to 14.228860 which is very high. In order to have better evaluation of model k-fold cross-validation is applied.

Cross-validation is used to analyze machine learning models with limited dataset. It is a re-sampling procedure, this procedure consists of a parameter K which consist of a positive integer value. This K value denotes the number of groups in which the dataset is to be divided. Then in K-1 groups model is trained and in the remaining group model is tested. This procedure is repeated K time until the model is trained efficiently. As such, the procedure is often called K-fold cross-validation. Based on the K value cross-validation is defined. Ex: If K=10 then it is called 10 fold cross-validation. Applying cross-validation on our data with K=10 outputs mean equal to 14.393915 and standard deviation equal to 1.529740 which is still a very large mean value. So, linear regression is rejected.

Experimenting Decision Tree on dataset. Decision Tree Regressor is one of the most commonly used approach for supervised learning. It can be used to solve both regression and classification task. It is a tree based approach in which every path starting from the root represents a sequence of multiple splits until a Boolean output is reached at the end i.e. at the leaf node [22]. Applying decision tree on dataset outputs root mean squared error as 0.111028. This value is clearly a sign of over-fitting. In order to overcome over-fitting cross validation with k=10 is applied, which gives MSE equal to 8.713161 and standard deviation equal to 1.358693 which is much better compared to linear regression. But mean is still big and can be reduced through Random Forest Regressor.

Random Forest is a technique which can perform both classification and regression task. It is a meta estimator that uses multiple decision trees in parallel to get the required output. In classification problem the final output is calculated by majority voting, whereas in case of regression the final output is the mean of outputs from different decision tree. Using random forest on the dataset, obtained value of root mean squared error is 2.497105 which is good. But for better evaluation k-fold cross-validation is important. Applying k-fold cross-validation with k=10, obtained MSE is equal to 6.791021 and standard deviation equal to 1.278437 which is much better compared to linear regression and decision tree. Also the r2_score value for Random Forest Regressor is 0.98 which shows that the data is fitted very well.

Comparing the outputs of all the three models applied, Random Forest Regressor is giving the best results. As mean squared error of random forest is far less compared to all the other models, we will choose Random Forest Regressor as our preferred model. Applying Random Forest on test data gives 5.622029 as mean and 0.0 as standard deviation.

V. Feature Importance

Beside the overall performance of prediction models, it is important to analyze the contribution of each input variable. Out of three models used, Decision Tree and Random Forest model can evaluate feature importance. Based on feature importance organizations can take decisions to add a particular feature in its decision models or drop a feature for providing optimal offers [23].

A. Decision Tree Feature Importance:

The importance of each feature while evaluating offer value using decision tree regressor is shown in Fig. 4. Feature with higher Importance percentage value influences the most.

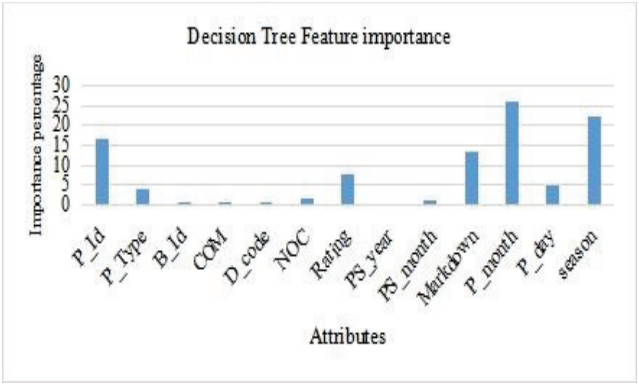

Fig. 4. Decision Tree Feature Importance

B. Random Forest Feature Importance:

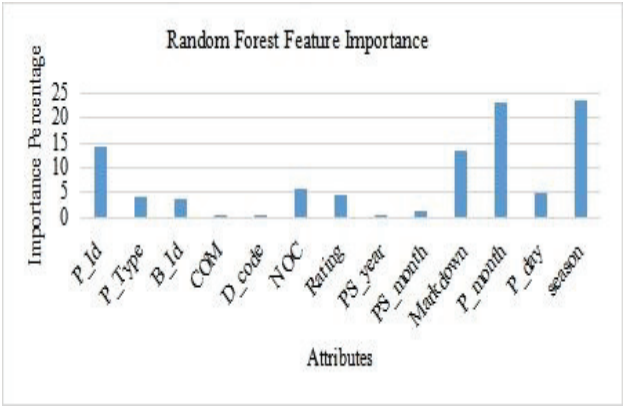

Fig. 5. Random Forest Feature Importance

Fig. 5 shows the importance of each feature while applying random forest regressor on dataset.

C. Comparison:

Comparing Fig. 4 and Fig. 5 it can be seen that the most important feature in Decision Tree is P_month (i.e. purchase month) and in Random Forest Regressor is P_month and season. By analyzing the two figures it can be concluded that variables like Product Type, Brand Id, Number Of Clicks, Rating, Markdown, Purchase month, Purchase day and season play a major role in setting optimal value of Offer for any product.

VI. Conclusion and Future Direction

The data used in this project is very less to properly train a machine learning model. To get the model trained efficiently dataset needs to be increased. Also we have tested only three models on dataset and predicted the output; number of other models can be used to get better predictions. There are 140

different products listed on the company's website which are huge for a dataset of only 1268 tuples. This shows that there may be products which would have occurred only once or twice in the dataset. So, for better outputs increasing the data sets remains the best option. Also finding out new price influencing features in the coming future will give better offer estimation for any product.

REFERENCES

[1] Y. Huang, Y. Chai, Y. Liu and J. Shen, "Architecture of next-generation e-commerce platform," in Tsinghua Science and Technology, vol. 24, no. 1, pp. 18-29, Feb. 2019, doi: 10.26599/TST.2018.9010067.

[2] Agarwal, H. and Dixit, S., 2017. The Encumbered Growth of E-commerce in India: Can we help?. Aweshkar Research Journal, 23(2).

[3] Mercuri, M.E. and Tisdale, J.O., Microsoft Corp, 2012. Presenting offers to consumers based on information from a vehicle. U.S. Patent Application 13/152,261.

[4] Alba, Joseph; Lynch, John; Weitz, Barton; Janiszewski, Chris; Lutz, Richard; Sawyer, Alan; Wood, Stacy (1997). Interactive Home Shopping: Consumer, Retailer, and Manufacturer Incentives to Participate in Electronic Marketplaces. Journal of Marketing, 61(3), 38–53. doi:10.1177/002224299706100303

[5] Pui-Lai To; Chechen Liao; Tzu-Hua Lin (2007). Shopping motivations on Internet: A study based on utilitarian and hedonic value. , 27(12), 0–787. doi:10.1016/j.technovation.2007.01.001

[6] Erik Brynjolfsson and Michael D. Smith (2000). Information Technology Industry || Frictionless Commerce? A Comparison of Internet and Conventional Retailers. Management Science, 46(4), 563–585. doi:10.2307/2661602

[7] Coy, P., 2000. The power of smart pricing. *Business Week*, (3676), pp.160-160.

[8] Dynamic Pricing in the Presence of Inventory Considerations: Research Overview, Current Practices, and Future Directions.Wedad Elmaghraby and Pinar Keskinocak, (2003) pp. 1-2

[9] Z. Zhao, J. Wang, H. Sun, Y. Liu, Z. Fan and F. Xuan, "What Factors Influence Online Product Sales? Online Reviews, Review System Curation, Online Promotional Marketing and Seller Guarantees Analysis," in IEEE Access, vol. 8, pp. 3920-3931, 2020, doi: 10.1109/ACCESS.2019.2963047.

[10] Kung, M., Monroe, K.B. and Cox, J.L., 2002. Pricing on the Internet. Journal of Product & Brand Management.

[11] Armstrong, M. and Vickers, J., 2001. Competitive price discrimination. rand Journal of economics, pp.579-605

[12] Ha-Brookshire, J. E. (2012). Country of Parts, Country of Manufacturing, and Country of Origin: Consumer Purchase Preferences and the Impact of Perceived Prices. Clothing and Textiles Research Journal, 30(1),19–34. doi:10.1177/0887302X11433502

[13] Suri, Rajneesh; Thakor, Mrugank V. (2013). "Made in Country" Versus "Made in County": Effects of Local Manufacturing Origins on Price Perceptions. Psychology & Marketing, 30(2), 121–132. doi:10.1002/mar.20592

[14] Tülin Erdem; Joffre Swait; Jordan Louviere (2002). The impact of brand credibility on consumer price sensitivity. , 19(1), 0–19. doi:10.1016/s0167-8116(01)00048-9

[15] Deloitte, L.L.P., 2012. New deloitte study shows inflection point for consumer products industry: Companies Must Learn to Compete in a More Transparent Age.

[16] Duan, W., Gu, B., and Whinston, A. B. 2008. "The Dynamics of Online WOM and Product Sales-An Empirical Investigation of the Movie Industry," Journal of Retailing (84:2), pp. 233-242.

[17] Clemons, E.K.; Guodong Gao, ; Hitt, L.M. (2006). [IEEE Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06) - Kauia, HI, USA (2006.01.4-2006.01.7)] Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06) - When Online Reviews Meet Hyper differentiation: A Study of Craft Beer Industry. , (), 116c–116c. doi:10.1109/hicss.2006.534

[18] Li, Xinxin; Hitt, Lorin M. (2008). Self-Selection and Information Role of Online Product Reviews. Information Systems Research, 19(4), 456–474. doi:10.1287/isre.1070.0154

[19] Callioni, G., de Montgros, X., Slagmulder, R., Van Wassenhove, L.N. and Wright, L., 2005. Inventory-driven costs. harvard business review, 83(3), pp.135-141.

[20] Gajewska, Teresa; Zimon, Dominik; Kaczor, Grzegorz; MadzÃk, Peter (2020). The impact of the level of customer satisfaction on the quality of e-commerce services. International Journal of Productivity and Performance Management, 69(4), 666–684. doi:10.1108/ijppm-01-2019-0018

[21] H. Lim, "A Linear Regression Approach to Modeling Software Characteristics for Classifying Similar Software," 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), Milwaukee, WI, USA, 2019, pp. 942-943, doi: 10.1109/COMPSAC.2019.00152.

[22] F. Yang, "An Extended Idea about Decision Trees," 2019 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2019, pp. 349-354, doi: 10.1109/CSCI49370.2019.00068.

[23] Prioritizing price sensitivity drivers using machine learning classification algorithms. Clint Rooijakkers, Edinhoven 0804047,2019 pp. 46-69