Proceedings of IAM, V.11, N.1, 2022 pp.50-61

A MACHINE LEARNING BASED APPROACH FOR PRICE ESTIMATION

E. DUYKU[1], M. S. GUZEL[1], E. BOSTANCI[1], I. ASKERZADE[1]

[1] Computer Engineering Department of Ankara University, Turkey

Email: imasker@eng.ankara.edu.tr

Abstract: In this paper a popular e-commerce web site, named, from Turkey will be mined to get a particular products sales volumes in different price ranges with ascending order. By using this ranges many other features such as rating number, rating point, discount ratio, etc. will be extracted to use in the random forest regression model. Machine learning algorithms will be used to find common patterns by using this historical data and anyone who wants to sell any product can use this approach to optimize the price of his or her desired product. At the end of the study we have the bag of words results of the dataset to use for sales analysis.

Keywords- Random Forest Regression, Bag of Words, Web Mining

AMS Subject Classification: 68T05, 68T20

1. INTRODUCTION

Pricing is the most important factor in the e-commerce community. People generally choose their requirements mainly according to their prices. So it is very important to adjust the price of the products accurately to meet the needs of people, e-commerce dynamics, cost of the retailer, shipping, etc. Also estimating the most purchased commodities in e-commerce web sites, gives opportunity of good inventory management to the sellers. Goods must be ready when there is a demand for them and stock costs would be high when there is no demand for a very long time for any good. So compensating these two issues can be handled by accurate price predictions. So accurate e-commerce price predictions can save money, time and many other resources of the sellers and the buyers. On the other hand data has been becoming very important tool in every area. When used in right place data can give much information about the dynamics of the e-commerce environment which can be very difficult by the human observation without interference of the data analysis. So especially e-commerce web-sites give us valuable data resources to use in machine learning algorithms.

2 . PROBLEM DEFINITION

The E-commerce environment has many different problems in its nature. First of all it is not possible to determine which factor is dominant other than prices on the product choices of people. E-commerce web sites give us many indicators such as customer evaluation number, rating number, rating point, discount ratio, seller point, etc. But there is no certain factor that affects customer behavior much more than the other factors. So in our model only one type of product, for instance sport shoes is analyzed. Because for different products such as shoes, mobile phones, clothes there can be many different reasons for customers to buy these products. Even in the same category ,for example shoes, there are plenty of different kinds of shoes such shoes, mobile phones, clothes there can be many different reasons for customers to buy these products. Even in the same category ,for example shoes, there are plenty of different kinds of shoes such as sport shoes, classic shoes, boots, etc. So for different categories there can be more specific reasons which cannot be analyzed by observing the web site data. But when it comes to analysis of a certain product for example sport shoe rather than the any kind of the shoe, it can be accepted that people generally have the same reason or motive for buying this more specific product which would be just doing sport. But this study is not limited for the sport shoes, on the contrary it can be applied many other products which are desired to be analyzed.

Another problem is that it is not possible to know the exact number of purchased products by the customers on the web sites. Even though a data analyst who working for a company can have detail information about his or her company but when it comes to rival companies sales volumes it is not possible anyway. So in our study we will use sales volume of a e-commerce website which is searched a particular product that is grouped by prices levels such as 0-49 ,50-99 ,100-149 etc. total 10 different sales volume ranges which can be seen at this website [6].

Any e-commerce web site can be used for this purpose if it has a search engine option based on the sales volume of a product. This searching has some drawbacks. For example for the minimum price range 0-50 , some very cheap and unrelated products can be found by the web site search engine and this result is not reasonable for our random forest regression model. So results of the cheaper products are taken into consideration carefully when preparing the data for the regression model. Moreover shipping costs have influence on the price-setting. So shipping options of the products will be handled comparing the different combination of features in the training of our random forest regression model.

3 . RELATED WORK

There are plenty of papers which give great importance on price optimization or dynamic pricing with the help of machine learning algorithms.

First of all, in this paper [2] by using statistical and machine learning models, the changing conditions of the prices are taken into consideration to predict whether the customer may buy the particular product or not. This prediction is made by using customer segmentation. This study states that price setting is made by using three different methods. These are agent based, data driven and auction based methods. The core method which includes machine learning models is used as data driven methods and these methods are the main methods explained by this paper. When it comes to the proposed model in this study it can be grouped in three steps; determining purchaser groups, proper pricing for each grouping and prediction of their purchase. Customer segmentation begins with collecting the necessary data from two different databases which are transaction database and offer database. Transaction database includes the main information about product, brand, date, size while offer database includes offer, price, quantity , etc. After collecting required data, data preprocessing is being made. Since variables of the model are not enough to draw meaningful conclusions from the database, new variables are being derived. For example using purchase and offer features, purchase by offer feature is derived. Like that purchase by company, purchase by brand ,etc. features are derived in this paper. Than the outliers which are no or very minor effect on the prices are being eliminated. These attributes are being used to search for similar patterns among the customers. K-means algorithm is being used for the determination of the same customer groups. After that second step is determining the price segmentation according to these customer groups. According to this study supervised learning is the best method for finding appropriate price levels. Every customer group is represented with the different price levels. At the three and the last step by using Logistic Regression model given a customer group and a dynamic price level, a customer may buy the product or not, will be predicted. The whole data set in the study is splitted as %80 for the training and %20 for the testing purpose. According to the study, it is resulted that comparing the revenue of the same product of fixed price and the dynamic pricing according to the customer segmentation, proposed model by the paper got higher revenue. Behavior prediction can play an important role in setting the price of each product. As a conclusion of this paper it can be said that machine learning methods can be used in setting price ranges of e-commerce products dynamically [2].

Secondly in this paper [3] comparison of the rule based pricing and the data driven pricing plays key role. Proposed model in this study is being applied by a book seller on Amazon and data driven model is much more profitable than the rule based model. For the seller it is required to take into

account of many factors such as price, brand, shipping opportunities ,rating, seller points, discounting. Purpose of this paper is to maximize the revenue of discounted books by using these factors. Proposed model is based on sales probabilities with the 10 factors for each seller must take into consideration. Logistic Regression is used to to see the relationship among the offer prices, the market conditions and the sales. The dependent feature is the number of books which are sold. So that possibilities of the sales can be predicted for any offer price and market condition. This prediction model is used later to optimize the prices by observing the current situation. This situation is recycled in every two hours. Results show that data driven model is much more profitable. Discount factor is of great importance for the success of the model [4].

Thirdly in this paper [3] sales prediction is based on a database which includes item, transaction, stocks and many different variables. Three different regression models (linear, decision tree and random forest) are applied one after another to find a suitable model to implement. This paper offers a system structure which produce an ensemble model for each item and some machine learning algorithms, mentioned above, are applied this model for the sale prediction. By applying linear regression, goal of the paper is to find a meaningful relationship between two features, by using decision tree regression its aim is to use tree like graph to draw reasonable decision and by using random forest regression its aim is to find a useful solution based on the whole picture rather than the individual tree [3]

Fourthly in this paper [5] some machine learning algorithms based on the regression are used together to constitute a stacking model especially under limited historical data and new product is taken into consideration. This paper emphasizes that when there is a huge uncertainty, improving accuracy is of great importance. According to this study there are problems about time series considerations on sales predictions such as limited data ,brand-new product, lots of outliers, missing information and many other factors that affect the prices. It strongly asserts that sale prediction is a regression problem instead of a time-series problem. By the way main idea behind the regression approach is that historical patterns may repeat in near future. This study consist of single model, machine learning generalization and stacking of different models. When it comes to data this study uses “Rossmann Store Sales" dataset. Sales distributions for each products and visualiton of the data are the main approach to find the correlations. Moreover in this paper it is stated that static data is base for the machine learning algorithms. This paper uses supervised approach and Random Forest algorithm to find bias and for the error estimation it uses relative mean absolute error. Correction of bias is made by using validation set and it is said that correction by using validation set is very important step on the decision of the iteration numbers. Machine learning generalization is explained by the advantages of the regression approaches. The time series, studied in this paper, shows that for a small number of historical data usage is

much more correct on prediction of the sales. This result is important for the sales of new products. Another crucial step of the study is using stacking of the many machine learning algorithms. By this way estimations of a validation set can be used as input data for another algorithm. On the other hand in time series studies this paper shows that cross validation approach has no usage. Training data and validation data must be in splitted into different periods of time. In short, in time-series problems stacking of the machine learning algorithms rises the level of accuracy [5].

Last paper [6] mainly put emphasis on the advantages of data mining in e-commerce sites. Pricing can be adjusted according to meaningful and related price prediction by data mining activities. According to paper; by using decision trees, proposed pricing policy can be made general and patterns in the previous data can be used in prediction of the customer behaviour. In this paper it is highlighted that pricing is the only element that gives rise to the revenues. After applying more than 1000 test dataset this paper found that there is a nearly -0.18 negative correlation between pricing and rating .It shows that when pricing becomes lower, rating numbers are becoming higher. Again with the same test dataset this paper found that there is nearly 0.065128 positive correlation between productivity and rating. After that this paper explains the using decision tree algorithm for the prediction of the ratings. From a popular web-site some features are being mined, such as item name, price, quantity, type and rating. Next step is determining the common traits of the features and grouping them appropriately. By this way general rules are provided by if-else statements. Than confusion matrix is used by this study to evaluate the model accuracy and at the end 86.4780% accuracy is achieved. For a conclusion, web mining techniques can be used for setting rules for drawing fruitful conclusions about the sales policy and price, product and production quality have a big effect on the customer online ratings [6]. More comprehensive machine learning based approach can be accessed in [3,4].

4 . DATASET DESCRIPTION

For this study e-commerce price dataset is prepared by data mining techniques from this website [6,9,10]. For this project Python 3.7 and mainly BeautifulSoup library is used in order to get the related features of pricing dynamics [2]. There are 10 features in our dataset. These features can be seen in Figure-1. In the dataset there are 280 entries for a price class and there are 10 different price ranges, so totally we have 2800 entries.

Unnamed: 0	product_name	new_price	old_price	discount_ratio	shipping	rating_point	rating_number	seller_name	seller_point	price_class
0	Tatami Minder 50*50 cm 13 mm …	7	17	59	noShi…	90	9	clk-06	100	1
1	Forza Günlük Bayan Spor Ayakk…	39	69	43	freeS…	80	12	forzaspor	96	1
2	Bağcıklı Yazlık Bez Babet-Ayakkabı	39	0	0	freeS…	100	2	Kandişavm	91	1
3	HALI SAHA FİLELERİ (TAVAN AĞLARI)	5	0	0	noShi…	90	6	kocgrass	100	1
4	Emek Erkek Keten Günlük Ayakk…	29	59	50	noShi…	60	1	eterlik	100	1
5	Erkek Çocuk Keten Spor Ayakkabı	29	89	67	freeS…	30	2	OdesaAya…	70	1

Figure-1 Price dynamics dataset features

First feature is product_ name used for controlling whether this mined item is the correct product or not. This feature will not be used in the training phase. For example the first product which is cushion is not a proper item for this project. But this issue is seen generally for the first price range which is the cheapest one. But for the other price ranges we don't see the same situation. So it will be taken into consideration while analyzing the price ranges. Second feature which is the new_ price feature that will be our prediction value. If our model predicts the price as close as this value, our model will be more accurate. Third feature is old _ price which is zero in some items so this item will not be used. Instead of old price value we will use discount ratio which is much more reasonable than old price value. Fifth feature is shipping having categorical values encoded into numeric values. Sixth to ninth features are as following rating point of the seller, rating number of the purchasers, seller name which is also categorical will be encoded into numeric values and seller points. Last and the tenth feature is price class which shows the different ranges of the particular product's sales price. By the way price class value for 1 (one) means the price of the sport shoe is within the 0-49 range, 2 (two) is within 50-99 range,10 (ten) is within 450-499 ,etc.

5. PROPOSED APPROACH

Analysis of the main features for a e-commerce site is the first step in our study. Even though it is difficult to estimate the purchaser inclinations, some data driven analysis and predictions can be made. But in order to get reasonable results we must set price ranges which can change from one person to another so in our study size of a range is determined as 50 . Than every feature is evaluated within its price range. So that we can have some common traits in the same range. By comparing the features with each other we can find some relationships or likeness between the features. After that we will use regression models instead of the classifier methods. Because we want to optimize our product's sales volume by keeping the distance with the higher

sales volume of the same product rather than to classify our product to the price ranges. Any seller can determine two or more price range which can fit his or her product than use our more accurate regression model with the same definite features like shipping, rating point, seller name, seller point, price class to find the best price offer. Than any seller can use these predicted prices for higher class as old price and lower class prediction price as new price for the desired product. In our study we will also use some combination of these features in order to find the most accurate prediction results. After the prediction we can use the bag of words results in order to get meaningful results about the price dataset by user types, sport types, brand name, etc. So applying machine learning approaches to the e-commerce data we can have a great chance to analyze the desired product in many aspects.

6. EXPERIMENTS AND RESULTS

Dataset analysis give us valuable information about the relationship between the discount ratio and the other features like rating point, rating number and seller point. For example when comparing the means of the four important feature values which are discount ratio, racing point, rating number and seller points obviously we can see the similarity between the lines of them in Figure-3.

Especially second price range (50-99) which is seen at the figure above is higher than other price ranges. So that we can infer that if discount ratio is high in range 2 other features such as seller point, rating point can be more high. But when it comes to 8 and more higher ranges there is no such relationship between those features. So analysis of the features is strongly dependant on the price ranges.

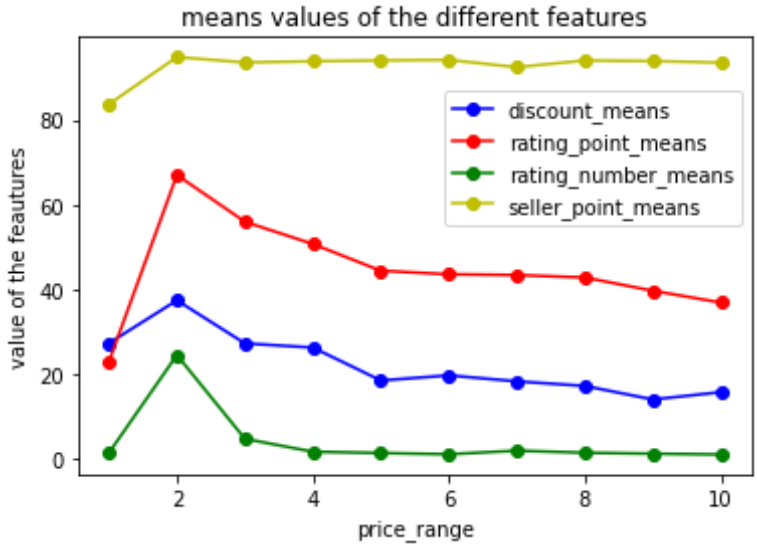

Figure-2 Means of the main features of price dynamics dataset

Moreover first price range (0-49) is more lower than the others except for the discount ratio mean. As stated before for this range there are some items which are not related to sport shoes can be take place in this group. So our regression model may not give accurate predictions about the first range prices. After analyzing the main features most important factor in our study that make the regression results higher is using the price ranges in our models. For example without using price class we use the other reasonable features from 5th to 9th in the figure (from shipping to seller point) in our random forest regression model as independent features. As dependent feature we use 2nd feature, named new_ price .Our training data is 80 % of the dataset and test data is 20 % of the whole data. When it comes to result we see that accuracy is 44.34 % and Mean Absolute Error is 82.7 degrees. On the other hand when we add the price class to those features and apply them to our model we get this result:

Accuracy is 92.76 % and Mean Absolute Error is 12.44 degrees. At the following figure we can see the effect of the features on the results. (Figure – 4)

Nu.	Feature	Accuracy	Mean Absolute Error
1	shipping,rating point,rating_number,seller_name,seller_point	44.34 %.	82.7
2	discount, **price class**	92.38 %	12..25
3	rating_number,seller_name,seller_point, **price class**	92.61 %	12..54
4	shipping,rating point,rating_number,seller_name,seller_point, **price class**	92.76 %	12..44

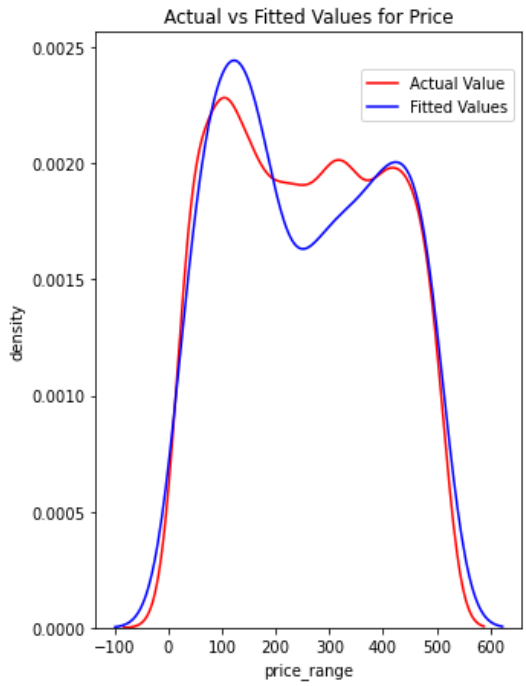

Figure-3 Accuracy of the regression model with different features

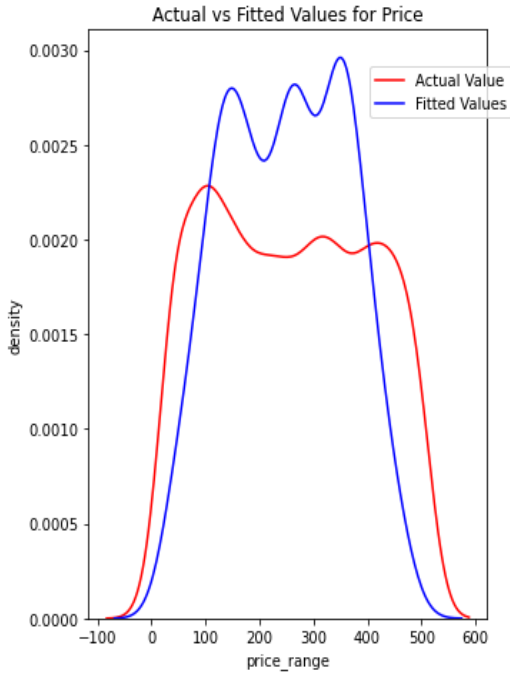

Figure 4.a row = 1 (without price class) Figure 4.b row = 4 (with price class)

In our study Random Forest Regression model takes a very important place. Because this model is an ensemble model which is explained in the related work studies we take advantage of the merging the decision trees together and get more accurate and reasonable results by getting the most accepted results of these trees.

Another aspect of our study is the analysis of the sport shoes according to information in the product name features of the dataset. Figure-5 belongs to the best-selling sports shoes on the inspected e-commerce web site. By using a bag of words approach we got the groupings made by most used words in the name of the products.

Considering the practical usage of this study we develop a website application. This application gives us options to get data from the e-commerce website, train our prediction model and have bag of words results [8].

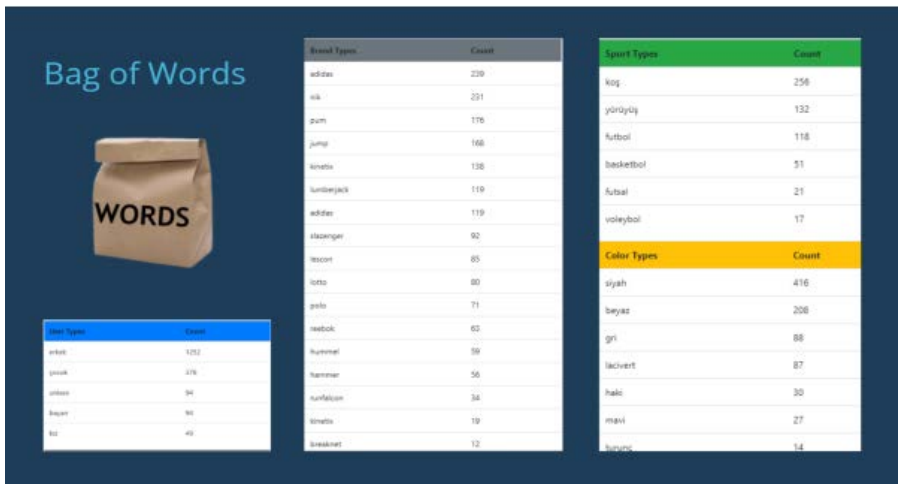

Figure 5. Bag of words approach results

7. CONCLUSION AND DISCUSSION

E-commerce is a very huge searching area for machine learning techniques to put theory into practice. Achieving real data is the main problem to find a reasonable solution to the price optimization problems. So in our study we use a public e-commerce website sales volume based searching engine facilities to construct our dataset named price dynamics. After analysing the main features it becomes obvious that there is a positive relationship between discount ratio and the ratings, seller point and rating numbers at some price ranges. This can be very useful information for the sellers. By the way our main approach for the machine learning model is using random forest regression model which takes advantages of the merging decision trees and get the most useful and stable predictions of them. And for this modelling price_ class feature which holds different prices ranges is indispensable for our study .If much more reasonable and related features are used with these features, there will be more accurate and stable prediction results.

On the other hand if a seller whose online store has already into the sales volume of the same website it is much more predictable for that seller to determine reasonable price for his or her product. Even though this study focused on the sport shoes it is possible to apply it to any other particular

product. So this method is very flexible in applying to any e-commerce website or to any product.

When it comes to drawbacks of the study, it must be said that especially for the first price range improper items must be taken out from the dataset. Also we have no any information about how the studied e-commerce website determine sales volume of the sport shoes is the main drawback of this study. So in order to enhance the focus of this study how to determine sales volume of an e-commerce website must be studied in detail.

REFERENCES

1. Anurag Bejju, “Sales Analysis of E-Commerce Websites using Data Mining Techniques” International Journal of Computer Applications, v.9, 2016,pp.11-19.
2. Bohdan M. Pavlyshenko, “Machine-Learning Models for Sales Time Series Forecasting” Published:18 January 2019.
3. Gupta Rajan and Chaitanya Pathak, "A Machine Learning Framework for Predicting Purchase by Online Customers based on Dynamic Pricing", Procedia Computer Science, vol. 36, 2014, pp. 599-605,
4. Karim A.M., Güzel M.S., Tolun M.R., Kaya H., Çelebi F.V., A new framework using deep auto-encoder and energy spectral density for medical waveform data classification and processing, Biocybernetics and Biomedical Engineering, v.39(1), 2019, pp.148-159.
5. Karim A.M., Güzel M.S., Tolun M.R., Kaya H., Çelebi F.V A New Generalized Deep Learning Framework Combining Sparse Autoencoder and Taguchi Method for Novel Data Classification and Processing, Mathematical Problems in Engineering, Article ID 3145947,2018, pp.1-13
6. N11 e-commerce website. [Online]. Available: https://www.n11.com/spor-giyim-ve-ayakkabi/spor ayakkabi? q=spor +ayakka%C4%B1&srt= SALES_VOLUME &minp=1&maxp=50& ref=auto&pg=1/, 2020
7. Schlosser, R., M. Boissier, A. Schober, M. Uflacker. “How to Survive Dynamic Pricing Competition in E-commerce”. Poster Proceedings of the 10th ACM.,2016
8. Shivani Balduva Kanchan A. Khedikar, Aayushi Jain, Ria Jain, Saumyata Sharma, “Sales Prediction for E-Commerce Site”,A Journal of Composition Theory, v. XIII, No IV,2020, pp. 62-71.
9. The Python website. [Online]. Available: https://pypi.org/project/beautifulsoup4/, 2020
10. Website. [Online]. Available: https://ml-project-master.herokuapp.com/, 2021