Price optimization in E-commerce using ML methodology

Mohamed Bekheet Abdelall Mohamed
Dept. Electrical Engineering
Ottawa University
Cairo, Egypt
mmoha383@uottawa.ca

Omar Ashraf Fathy Ibrahim
Dept. Electrical Engineering
Ottawa University
Cairo, Egypt
oibra063@uottawa.ca

Eslam Abdelraheem Shabaan Khalaf
Dept. Electrical Engineering
Ottawa University
Cairo, Egypt
ekhal066@uottawa.ca

Rokaya Ismail Hussein Mohamed
Dept. Electrical Engineering
Ottawa University
Cairo, Egypt
rmoha121@uottawa.ca

Amjad Dife Friend Dife
Dept. Electrical Engineering
Ottawa University
Cairo, Egypt
adife035@uottawa.ca

***Abstract*—This project tries to solve a price optimization problem .this problem has a conflict between optimization and maximization so a machine learning-based solution has been introduced to solve these problems. the historical data of e-commerce stores are limited so a dataset has been used in this project to solve this problem. a prepossessing step has been developed to put the data in the shape that the algorithms could deal with. clustering techniques techniques are used to assign items to sub-styles. also Regression techniques have been used to predict the quantity of an item to be sold. the results coming from the regressors used to expect the revenue. a revenue matrix associated with a price matrix was developed to be an input for the optimization algorithms.**

I. Introduction

Operating a business at a time when changes are occurring daily, you are unsure of which bandwagons to join and which deals can be risky. Even while traditional commerce offers benefits, many businesses rely on e-commerce or e-business since it requires less capital and gives customers more convenience also the basics of purchasing and selling have entirely changed as a result of the expansion of the internet and technology. Most clients now choose e-commerce websites or stores over traditional brick-and-mortar establishments since they have gained huge popularity on social media, news portals, and search engines so e-commerce is most popular nowadays than traditional trading as shown in figure(1), The Amount of money spent on E-Commerce sites by year (Source: https://www.paldesk.com/will-ecommerce-overtake-traditonal-retail/).

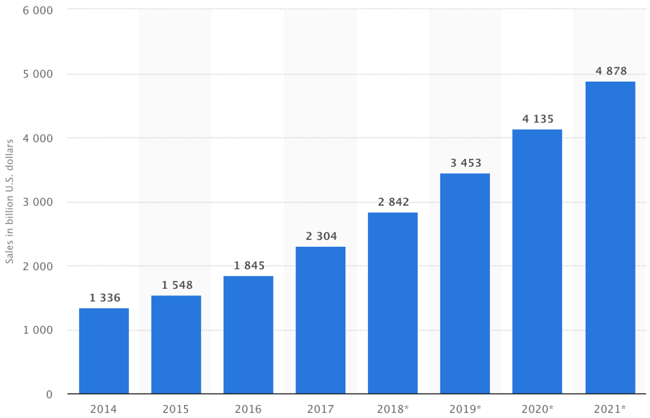

Fig. 1. Amount of money spent on E-Commerce sites by year.

If online sales plan does not incorporate e-commerce pricing optimization, it is still insufficient even if having the ideal product for a captive audience. Pricing is one of the most important ways businesses communicate with their clients, thus pricing optimization should be an important pillar in the business plan, E-commerce pricing optimization, to put it simply, is the act of leveraging data to make sure that products sold online and through applications are priced as competitively as feasible so this work goal is to balance between product pricing and other factors that will be considered in our problem are shown in fig(2) (Source: https://www.openpricer.com/blog/what-is-price-optimization/) .

Price Optimization Factors:

- The customer's willingness to pay should be the first thing taken into account. It depends on the buyer's traits and motivations, as well as the value portrayed by the brands, settings, and prices of competitors.
- The transaction's incremental margin, which includes service costs that will be recorded as variable sales charges, as well as the opportunity cost, which includes future demand at higher rates that will be shifted owing to capacity restrictions.
- The optimization objective function, such as the contribution to profit, revenue, or market share, as well as

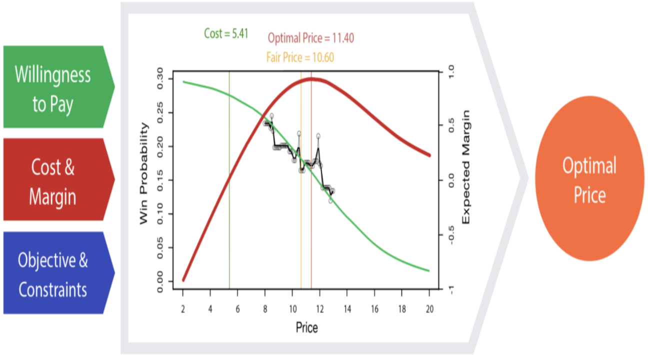

Fig. 2. Price optimization factors.

the restrictions, which in this case may be the minimum and maximum stock capacity rates, must be taken into account.

So the goal is to balance these factors by calculating e-commerce products price that make best revenue by using data science methods.

II. Literature Review

The Algorithm proposed by [1] is one of the algorithms that is applied to predict the optimal pricing in e-commerce.

The main components of the algorithm is illustrated in the following sections followed by the design of the final algorithm and why it is hard to apply such algorithm.

1) The Nadaraya-Watson estimator. The Nadaraya-Watson estimator (non parametric method) is used to predict the revenue or profit of a product depending on the profit percentage of this product. Parametric methods are not used as the profit and revenue functions are complex (having multiple local optima) and different products will have different functions. The Nadaraya-Watson estimator has only one hyper-parameter to be tuned that's why it is used instead of the other non-parametric approaches that have multiple hyperparameters (hence need more data). To predict the revenue or profit of a product given the profit percentage of that product, the estimator uses different profit percentages and the corresponding profit/revenue (historical data). The bias-variance trade-off is handled by the bandwidth (the only hyper-parameter). Choosing low value for the bandwidth will cause overfitting while choosing a high value will cause underfitting. The leave one cross validation is used to determine the bandwidth. The Gaussian kernel is used in the estimator to give more weights to the points that are near to the new profit percentage point [1].

2) Bootstrap-based confidence estimation Bootstrap-based approach is used to predict the probability that a certain profit percentage point is optimal. Instead of using revenue or profit as the target value (the value that we want to optimize), a combination of these two values is used. So, the target value is equally weighted sum of the revenue and profit. The probability of a profit percentage to be optimal is calculated by estimating the probability that the target value that corresponds to this profit percentage is higher than the threshold (nine tenths the highest target value in the past data). These probabilities are used in the Nadaraya-Watson estimator instead of the revenue/profit historical data. Bootstrapping is used to address the problem of the unknown distribution of the probability. Bootstrapping is done by taking samples with replacement from the data points (target values as a function of single price and the corresponding days with this price) with the same size as the dataset. One thousand samples are taken then the average of the target values is calculated. The percentage of samples that have average target value higher than the threshold is calculated and hence this percentage would be the probability that this profit percentage is optimal. The previous steps are repeated with each profit percentage to calculate the corresponding probability [1].

3) Bayesian approach and decision tree. Until now, the whole previous steps are held for individual products. The total amount of data points for individual products may not be enough to apply the previous algorithms or may be there is not any data points at all at new products case. So, the optimal profit percentage for the entire sub-category of items is calculated. The information about competitors' prices and unit cost is also considered while calculating the optimal profit percentage for the entire subcategory. Decision tree algorithm is used to tackle this scenario using the unit cost, competitors' prices, and the profit percentage and hence the subcategory price can be used as a starting point for the product. Bayesian inference approach is used to combine the predictions of the product level with the predictions of the subcategory level. As more data is acquired the importance of the subcategory level estimates is decreased while the importance of the product level estimates is increased, and this scenario is handled by the Bayesian inference approach [1].
4) Metropolis-Hastings-based algorithm. Choosing the prices that will be tested for the optimality is a challenge because the optimal price may be in unexplored area of prices. So, Metropolis-Hastings method with truncated normal distribution is used to determine how to draw new sample given the previous drawn one [1].
5) The final algorithm. For every product, a bootstrap-based method is applied to estimate the probability that the given price is optimal for each price in the dataset based on the target value. Then these probabilities are used by Nadaraya-Watson estimator to get the estimated target value function. The data about all products in the subcategory is used to generate the

decision tree. Then the Metropolis-Hastings method is used to sample a new profit percentage. The functions that are generated by Nadaraya-Watson estimator and decision tree will use this new profit percentage. Then the Bayesian inference method will combine the results of these two functions. The acceptance ratio will be calculated based on the result of the Bayesian inference method. The acceptance ratio is used to determine whether the new sampled profit percentage is better than the previous one or not. The Metropolis-Hastings will sample a new profit percentage again and repeat the previous scenario again. The Metropolis-Hastings method will repeat this scenario for one hundred times to get the optimal price for each product [1]. To apply the previous algorithm, the profit percentage of each product should be contained in the data and because it is hard to acquire such data, we didn't manage to apply this algorithm.

III. System Architecture

in the following arch represent the main step in the price optimization system

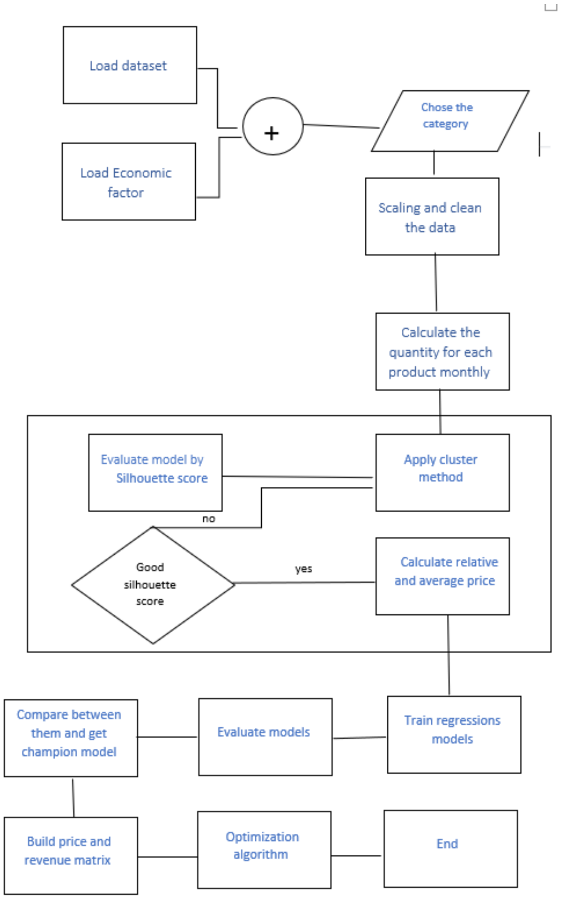
Load dataset
Load Economic factor
+
Chose the category
Scaling and clean the data
Calculate the quantity for each product monthly
Evaluate model by Silhouette score
Apply cluster method
no
Good silhouette score
yes
Calculate relative and average price
Compare between them and get champion model
Evaluate models
Train regressions models
Build price and revenue matrix
Optimization algorithm
End

Fig. 3. the system architecture of price optimizer

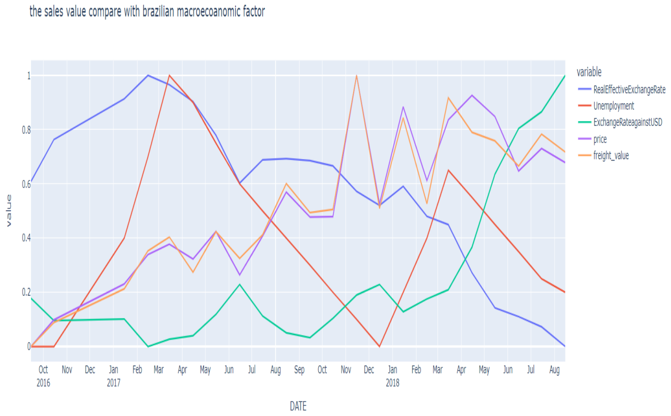

Fig. 4. the effect of economic factors on sales

figure cat represents the product categories included in the dataset

A. Dataset

Olist store Dataset has been used for this project, The dataset contains details on 100k orders placed at several Brazilian marketplaces between 2016 and 2018. it contains multiple datasets Olist order payments dataset in which The order payment alternatives are covered, the Olist order reviews dataset in which customers reviews are included, the Olist orders dataset which is the core dataset, the Olist order items dataset in which Data from each order's purchases are included, the Olist products dataset which contains data about products that have been sold by the store, the Olist seller's dataset which contains sellers data, Olist order customer dataset in which The customer's location and other details about the customer are included and the Olist geolocation dataset in which Brazilian zip codes and their coordinates are included, it has 36 different categories of products, 32951 products, and 99441 orders.

B. prepare the Dataset

The data from different Olist datasets have been merged to get the data required for the modeling phase, the core dataset was used to merge the other datasets. the Brazilian categories' names have been mapped to English words; a web scrapping technique has been used to collect the data about the economic factors from the web to understand the effect of the economic factors on sales. the results shown in eca represent the comparison of the prices with unemployment, Exchange rate against USD, and Real change rate factors[2].

. to eliminate the processing phase the furniture category has been chosen to work on. the values of some features were not on the same scale so, a feature scaling technique was applied to these features. filtering techniques used to count the quantity have been sold in a predefined time interval (monthly)

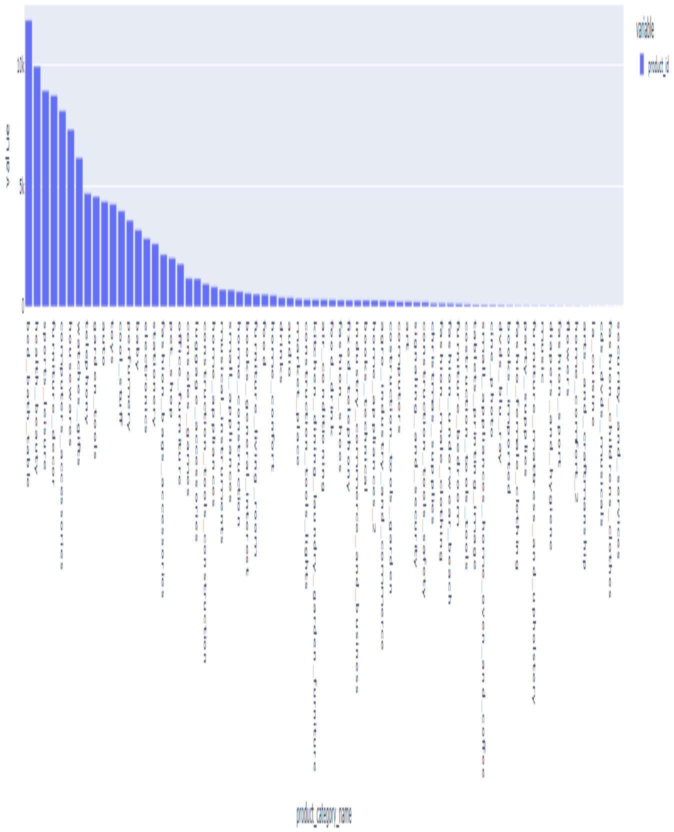

Fig. 5. the sales of each category

IV. MODELING

A. Cluster products and generate substyle

cluster methods are used to collect items that have similar features together to build substyles and the chosen features effective on price decisions so it's better to use an expert in this domain to determine the importance of each feature, the price of each product can't be assigned in isolation, because the price of an item in a cluster affect on and effected by the prices of the other items in the same cluster because the products in the same substyle compete with each item in same substyles. So a clustering technique such as k-means, Expectation maximization, and hierarchical clustering has been used to assign each row in the data set to a substyle. the clustering evaluation criteria such as the elbow method and the silhouette score were different with these different techniques and use them to detect the best numbers of clusters that achieved high between clusters and lower distance into clusters. the figure kmean shows the clustering evaluation methods using K-means, EM, and hierarchical clustering algorithms simultaneously. tsne plots were used to show the results of clustering as shown in figure tsn

B. calculate relative and average price

two factors have been introduced the average price and the relative price; the average price has been introduced as the average of the prices in the same substyle; the relative price has been introduced as the price of an item in the substyle divided by the average price. the characteristics of an item, the average price, and the relative price[2].

C. demand prediction

using the relative and average price and the characteristics of an item were used as features and applied multiple regression algorithms like train Regression models such as Random Forest, Decision tree, Linear Regression, and the Bayesian Ridge to predict the quantity that may be sold used the demand prediction to build a revenue and price matrix in the df shown the decision tree

- initializes the minimum value is less than the given price value with 10 % of the given price and the maximum value is greater than the given price value with 10 %.
- calculates the value equal number of items in a subclass in the range between min and max value and assigns the values in the price matrix.
- predicts the quantity for each price and rounds up the value to be an integer and calculates revenue by multiple quantities with a price.

used the price matrix and revenue matrix as an input for the next step (optimization algorithm)

D. optimization algorithm

the optimization process was introduced to find the price of each item that maximizes the total revenue of a substyle of products as introduced in []there are two ways to apply the optimization the first one using a branch and bound technique the second one using cplex library introduced by IBM can solve these types of problems. as a preparation for the

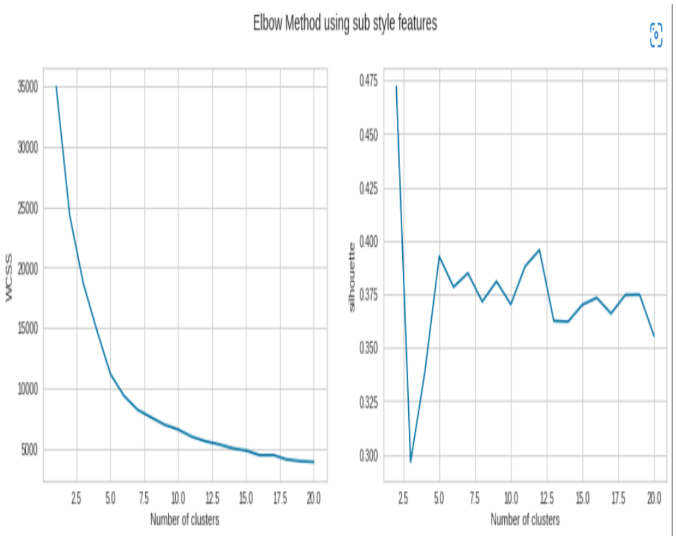

Fig. 6. choose the best number of k cluster

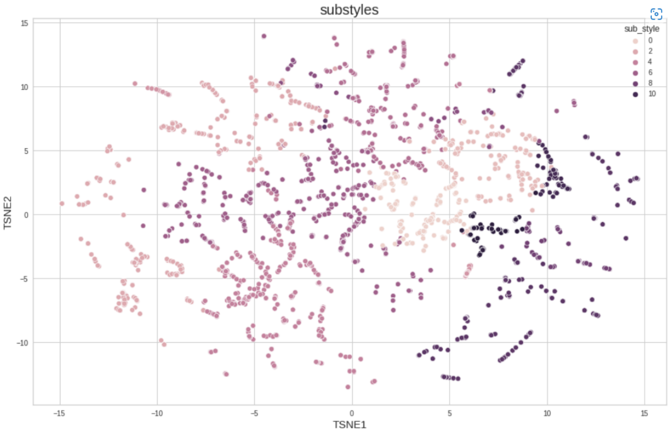

Fig. 7. represent data after applying k-means

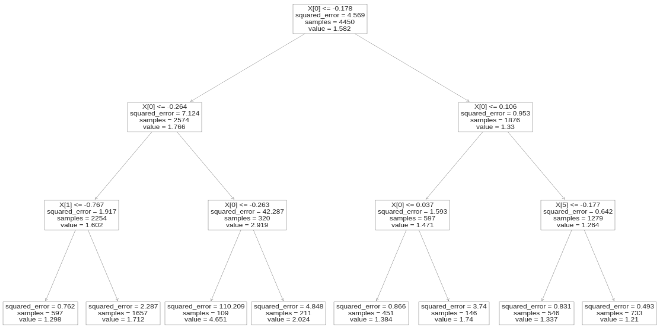

Fig. 8. regression decision tree model

optimization phase, a price set containing a range of prices for each substyle was developed so each price in it can be assigned to each item in the same substyle during the optimization phase. by assigning each item each value in the price set we got the price matrix. using the price matrix and the champion regressor the revenue matrix has been introduced. then cplex library has been used to solve this problem and find the mixed set of prices that maximize the revenue matrix[2].

V. Evaluation

in generator substyles, can find the best value for the number of classes is 12 and after that use, this substyle as shown in tsn and choose for example substyle 3 that contains 610 items and build price set for each item contains the same number of items and predict the quantity for each item with each price and calculate the revenue matrix and apply optimizer algorithm on it

the following figure shows the comparison between four regression models by evaluating them on test data and is used in demand prediction. the champion model is a random forest, it achieved less mean squared error.

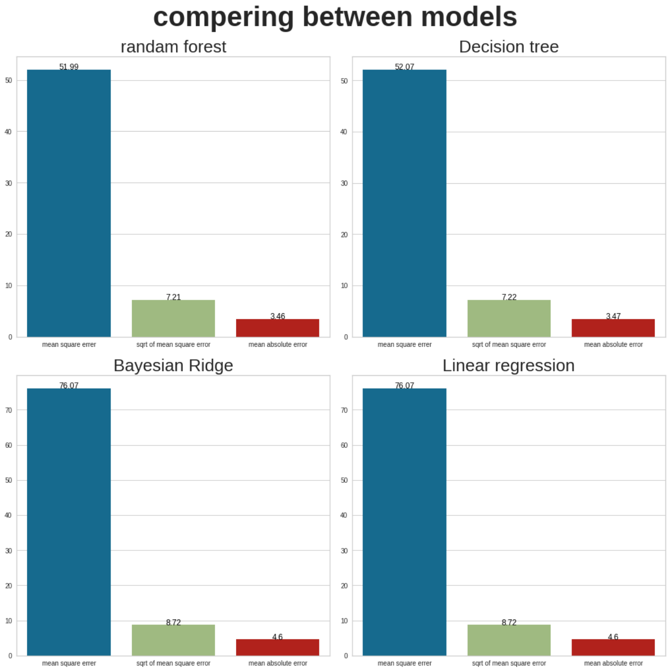

Fig. 9. compare between demand prediction model

VI. Conclusion

The new dataset has been developed by merging the datasets from olist stores datasets. a preprocessing step has been accomplished to remove the missing values and null values. also, the redundancy has been eliminated. to assign the items to a substyle three clustering techniques have been used and the selected champion model has been used to cluster the data. to predict the quantity four regression techniques have been developed and compared and the selected one was used to predict the quantity. for the optimization phase, a cplex library introduced by IBM has been used to solve the optimization task

the data leakage is one of the main limitations of such these problems. as future work we suggest using different data from different stores that contain more information about a specific product during a predefined time interval. deep learning techniques can be applied to predict the demand instead of the traditional machine learning methods. different optimization techniques can be applied like trying to optimize the discounts on the items at a specific time interval. Finally, it would be good to study the relationship between items and their prices which can be accomplished by association rule mining techniques.

References

[1] Bauer, J., Jannach, D. Optimal pricing in e-commerce based on sparse and noisy data. Decision support systems, 106, 53-63. (2018).

[2] Qu, T., Zhang, J. H., Chan, F. T., Srivastava, R. S., Tiwari, M. K., Park, W. Y.Demand prediction and price optimization for semi-luxury supermarket segment. Computers industrial engineering, 113, 91-102. (2017).