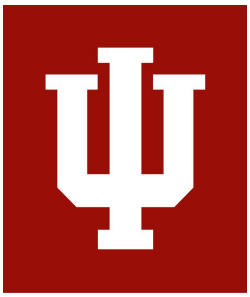

B565 - Data Mining
Final Project

Project Report - Fall 2022

Boosting sales margin of a product using price optimization

Gauri Chaudhari (gchaudh@iu.edu),
Pooja Parab (pooparab@iu.edu),
Divya Manoj (dmanoj@iu.edu)

Under the guidance of
Prof. Yuzhen Ye

Abstract

While shopping on E-Commerce websites, many customers tend to view the same product from multiple sellers or online stores to find the best affordable price. Retail businesses too acknowledge this trait of their customers and offer promotional codes and discounts to increase their sales. Historical data of transaction conveys that the pricing of a product depends on how much it is in demand. Using this perception and by understanding the customer's reaction to price drop and price hikes of a product in the past one year, we have attempted to predict the best optimal price for this product yet increase their profit margin using data mining techniques, linear regression, ridge regression, lasso regression and demand curve.

Keywords

E-Commerce; product; price; retail; data; demand; customer; optimal; linear regression; ridge regression; lasso regression; demand curve; profit; margin; MSE; OLS; prediction;

Contents

1 Introduction

What is price optimization?

There are a lot of factors that may influence the price of a product and the likelihood of the customer to buy that product at a price point. Price optimization deals with analyzing the historical transaction data and figuring out trends and abnormalities in the sales of a product to predict the best price of that product that would not only increase the likelihood of the customer purchasing the product but at the same time increases the profit margin of the company by increasing the sales. This is achieved my modeling and training historical data using various machine learning algorithms resulting into a demand curve that would pin point the approximate best price of the product. We are considering sales volume as our major factor while trying to reach to an optimal price point.

2 Dataset

The Brazilian public E-commerce dataset is a real transactional data with the customers who make purchases at the Olist store based in brazil. Dataset has recorded more than 100k records and 40 features within the 2016-2018 timeframe. The original public dataset was distributed over 8 CSV files which we consolidated into one file with the required features based on our project requirements. For simplification, We are only considering 'delivered' order status records for our analysis.

3 Methodology

The proposed methodology has five phases, the first phase deals with the cleaning and pre-processing of data. We worked on a Brazilian e-commerce public dataset which we obtained from Kaggle. Once the dataset is obtained, cleaning and pre-processing steps are applied to it. The second phase includes Dimensionality Reduction. This is done by Principal Component Analysis. The algorithm is applied to the pre-processed dataset to reduce number of features and obtain the important features that are correlated. The third phase is about Exploratory Data Analysis. After obtaining important features from the previous phase, we performed data analysis to answer some business use cases. The fourth phase is about Modelling. The dataset was split into training and testing data. The data was trained using algorithms like Linear Regression, Ridge, and Lasso

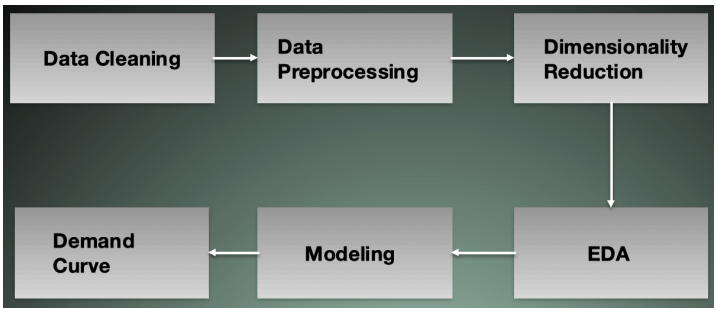

Figure 1: Workflow of this project

algorithms were used for predicting prices. The fifth phase is plotting Demand Curve, where we answer relation between demand and price. The workflow is depicted in fig1.

3.1 Cleaning and Preprocessing of Data

The data was cleaned and preprocessed to remove irrelevant data and generate some useful data. For our data we did the following cleaning and preprocessing -

- Removed incomplete records
- Removed duplicate records
- Extracted features like year, month, year-month from timestamp data
- Filtered data based on order status "Delivered"
- Added order_count column for calculating total sales

3.2 Dimensionality Reduction

It is used to transform data from high-dimensional space to low-dimensional space to retain some meaningful properties of data. We used Principal Component Analysis for dimensionality reduction. The purpose is to retain features that would be more important for the model. We had 40 features, of which we applied PCA on 24 features. The categorical features were encoded using Label Encoding. We computed variance and variance percentage by considering different number of components as shown in fig 2 and fig 3.

When we considered 17 components, there was 90% variance hence we reduced number of features to 17.

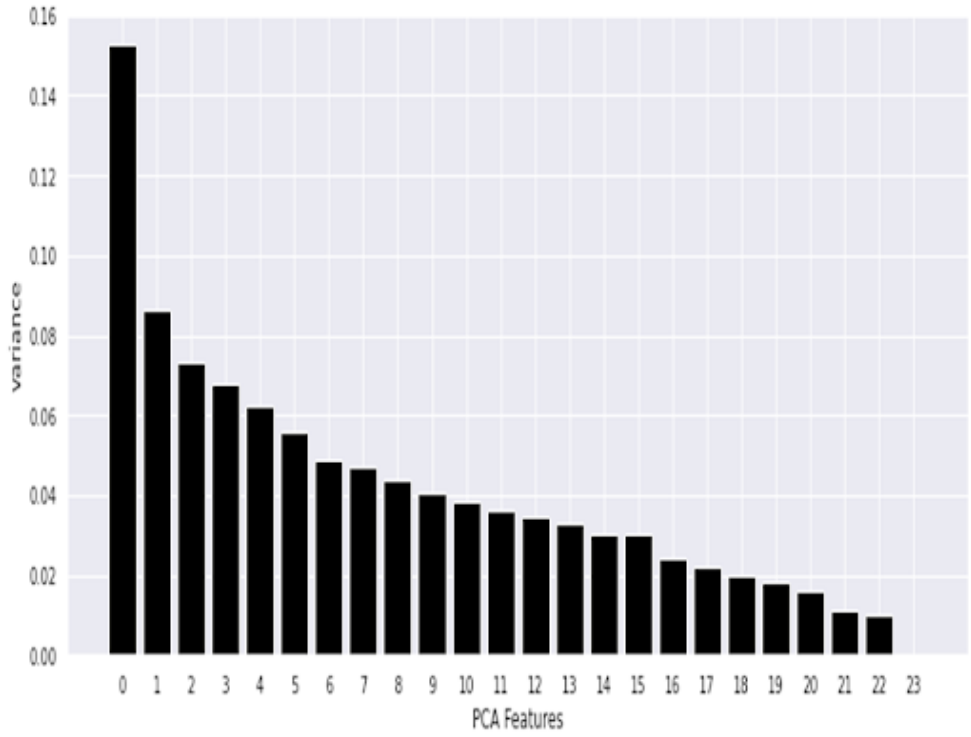

Figure 2: Variance of different features

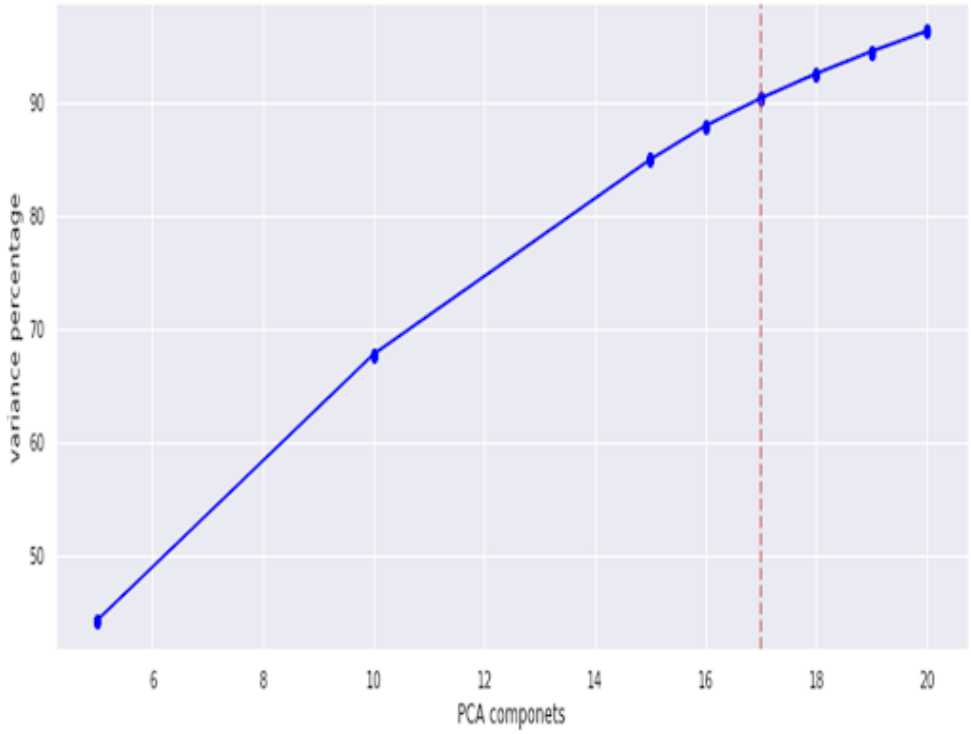

Figure 3: Variance Percentage of different number of components

3.3 Exploratory Data Analysis

We analyzed data and got some meaningful insights.

- We had 71 categories of product.
- The order_count increased with years but suddenly seemed to drop

	year	month	year_month	price	order_count
0	2016	10	2016-10	34642.15	295
1	2016	12	2016-12	10.90	1
2	2017	1	2017-01	108069.81	883
3	2017	2	2017-02	212085.93	1751
4	2017	3	2017-03	336937.19	2788
5	2017	4	2017-04	295445.81	2443
6	2017	5	2017-05	464600.44	3907
7	2017	6	2017-06	393688.89	3412
8	2017	7	2017-07	471900.29	4335
9	2017	8	2017-08	524968.29	4572
10	2017	9	2017-09	584714.56	4504
11	2017	10	2017-10	584238.58	4966
12	2017	11	2017-11	908192.07	8102
13	2017	12	2017-12	642083.75	5762
14	2018	1	2018-01	835396.74	7593
15	2018	2	2018-02	760266.61	7126
16	2018	3	2018-03	881405.22	7616
17	2018	4	2018-04	883044.87	7413
18	2018	5	2018-05	891900.80	7483
19	2018	6	2018-06	791001.53	6788
20	2018	7	2018-07	765577.77	6544
21	2018	8	2018-08	747939.92	6611

Figure 4: Sales volume increasing with months and then reducing again

- In 2018, health_beauty category generated the most revenue.

	category	revenue
43	health_beauty	744735.73
7	bed_bath_table	553929.63
69	watches_gifts	532080.96
15	computers_accessories	516811.72
49	housewares	408959.27

Figure 5: Top categories generating the most revenue

- Black Friday had most number of sales. There is a possibility that the sellers had given out a lot of discounts and promotional offers on this day which resulted in more sales.

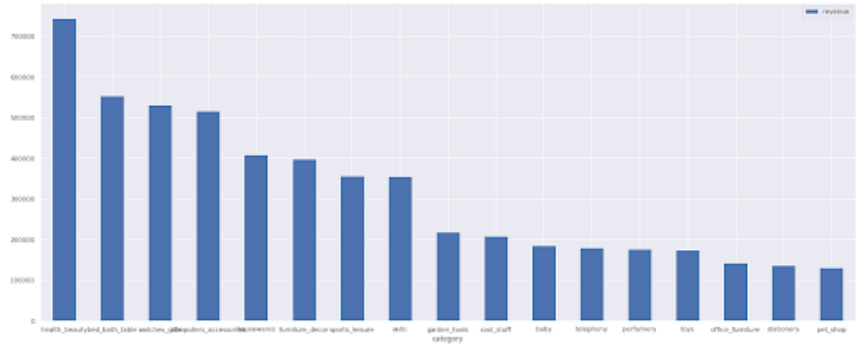

Figure 6: Bar graph of top 17 categories contributing to 80% of the sales revenue

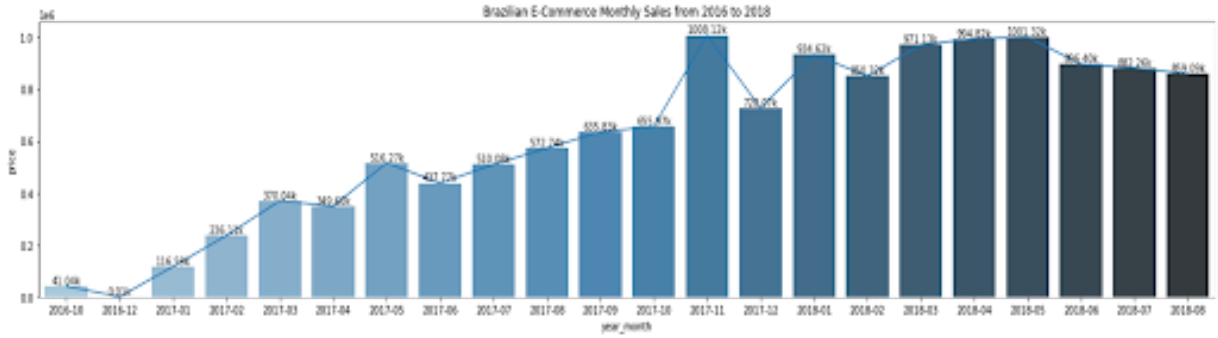

Figure 7: Top revenue generating day is black friday of 2017

- Which product from health_beauty has undergone the most price changes in 2018?

	price
product_id	
437c05a395e9e47f9762e677a7068ce7	31
921d31a1daa51460b7a95ea5f3ab64d5	13
8c292ca193d326152e335d77176746f0	9
af0a99476d96dcc1a1baa7c0d9ff6b9d	8
2fea0f2cec6b6324a277d4a61c2ed2c6	8
...	...
60c031bf1162848b7ee14f56f432285b	1
609c35bf8122d5ab8186ed7a6bfcd843	1
608f44934fbb70de5e05998ae59f4e46	1
607fc06d7055bbb7c7649810df3bc0fe	1
fff81cc3158d2725c0655ab9ba0f712c	1

2390 rows × 1 columns

Figure 8: Products from health_beauty undergoing the most price changes

- What is the average price of the product 437c05a395e9e47f9762e677a7068ce7 and how many units were sold?

	price	units_sold
year		
2018	50.037895	152

Figure 9: Metrics of the product most sold

3.4 Modeling

In this section, we have attempted to change the price of the product 437c05a395e9e47f9762e677a7068ce7 from $50.03 to a price where significantly more than 152 units would be sold.

We have considered 17 features as our independent variables and price as our dependent variable. we have split the data in the ratio of 7:3 where training data is 70% and testing data is 30%. We are using StandardScaler() to standardize both training data and testing data. We performed this to obtain a zero mean and standard deviation of 1.

We have used linear regression to predict the demand curve. The MSE obtained with linear regression is 2.79. The dataset is obtained from a real transaction log so the model is not a accurate fit but we still attempted to predict the price of a product that has undergone 31 price changes (assuming due to various reasons) in the year 2018 alone. This product (437c05a395e9e47f9762e677a7068ce7) is the most selling item from the health_beauty category which generates the most revenue for the store. The MSE of Ridge and Lasso are almost similar to linear with a slight improvement.

3.4.1 Profit Function

We have considered the following formulae to create our profit function.

$$profit = volume - cost * price$$

We are considering cost as 0 for simple calculations. We are randomly generating price points that negatively that are close to the price ranges($44-$60) of last year with some margin additions at both ends.

3.4.2 Demand Function and Demand Curve

Finding connection between demand and price of a product. We have used Linear Regression to train our data and generate a demand curve. The demand curve is generated to predict the

optimal price of product 437c05a395e9e47f9762e677a7068ce7. The intercept and slope obtained with linear regression is aiding in predicting the demand curve.

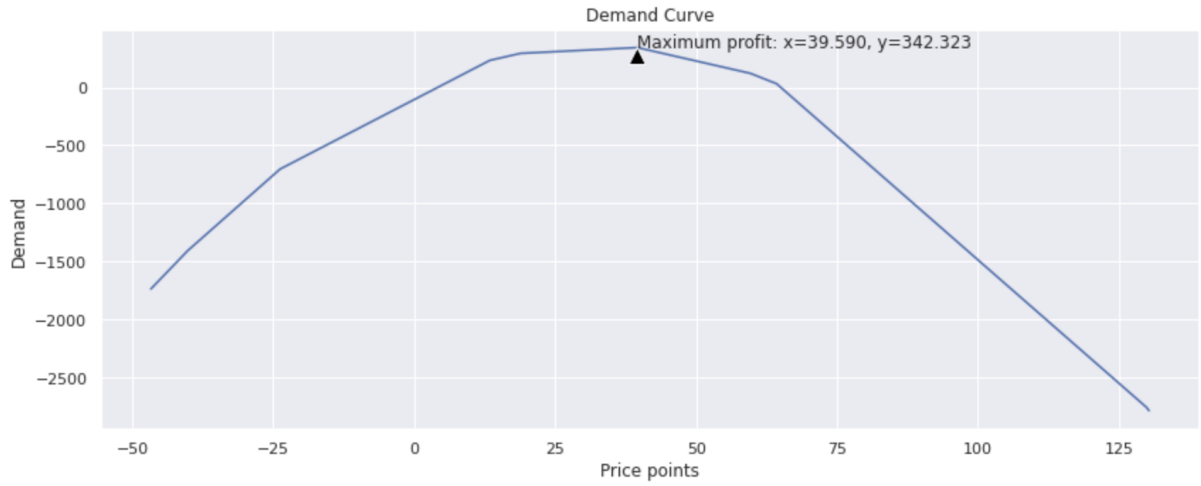

Figure 10: Optimal Price prediction of the product 437c05a395e9e47f9762e677a7068ce7

4 Results and Discussion

Based on fig. 10, if the prices are reduced to $39.59, the store is predcited to sell more than 300 units of the product. This is a 45% increase in sales volume from year 2018 of this product. We calculated the margin percentage with below formula:

$$percentage\_increase = (predicted\_revenue - 2018\_revenue/2018\_revenue) * 100$$

Calculating margin percent if we set this new price

```
revenue_in_2018 = revenue_of_product_2018['price'] * revenue_of_product_2018['units_sold']
revenue_predicted = xmax * ymax
margin_percent = (revenue_predicted - revenue_in_2018)/revenue_in_2018*100
margin_percent
```

```
year
2018    78.188228
dtype: float64
```

Figure 11: Margin percentage increase with new price

Based on fig. 11, if the demand of the product increases by 45% then the revenue is predicted to increase by 78%.

5 Conclusion

Although linear regression somewhat works on real transaction data to predict an optimal value, the price can still be optimized further and may give in better results with neural network or a hybrid of algorithms.

References

- https://tryolabs.com/blog/price-optimization-machine-learning
- https://medium.com/hamoye-blogs/unraveling-brazilian-e-commerce-dataset-e78463d77340
- https://arxiv.org/pdf/2007.05216v2.pdf
- https://www.semanticscholar.org/paper/Price-Optimization-in-Fashion-E-commerce-Kedia-Jain/69d699ca6ac62c759c6372aa86a10756c8f509ce
- https://arxiv.org/abs/2007.05216
- https://7learnings.com/blog/price-optimization-with-machine-learning-what-every-retailer-should-know/