UKRAINIAN CATHOLIC UNIVERSITY

BACHELOR THESIS

Prognostication model and pricing strategy optimization for restaurant businesses

Author:
Solomiya SHUPTAR

Supervisor:
PhD Taras FIRMAN

A thesis submitted in fulfillment of the requirements
for the degree of Bachelor of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2022

Declaration of Authorship

I, Solomiya SHUPTAR, declare that this thesis titled, “Prognostication model and pricing strategy optimization for restaurant businesses” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"If you are thinking of doing something, someone else has already done it."

My brother

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

Prognostication model and pricing strategy optimization for restaurant businesses

by Solomiya SHUPTAR

Abstract

The aim of this thesis is to study price elasticity for this type of consumer product such as confectionery, in order to find out the relationship between price and demand. Since very few companies change the price in both directions for research, we will need to simulate how the demand would change if the price will be moved by an arbitrary percentage rate.

The result of this work will be a revealing of the correct elasticity between demand and price for every product, which will allow to improve current pricing strategy and optimize the model of sales with relation to retail prices.

Acknowledgements

First of all, I would like to thank Taras Firman for his enormous contribution to this work and to the research I conducted and described in this work.

Also, I want to thank my parents, friends and just loved ones for their constant support in difficult times and for the fact that they are always with me…

Contents

List of Figures

List of Tables

List of Abbreviations

HORECA	**HO**tel **RE**staurant **CA**tering
ML	**M**achine **L**earning
AI	**A**rtificial **I**ntelligence
OLS	**O**rdinary **L**east **S**quares
RF	**R**andom **F**orest

Dedicated to the people who keep strong and continue working in, creating, investing and developing Ukrainian business…

Chapter 1

Introduction

1.1 Context

In a modern world, success of the business is determined by a big pull of essential factors which make a harmonious working system: from initial business model to the perfect organization and continuous improvements on all the company levels.

When a business wants to maximize its profit, the first steps to take are basics such as cost-cutting, increase in production, etc. But after elementary actions are taken, the search for more sophisticated methods begins.

In the past years more and more companies address information technology specialists to help them bring the business to the new level. Not without a reason, as seen in practice, recent advances in machine learning have had a great impact on maximizing business efficiency in almost all industries.

For instance, the marketing and sales industry has greatly benefited from the advance of machine learning technology. Specifically, the success of a modern e-commerce company in a fastly-growing online retail field can largely depend on the performance of ML programs. That is, ML algorithms are frequently introduced on the company websites in order to capture and analyze the preferences of their customers. This way, businesses can provide personalized recommendations to their website users, which in turn might increase both the sales and customer retention rate. Indeed, according to Forbes, 57% of enterprise executives believe that the improvement of customer support and experience will be one of the most important growth benefits of AI and ML.[8] Another field where machine learning has demonstrated significant impact is logistics. In fact, machine learning is probably one of the critical components of supply chain management. ML algorithms enable companies to improve the efficiency of carrier selection and to optimize the routing schedules. Essentially, the ability of ML algorithms to operate with huge amounts of retail data in little time simplifies the business processes and promotes cost reduction for a company.

Machine learning has also made its way into the manufacturing industry. By tackling the complex inventory management and predictive maintenance processes, ML algorithms automatically reduce the inefficiencies and potential errors in company operations.

The afore-mentioned industries do not comprise a complete and exclusive list of industries, which benefited from the advance of machine learning. In fact, the list extends beyond, including such industries as finance, government, and even healthcare. In food industries machine learning methods had already shown big expectations in terms of dealing with food waste.

At the same time, despite the advantages that come with incorporating ML systems, there are no representative statistics or data about its methods being implemented in practice, which demonstrates agricultural and HORECA industry reluctance toward the integration of machine learning into its business practices.

1.2 Problem

While more and more companies integrate machine learning as a part of their sales analytics, there are some industries which are far from implementing such approaches in their businesses. To be specific, HORECA industry is lacking systematic and accurate technological analysis of their financial results.

At the same time, sales analytics, powered by machine learning technology, can play a significant role in the financial success of an organization. That is, the analysis of sales data can uncover the potential problems in the business strategy of a particular company. Specifically, the sales analytics can indicate some flaws within the pricing strategy of a business. Using the results, companies might then decide to make changes to its existing pricing plan or devise new business practices to increase their sales and revenue. If successful, organizations can decide to keep the new strategy, and if not, they can continue experimenting. Therefore, sales analytics enables companies to continuously monitor and enhance the existing business practices, including their pricing plan, as well as adapt their business model to changing market circumstances over time.

In fact, It is very common to perform different types of analytics over data, including sales analytics, in most online retail businesses. Statistics demonstrate that more than 28% of companies in the retail industry incorporated ML into their technology systems in some way in 2018, which is a 600% increase compared to the 2016 statistics. These dynamics are expected to hold in the future as well. In addition, only 26% of implemented ML technology in retail companies directly interacts with clients, while the remaining 74% are dedicated to the internal company affairs, such as data analytics.[2]

One of the reasons explaining the prevalence of machine learning in retail is a big advantage of organizations in this industry to accurately collect huge amounts of data on metrics of interest. This way, retail companies can analyze their sales data and experiment with their prices instantaneously in order to study its influence on sales. On the other hand, restaurants do not have the privilege to change their prices quickly. Pricing strategy in HORECA business is commonly based on the cost price of a product itself, the labor cost, the logistics cost etc. The retail price is then determined by designating a fixed percentage rate to the combined cost value (e.g. 30% of retail price). As a result, the mechanics of this pricing strategy and the low variability of prices prevent the businesses in HORECA industry from accurately calculating the demand elasticity and, in turn, facilitate the optimal pricing strategy.

1.3 Aim and Goals

In this project, we focus on studying the interdependency between price and demand specifically in the HORECA industry. The aim of this project is to determine the actual price elasticity of various products offered by a café, as it could have been if the cafe experimented with their prices. As stated in Section 1.2, restaurants do not generally experiment with prices to determine the price-demand relationship in practice. In this work, we try to simulate how the demand would change if the price moved by an arbitrary percentage rate.

More precisely, we simulate the modified prices in the vicinity of the actual retail price at the moment. To support the trend from the real data, we also adjust the sales with respect to the simulated prices and explore the appropriate interconnection.

The result of this work will be a revealing of the correct elasticity between demand and price for every product, which will allow to improve current pricing strategy and optimize the model of sales with relation to retail prices.

Chapter 2

Background

This chapter will explain in general what pricing is and why this concept should not be neglected when creating or running your own business. We will also talk about the interdependence and sensitivity of price to demand and why not all sales data are ready for analysis.

2.1 Pricing

One of the primary roles of any business is to create value. By offering a product or service, whose value to customers exceeds the cost of production, companies are able to make money and stay profitable. In the world of business, the value of a product or service is captured in a price tag, which in turn is determined by the company itself. In fact, pricing is one of the critical components of any business strategy and has a strong influence on the financial performance of any organization.

On one hand, the price set for a product or service should encapsulate the costs inflicted by the production and maintenance of that product or service, so that the companies can break-even. However, in order to make profit out of the sale of their products or services, companies should set their prices to exceed the amount of induced costs - place a profit margin, in other words. On the other hand, the price of a product or service should coincide with the amount that a customer is willing to pay for it, so that the transaction can happen. Otherwise, the company would not be able to sell its product and thus would receive no revenue. Therefore, the question of the right price is one of the critical determinants of organizational success.

Indeed, the 1992 study by Michael Marn and Robert Rosiello, the senior pricing analysts at McKinsey and Company, demonstrates the extent to which the correct pricing can impact business performance. By examining the unit economics of 2,463 companies, they revealed that a 1% improvement in price results in an 11.1% increase in operating profit. Furthermore, the similar improvements in other business aspects did not reach the same level of influence on the operating profit, according to their results (see Figure 2.1).

At the same time that the right price can make your business profitable, the wrong price on the other hand can ruin your work. Setting the right price for your product can be a daunting task and many executives decide on it based on their intuition. However, the human factor is often wrong and this practice may not be reliable and harmful for a particular business. In addition, you have competitors who also play with you in this market and the inflated price you have can lead to the fact that your customers are simply lured. That is why in today's world it is worth trying new methods of analysis of pricing policy and trust mathematics more than "your senses".

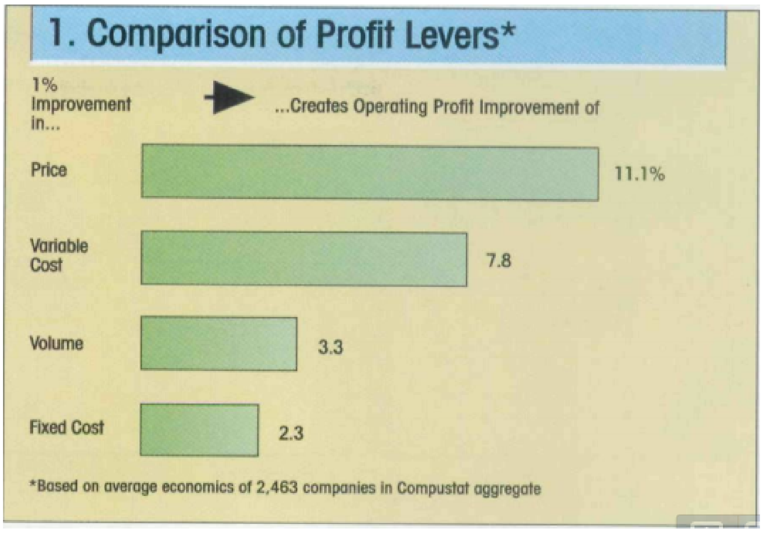


FIGURE 2.1: Comparison of Profit Levers [3]

2.2 Price Elasticity of Demand

Price elasticity of demand is a measurement of how sensitive consumption is to changes in price. In other words, the price elasticity gives the percentage change in quantity demanded when there is a one percent increase in price. The more negative elasticity is - the more decline in demand is expected after price increase.

$$e = \frac{dQ/Q}{dP/P} \tag{2.1}$$

For business, this concept carries one main message - "How much we can sell if we increase the price by n%?". From the elasticity formula we can derive that quantity sold equals price change multiplied by elasticity value. Empirically, demand can be approximated as an exponential function of price.

$$q(p) = const * p^e \tag{2.2}$$

We solve this equation by logarithmization to have a simple equation of kind:

$$\log q(p) = \log(const * p^e) = \log(const) + e\log(p) \tag{2.3}$$

As a result of log-log simplification the elasticity becomes a simple coefficient in a linear regression equation.

2.3 Causal Inference and Confounders Problem

In the real world we face more interesting challenges when working with sales data. Main goal of analyzing sales data is to discover and study causal effects of parameters in relationship with quantity sold. Naive approach would be to include price as the only factor to influence sales. But estimating causal effects from observational data is difficult because of confounding.

In statistics, a confounder is a variable that influences both the dependent and independent variables. In a simple relationship between price and quantity sold there is a pull of different confounders (not specified on a graph) such as product quality, location of a cafe, holidays, etc. (see Figure 2.2).

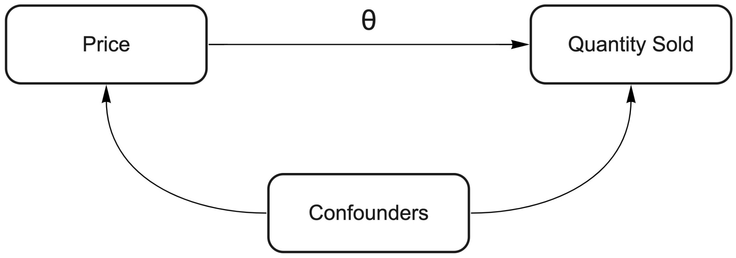


FIGURE 2.2: Relationship between Price, Quantity and Confounders

Consequently, ignoring or misinterpreting the effect of confounding variables on price can significantly bias the estimate of θ (estimate of price on quantity). Moreover, it leads to wrong conclusions about demand reaction to price and therefore to wrong pricing strategy.

Chapter 3

Technical Background

3.1 Double ML

At the beginning, it is worth mentioning that there are many statistical and machine learning methods that might be used for this problem and show good performance. Machine learning has been a matter of choice especially for prediction purposes. However, its methodologies show great effectiveness in practice, only if chosen correctly. Double ML is a relatively new approach introduced in 2016 which exactly addresses the problem of confounders. Not all ML methods can accurately estimate the effect of causal parameters, which eventually leads to wrong business decisions.

$$Y = D\theta_0 + g_0(Z) + U, \quad \mathbb{E}[U|Z, D] = 0, \tag{3.1}$$

$$D = m_0(Z) + V, \quad \mathbb{E}[V|Z] = 0 \tag{3.2}$$

Double ML considers a partially linear model where Y is the outcome variable, D is the policy/treatment variable of interest, Z is a vector of other covariates or "control" variables, and U and V are disturbances. (equation 3.1).

In the partially linear model parameter D has linear relationship with outcome variable Y, but control variables Z might have nonlinear effect on Y, therefore are modeled via the function $g_0(Z)$.

Yet the equation 3.3 describes the dependent relationship of treatment variable D on covariates, in other words, those confounding variables. The confounding variables Z influence treatment variable D with the function $m_0(Z)$ and the outcome variable via the function $g_0(Z)$.

The goal of Double ML method is to find root-n consistent estimate for θ.

Let $X_1, X_2, ...$ be a sequence of iid RVs drawn from a distribution with parameter θ and $\hat{\theta}$ an estimator for θ. We say that $\hat{\theta}$ is consistent as an estimator of θ if $\hat{\theta} \rightarrow \theta$ in probability or

$$\lim_{n \to \infty} (P|\hat{\theta}(X_1, ..., X_n) - \theta| \leq \epsilon) = 1 \quad \forall \epsilon > 0 \tag{3.3}$$

Which, in other words, states that θ is consistent if $T_n = \frac{X_1 + ... + X_n}{n}$ converges in probability to θ.

$\sqrt{n}$-consistent ("root n consistent") describes how quickly θ converges to $\hat{\theta}$. Estimator is root-n consistent if $T_n - \theta = O_p(\frac{1}{\sqrt{n}})$.

3.2 Problem of Standard Methods

As mentioned earlier, when dealing with confounders the biggest problem is the bias in learning $g_0(Z)$. Running simple Random Forest regression to estimate $g_0(Z)$

and OLS regression to estimate θ does not solve the problem. In fact, the Figure 3.1 explains the bias.

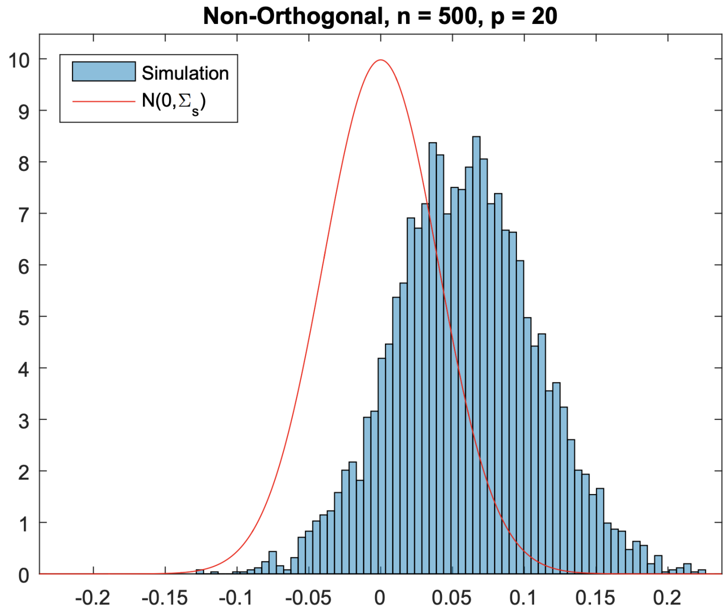


FIGURE 3.1: Behavior of a conventional (non-orthogonal) ML estimator $\hat{\theta}_0$ in the partially linear model [4]

By definition of consistency (Theorem 1) and Chebyshev's inequality we know that consistent estimator is the one where $P(|\theta - \hat{\theta}| > e) \to 0$. However, in Figure 3.1 we see behavior of a conventional estimator $\hat{\theta}$ in an experiment using just simple random forest to study g_0. The graph displays distribution of $\hat{\theta} - \theta_0$ vs normal distribution with mean 0. Normal distribution with mean 0 assumes that the bias of the estimator is negligible.

It is noticeable that the value is strongly biased, since its distribution is shifted to the right compared to the normal distribution shown with the red curve, which proves the inefficiency of the usual method when estimating $\hat{\theta}$ and $g_0(Z)$ in a partially linear model.

The bias is caused by two main reasons - regularization and overfitting. Double ML fights both of them.

3.3 Orthogonalization and the Neyman Orthogonality Condition

The underlying concept of orthogonal-ML is explained in Frisch–Waugh–Lovell theorem.

Assume we have a resposnse y and two data matrices X_1 and X_2. On one hand, we could perform ordinary least squared(OLS) of $y \in \mathbb{R}^n$ on $X_1 \in \mathbb{R}^{n \times p_1}$ and $X_2 \in \mathbb{R}^{n \times p_2}$ to get coefficients $\hat{\beta}_1 \in \mathbb{R}^{p_1}$ and $\hat{\beta}_2 \in \mathbb{R}^{p_2}$. (The implicit model here is $y = X_1 B_1 + X_2 B_2 + u$).

Alternative approach is:

1. Perform OLS of y on X_1 to get residuals $\tilde{y}$.

2. Perform OLS of X_2 on X_1 to get residuals $\tilde{X_2}$.

3. Perform OLS of $\tilde{y}$ on $\tilde{X}_2$ to get coefficient vector $\tilde{\beta}_2$.

The Frisch–Waugh–Lovell theorem states that $\hat{\beta}_2 = \tilde{\beta}_2$.

In case of partially linear model (equation 2.1 and 2.2), the intuition supported by the Frisch–Waugh–Lovell theorem is the following:

1. Predict D based on Z using ML.
2. Predict Y based on Z using ML.
3. Linear regression of the residuals from step 2 on the residuals from step 1, for getting an estimate of θ_0.

This algorithm directly partial out the effect of Z from both Y and D.

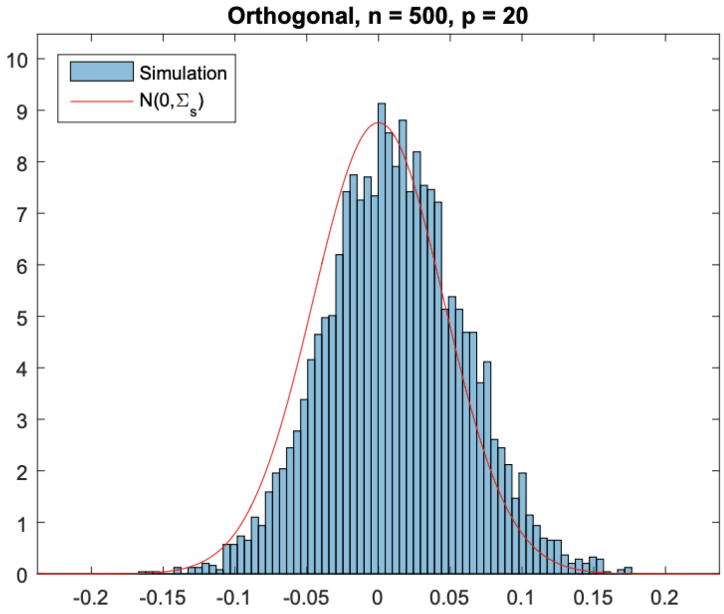


FIGURE 3.2: Behavior of an orthogonal double ML estimator $\hat{\theta}_0$ in the partially linear model [4]

As the result, we get an unbiased estimate $\hat{\theta}$. The distribution of unbiased $\hat{\theta}$ - θ_0 vs normal distribution is illustrated on the Figure 3.2.

Chapter 4

Related Works

There are several works related to the concepts applied in this project. The study by Chernozhukov et al. describes the double machine learning technique for treatment and causal parameters [4], which is also used in this work. In particular, Chernozhukov et al. claim in their study that the conventional application of machine learning algorithms for prediction tasks possesses several drawbacks, including the fact that they do not necessarily deliver accurate estimators of causal parameters despite the good prediction accuracy overall. Even more specifically, the authors mention that estimators resulting from the naive application of ML algorithms might suffer from the lack of convergence in relation to the sample size due to the regularization bias. Nevertheless, Chernozhukov et al. suggest that the synthesis of primary, as well as auxiliary prediction tasks might serve as a solution to the problem, and call this technique ‘Double Machine Learning’. Later in the work, the authors apply the double machine learning algorithm to the data: they evaluate the effect of 401(k) eligibility on accumulated financial assets, additionally applying such sampling-splitting methods as K-fold cross-validation along the way to avoid the problem of overfitting.

A different study by Lars Roemheld applies the double machine learning technique specifically to the problem of elasticity pricing [7]. In particular, the author first describes in his work the importance of price elasticity in the pricing strategy of a business, and later discusses the structure of the double machine learning algorithm. Roemheld also brings up the problem of confounding and explains why it complicates the analysis of historical data and undermines the causal inference tasks. The author then applies double ML to the product sales dataset in order to calculate the price elasticity of demand of retail store offerings: every data entry includes such information as the product ID number, the date of the sale, the product name, the day of the week, the number of days a product was in inventory, as well as the quantity sold and price of the product.

Guy Lebanon in his work puts a special emphasis on the discussion of the consistency of estimators in prediction tasks [6]. Specifically, he addresses the problem of divergence in estimators, which leads to inconsistent results with increasing sample size. Lebanon also includes some mathematical strategies, which can be applied to prove the consistency of estimators.

Chapter 5

Dataset

This chapter describes the dataset used in the experiment. This is a real time-series data downloaded from the accounting software system of one of the largest confectionery chains in Lviv bakery-confectionery SHOco..

5.1 Description

The preliminary request for data included metrics for quantity of sales, prices for respective products, its cost, revenues, unique ID etc.. Time-series data is collected in a time period between 1st of June 2019 and 16th of April 2022. Data is collected daily for a pull of different products, so the information is stored in one csv-file for a respective day. These files were combined into one large data-frame with date as a parameter. As a result, final dataset consists of the following columns described in the Table 5.1.

5.2 General Exploration

5.2.1 Sales Dynamics

Having combined all the files into one big time-series dataset, we start exploring data characteristics. First of all, we need to deal with data cleaning and preparing it for further analysis. Hence, the following facts are taken into consideration:

1. There are 5836 cases where RetailPrice is absent, so those records are omitted.
2. Number of unique GoodId-s is 393 which is quite a big collection of products to be analyzed. Therefore, we can expect a big difference in metrics across those products.
3. However, every product has its GoodParent. There are 14 unique good 'parents' in total.

In the case of 393 unique goods, GoodParent is a more suitable feature to start explorational analysis. This way we can determine which products are most in demand, and assess the sales dynamics and select data for more detailed analysis.

The core product of the bakery-confectionery SHOco. is indeed pastry as visualized on the Figure 5.1. In total sales during 2019 - 2022 make up 24 million UAH. Next two vectors of products to be considered are desserts and eclairs.

In fact, those three product groups were leading every year making most of the sales. As we can observe on the Figure 5.2, 2021 year was also a year of big improvements and record high revenue.

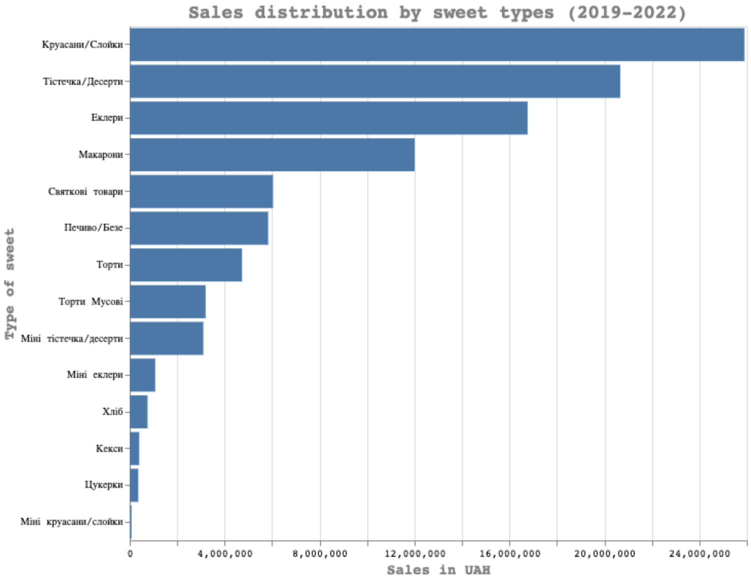


FIGURE 5.1: Sales Distribution by Good's Category (GoodParent) in the time period 2019-2022

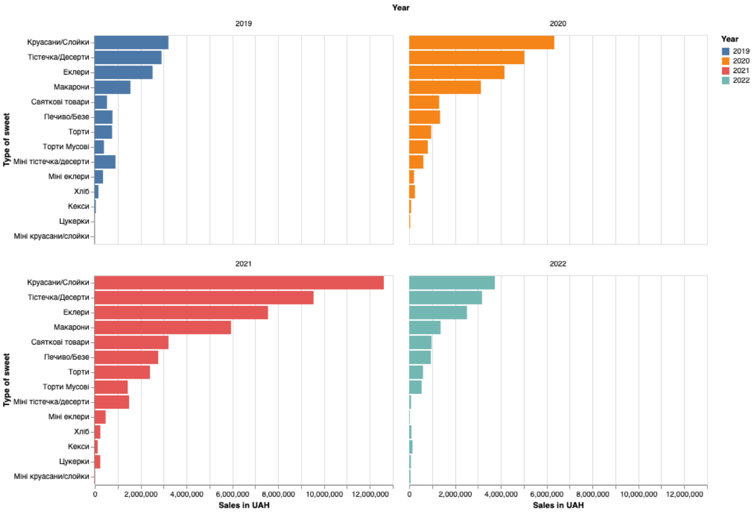


FIGURE 5.2: Sales of Goods in each year

	Column name	Column Description	Type
1	Date	Date in format yyyy-mm-dd when metrics were recorded	Categorical
2	GoodId	Unique ID of a product	Categorical
3	GoodName	Name of a product	Categorical
4	GoodParent	Category of a product	Categorical
5	RetailPrice	Actual price in the moment of sale	Numerical
6	Department	One out of 4 cafe locations	Categorical
7	Quantity	Quantity sold	Numerical
8	SumOfSale	Revenue from sales of a particular product	Numerical
9	Cost	Cost associated with a product	Numerical

TABLE 5.1: Description of the Attributes in the Bakery's Dataset

Moreover, as our data is available until 16th of April 2022, to analyze objective total performance, we build Figure 5.3 with comparison between first four month of 2021 and 2022.

In this year SHOco. started to show even greater results compared to 2021, although a big downgrade in sales after February 24th, the total sum of sales by April 2022 has increased by 23%.

5.2.2 Sales Distribution by Departments

In the available data there are 6 unique departments where SHOco. goods are being sold. However, two of them are negligible as those are other restaurants which buy SHOco. products by partner prices, therefore accurate and final information about sales is not stored within SHOco. accounting system. Later on only 4 of SHOco. own locations will be considered in the modeling, due to the fact that others do not generate enough sales data to take them into account

5.2.3 Discovery of Quantity Sold and Respective Retail Prices

Before actually diving into quantity and price modeling, we analyze the overall statistics on those two metrics. On the Figure 5.5 we observe that the biggest turnover is inherent in the products with prices up to 100 UAH, which actually equals the

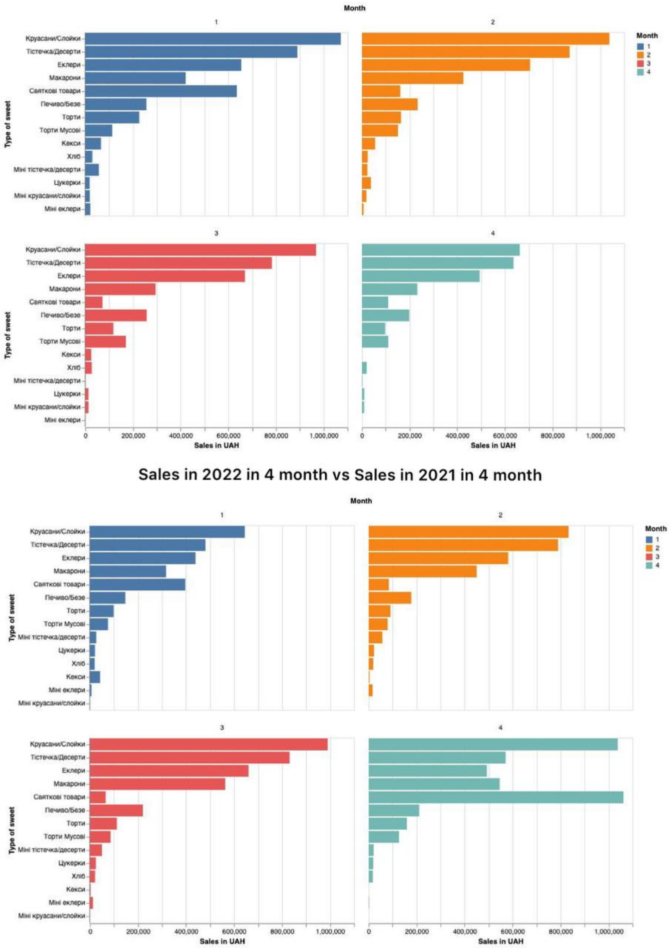


FIGURE 5.3: Sales of Goods 2022 vs 2021 during first four months

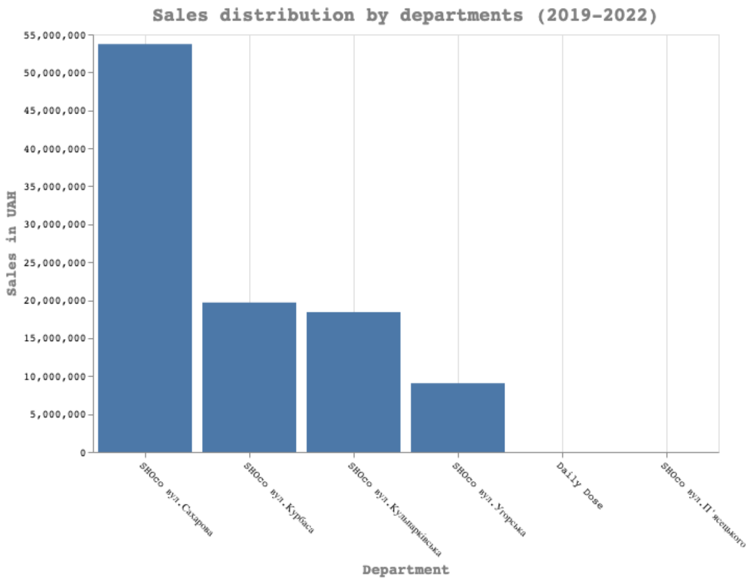


FIGURE 5.4: Sales of Goods by departments

mean price of all pastries and desserts in SHOco. As mentioned above, pastry and desserts are the most sold goods in the cafe.

However, it cannot be considered as correlation or causal effect between price and quantity sold, as we haven't done any appropriate modeling yet.

The Figure 5.5 illustrates that RetailPrice deviation is very high, therefore, to achieve greater effectiveness one should concentrate on core products of this particular business.

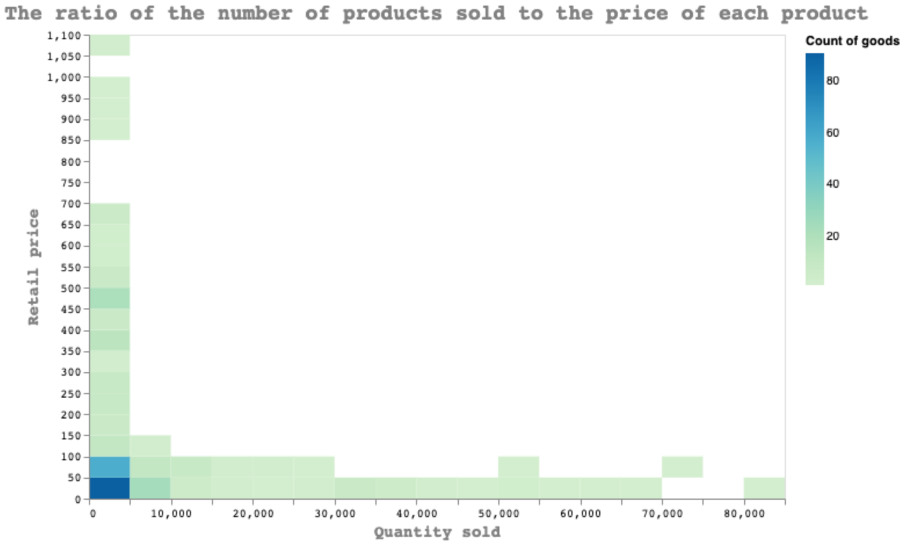


FIGURE 5.5: Sales of Goods relative to their prices

Chapter 6

Quantity Sold Modeling & Price Simulation

As discovered before, there are only some groups of the products worth modeling, due to the availability of sufficient data, etc. Therefore, we group our dataset by GoodId and for every goodId sum over quantity sold during the day and and take the retail price. Considering products with a high turnover, we select only those with more than 1000 records. Thus, the number of products selected for further analysis and modeling is 30 - top products of SHOco..

6.1 Quantity & Retail Price behavior

As we noted in the problem description, the main problem of the restaurant business in optimizing pricing and determining the correct elasticity, is the lack of data on price. In particular, when speaking about the lack of data, we mean that restaurants almost never experiment with the prices close to the initially set one, during the sales period.

This is exactly the situation for each of the products in the dataset. We observe high volatility in sales and almost a static price, which was apparently just revised quarterly at best. Accordingly, the only change in price that we can observe is an increase in the price of the product by several UAH in accordance with the increase in cost.

With the example of two products (Mohito Croissant and Coconut Eclair) on the Figure 6.1 and Figure 6.2 we illustrate the behavior of quantity sold and its retail price respectively.

Of course, with such price dynamics, it is impossible to conduct a correct analysis of elasticity and review pricing strategy.

6.2 Trend, Seasonal and Residual Components of Quantity

Time series data can exhibit a variety of patterns, and it is often helpful to split a time series into several components, each representing an underlying pattern category[5]. Assuming additive decomposition the equation looks like

$$y_t = S_t + T_t + R_t, \tag{6.1}$$

where y_t is the outcome variable, S_t is seasonal component, T_t is trend component and R_t is residuals, all at period t. Reasonable, multiplicative decomposition has the equation

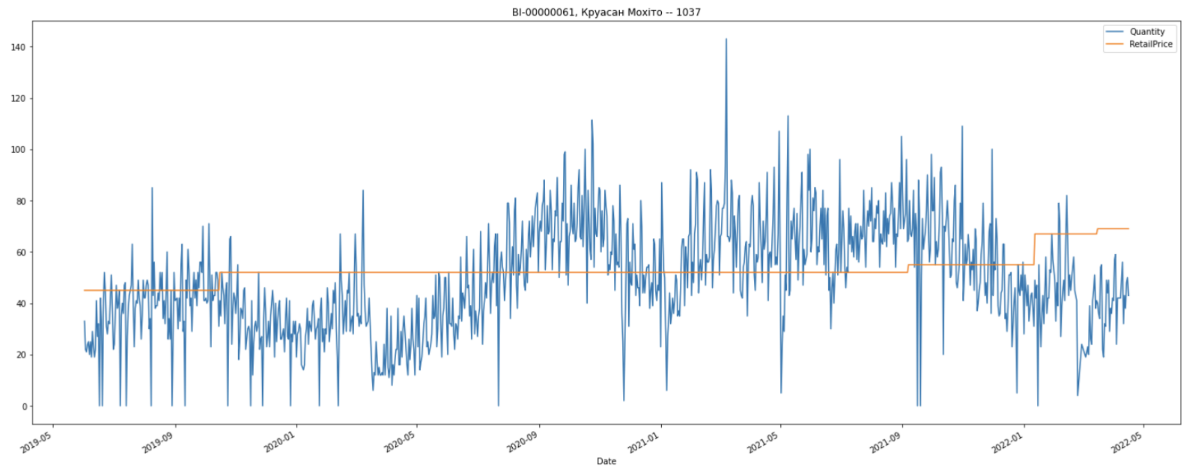


FIGURE 6.1: Quantity vs Price time-seris behavior for Mohito Croissant

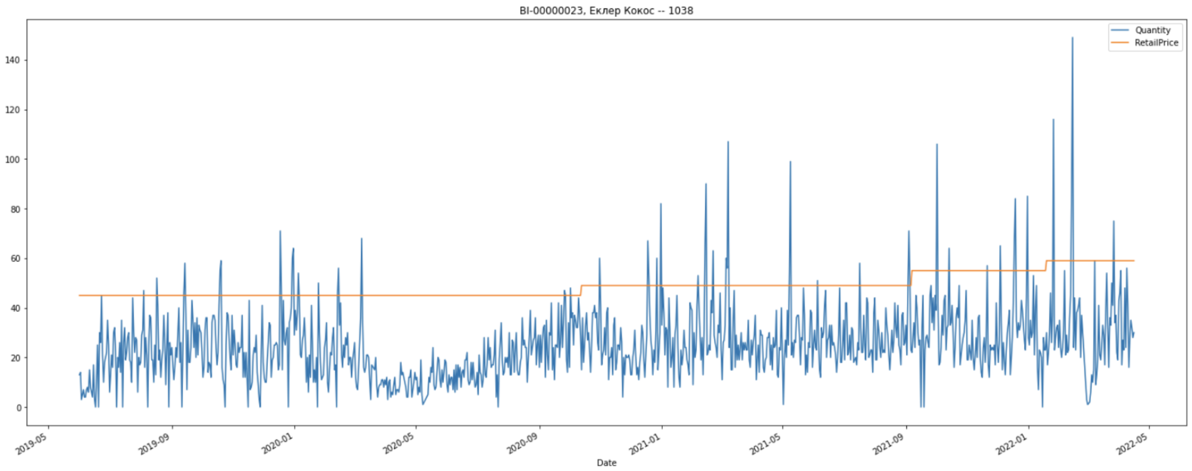


FIGURE 6.2: Quantity vs Price time-seris behavior for Coconut eclair

$$y_t = S_t * T_t * R_t \tag{6.2}$$

An additive model is used when changes in the magnitude of the seasonal fluctuations, or the variation around the trend-cycle are linear over time. Multiplicative model usually suits economic time-series data, where factors increase or decrease their impact over time. A multiplicative model is nonlinear, such as quadratic or exponential. That is exactly the case for SHOco. data. As a result of seasonal decomposition on raw quantity data, we can extract seasonal, trend and noise components from a data array and study its behavior. On the example of products from the previous figure, I will also show a clear example how seasonal decomposition performs in our case (Figure 6.3).

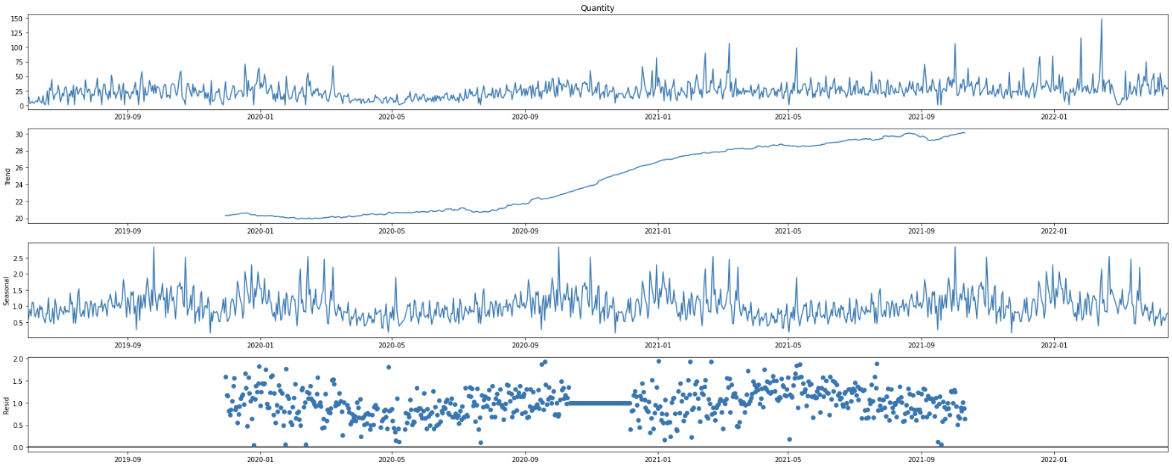


FIGURE 6.3: Seasonal decomposition of Quantity data (Coconut Eclair)

On the figure we clearly see a bright pattern of seasonality, with sales picks in October, falls in late November and declining sales during the summer. According to the directors' insights, they also observe a decrease in consumption of pastries and sweets during the summer period. Dramatically increasing trend starting from 2021 also proves our observations at the initial stage of the study, where we clearly saw significant changes in sales in 2021 (back on Figure 5.2).

6.3 Linear Regression model on Quantity. Predict trend

6.3.1 Seasonally adjusted data

If the seasonal component is removed from the original data, the resulting values are called "seasonally adjusted" data. For an additive decomposition, the seasonally adjusted data are given by $y_t - S_t$, but for multiplicative decomposition we can extract seasonality component by division y_t / S_t .

It is useful to remove seasonal component, because otherwise we cannot study the variation in data caused by other factors. Consequently, to analyze non-seasonal variation SHOco. quantity data should be seasonally adjusted.

However, trend and residual components are still included in the data array.

Figure 6.4 shows how quantity looks after being seasonally adjusted.

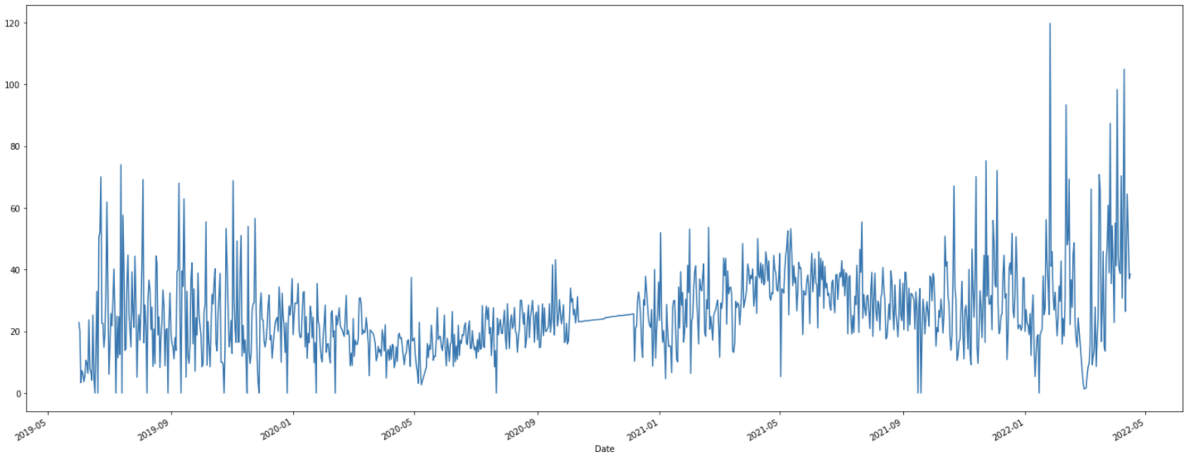


FIGURE 6.4: Seasonally adjusted quantity data for coconut eclair

6.3.2 Linear regression model

Linear regression is a statistical tool used to help predict future values from past values. The great advantage of regression models is that they can be used to capture important relationships between the forecast variable of interest and the predictor variables. In our case we use linear regression to predict the trend of quantity sold, in other words, get the closest trend line possible to the real values of quantity.

Again, we perform linear regression for every product separately and collect the trend predicted by the model for further usage. This type of regression is performed on seasonally adjusted data of quantity. Regression line captures the main trend as seen on the Figure 6.5 , however, it is obvious that we can't expect highly accurate score results from such a model.

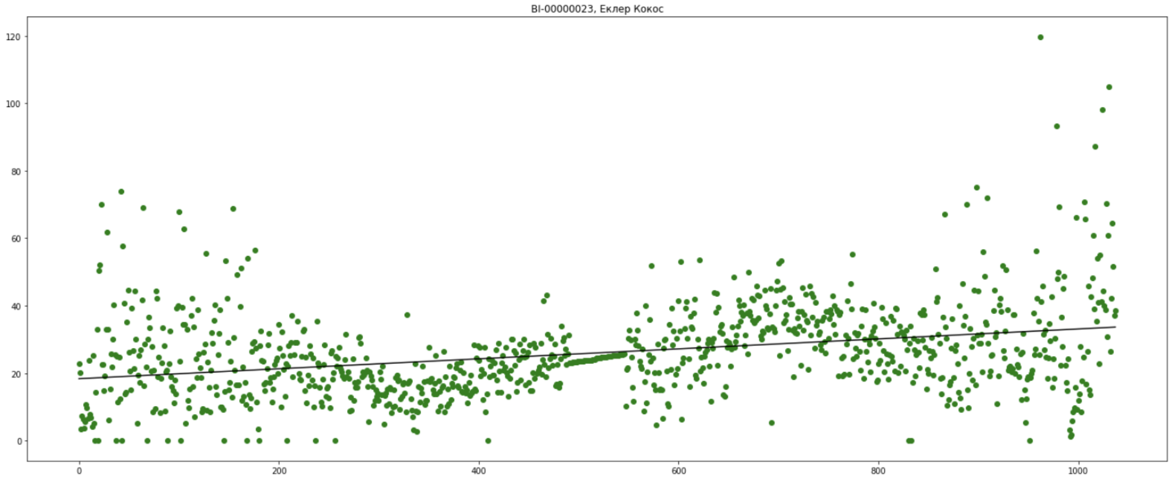


FIGURE 6.5: Linear regression results. Prediction line and seasonally adjusted quantity values

6.4 Price Simulation

On the next step we are dealing with the problem of static price. Our solution is to simulate a price change in increments of up to 5 UAH around the real price at that moment. Of course, this decision is not complete and correct without next steps that need to be implemented to preserve the logic and integrity of our initial data. To be precise, we need to account for new residuals adjusted to the new generated price.

It is a crucial point, because our goal is to make this simulation as natural as possible. Further in the work, those adjusted residuals will be used to adjust the quantity sold.

Residuals are obtained via fragmentation of seasonal component and trend taken from regression in a previous step, not just a simple decomposition. To adjust the residuals we multiply them by exponential function:

$$y_t = e^{10\alpha(1-x_t)}, \quad x_t = \frac{\hat{p}_t}{p_t}, \quad \alpha \in [1.5, 1.9] \tag{6.3}$$

where $\hat{p}_t$ is generated price and p_t is a real price at time-period t.

The Figure 6.6 shows the difference between relationships of of generated prices vs residuals and original retail prices vs residuals.

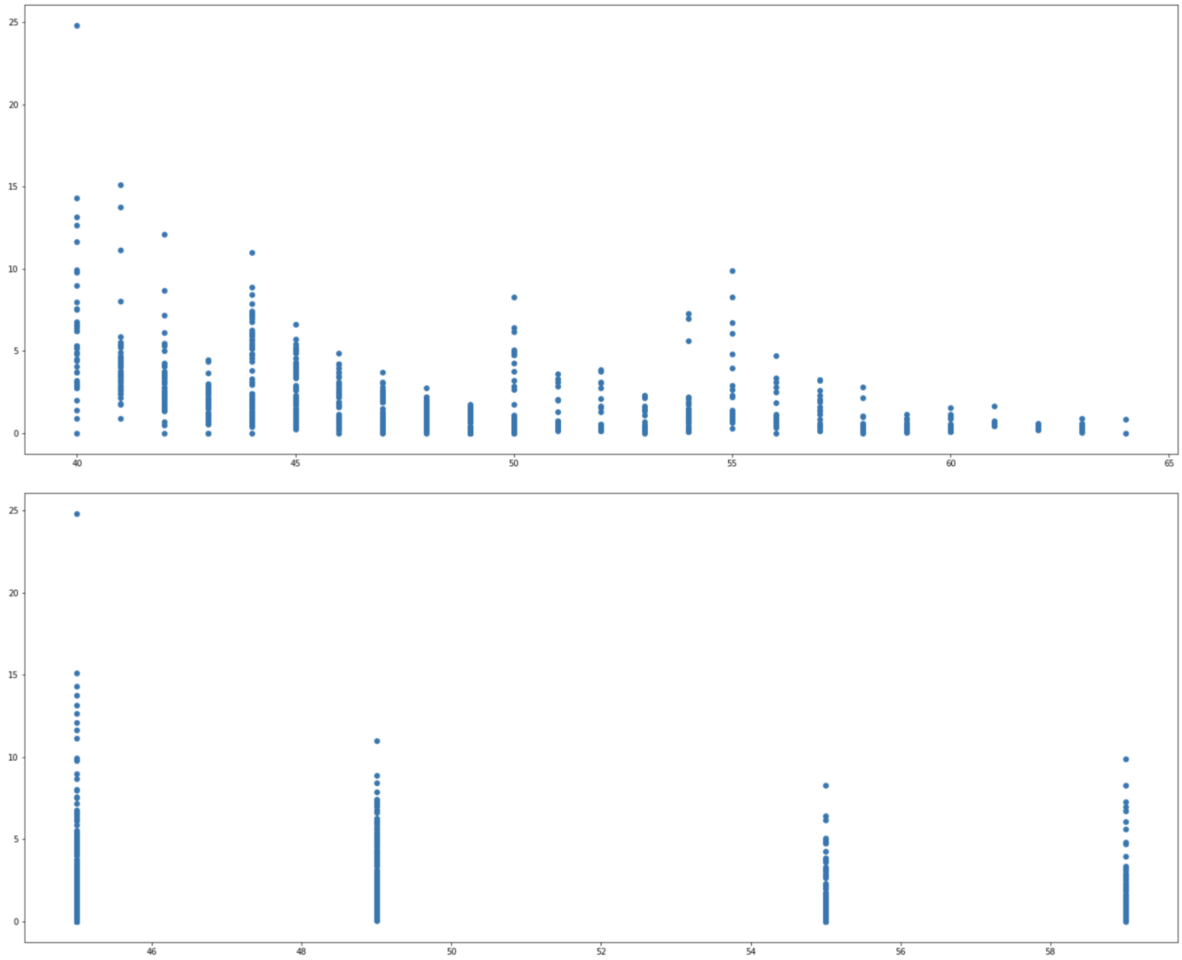


FIGURE 6.6: Plot of generated prices vs residuals and original retail prices vs residuals

Now we can see exactly what price we managed to generate compared to the original, on the example of such a product as vanilla eclair. The orange line, as we noted earlier, shows changes in the original price, and as we see it has changed frequently, as opposed to the blue line, which we neglected for further study (see Figure 6.7).

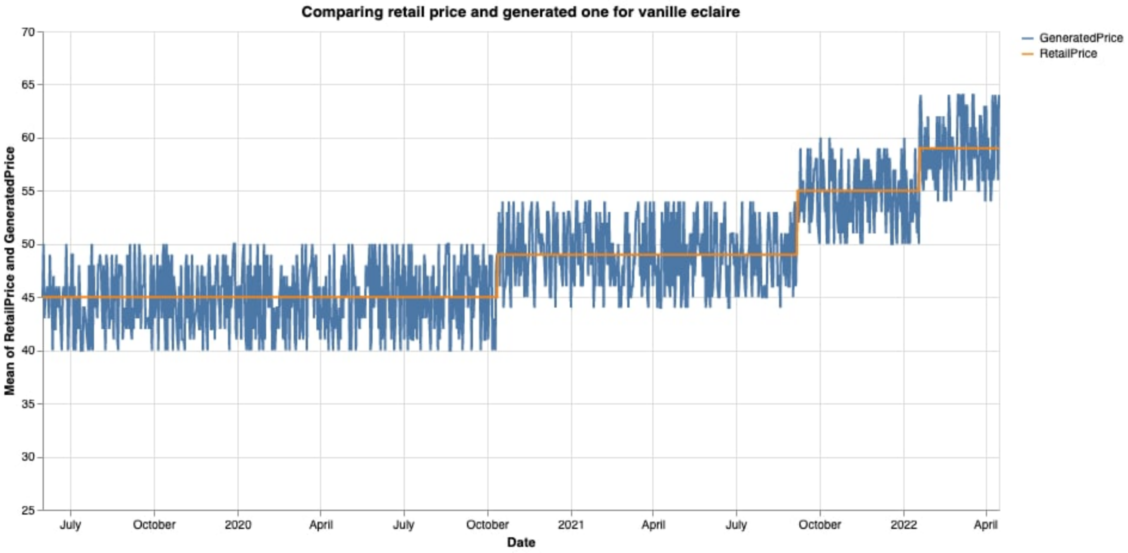


FIGURE 6.7: Original prices vs Generated prices

Chapter 7

Double ML Implementation

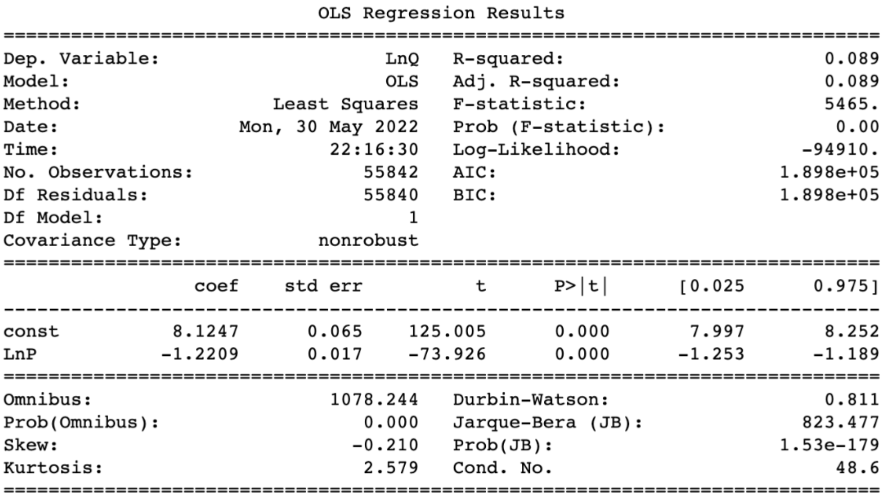
7.1 Feature Engineering

For the ML model it is important to have a pull of features which influence the output variable. Therefore, we try to extract as many factors as possible from the available data.

Moreover, in our problem, those factors may be the confounders hidden in the data, which influence both outcome variable quantity and price, so for us it is important to take them into consideration. Some of the new important factors included and their description is commented in the table (see Table 7.1).

7.2 Naive OLS on Quantity with Price

Let's perform naive OLS regression on log quantity (LnQ) with LnP as a predictor to see the results as a proof that conventional methods do not work with this problem. Figure specifies the outcome of OLS Regression with quite disappointing R-squared. The naive elasticity estimate corresponds to $\hat{\beta}_1$ coefficient, in this case the coefficient of LnP which equals -1.2209 (see Figure 7.1).



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    LnQ   R-squared:                       0.089
Model:                            OLS   Adj. R-squared:                  0.089
Method:                 Least Squares   F-statistic:                     5465.
Date:                Mon, 30 May 2022   Prob (F-statistic):               0.00
Time:                        22:16:30   Log-Likelihood:                -94910.
No. Observations:               55842   AIC:                         1.898e+05
Df Residuals:                   55840   BIC:                         1.898e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          8.1247      0.065    125.005      0.000       7.997       8.252
LnP           -1.2209      0.017    -73.926      0.000      -1.253      -1.189
==============================================================================
Omnibus:                     1078.244   Durbin-Watson:                   0.811
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              823.477
Skew:                          -0.210   Prob(JB):                    1.53e-179
Kurtosis:                       2.579   Cond. No.                         48.6
==============================================================================
```

FIGURE 7.1: OLS Regression Results

	Feature name	Feature Description
1	QuantitySold	New quantity values adjusted from predicted quantity adjusted on seasonality and new residuals
2	GeneratedPrice	Simulated price
3	LnQ	Logarithm of Quantity Sold
4	LnP	Logarithm of GeneratedPrice
5	Seasonality	Seasonal component of quantity
6	month	Number of month
7	DoM	Day of the month
8	DoW	Day of the week
9	department_avg_sold	Average sold quantity by department
10	department_avg_price	Average retail price(generated) by department
11-17	Mon/.../Sun	Dummy variables for days of the week
18	is_holiday	Boolean variable indicating the holiday (taken by Polish calendar)
19	GoodIdEnc	GoodId Encoded with Label Encoder
20-23	Department_Sakharova/...	Dummy variables for the department

TABLE 7.1: Description of the Features Used

7.3 Orthogonalization of Quantity and Price

Revising the orthogonalization approach explained in the technical background section, we now implement the core and most important idea of the Double ML approach. We orthogonalize Quantity on the pull of possible confounders and orthogonalize Price on the same pull of factors.

7.3.1 Random Forest regression on Quantity

For the RF regression we split the data in train, test and validation samples to avoid overfitting. With a pull of above mentioned features and fitting the regression on different samples, we get the following results illustrated on the Figure and Figure explaining the feature importance.

To assess the impact of a feature, we used the SHAP method, which calculates Shapley values from coalitional game theory. In short, this method explains how much a single feature has affected the prediction of our model.

Let's move on to the results of assessing the importance of features. It is very simple, a feature that has a more absolute SHAP value and is more important. The first graph is informative only in the sense that it allows you to see which features had the greatest impact, but it does not show the positive or negative side of this feature (see Figure 7.2). We can consider a simple example from the day of the week, namely Monday. SHOco knows that statistically the lowest number of sales occurs on this day, and looking at the first chart, you might think that this feature should be present because of its importance.

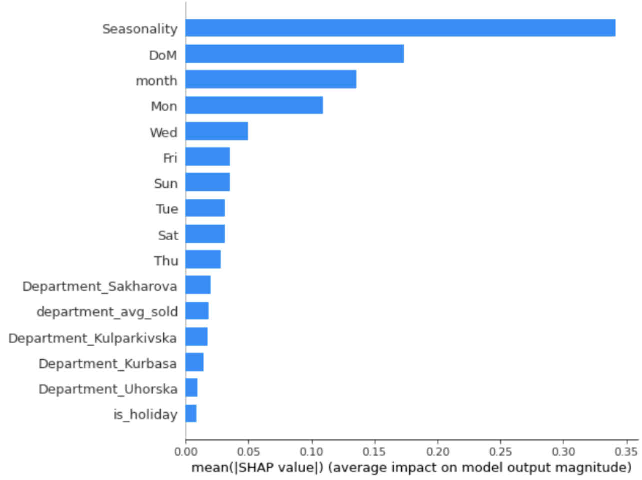


FIGURE 7.2: SHAP Absolute Value for each Feature (RF on Quantity)

However, when we need to decide on a feature, it is better to consider SHAP summary plot. It is on this chart that we can see that the same feature from Monday is very influential, but in a negative direction. This graph allows you to see for the first time in which direction the impact is taking place (see Figure 7.3).

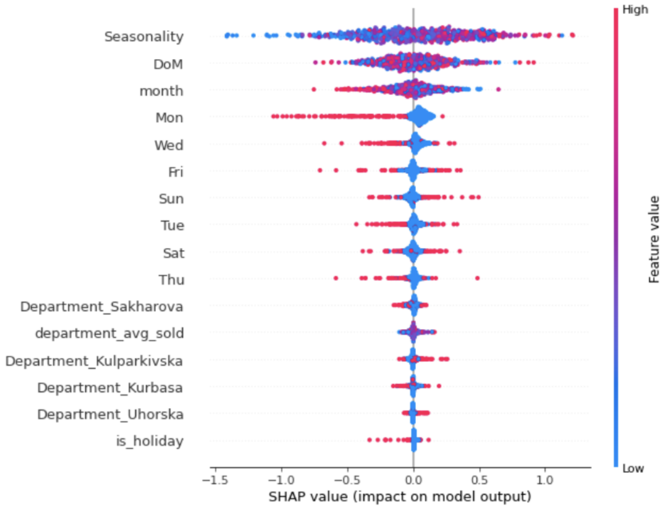


FIGURE 7.3: SHAP Value for each Feature (RF on Quantity)

However, this graph shows only how the feature affects, but each of its individual values can not be explored. In this model, it is clear that the day of the month is the second most important feature and should be considered with another graph - dependence plot (see Figure 7.4), in which you can see each individual value of this feature and the impact on the model. The results can be interpreted as follows: the first days of the month have the greatest impact, which is why they show the best positive result on the model. And by the end of the month, this trend is declining and the last weeks of the month affect the model more in a negative way, although there are exceptions.

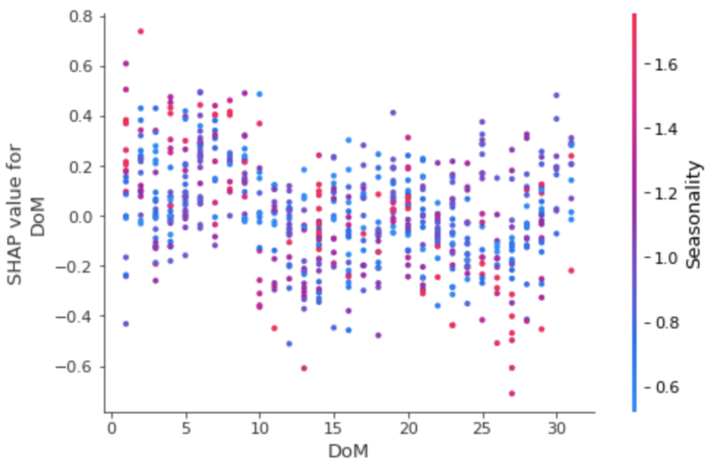


FIGURE 7.4: SHAP Dependence Plot of "DoM" Feature (RF on Quantity)

7.3.2 Random Forest regression on Price

Again, with the help of SHAP we analyze the importance of features and their impact now depending on the regression of the model on the price. The importance in this case has shifted from seasonality and the day of the week more to the month (see Figure 7.5).

Although as you can see on the graph (see Figure 7.6), the month has approximately the same effect on both positive and negative results of the model. Having researched more deeply and built a dependence plot, we can see that at the beginning and end of the year the model has a positive impact, because in these months there are more sales due to the holidays. The opposite trend can be seen in late spring and summer, where there are almost no such big holidays. This is also confirmed by the positive influence of the "is_holiday" feature and seasonality. Compared to Monday, we see the opposite trend compared to the previous model, because here we were interested in the price, not the number of sales, and as we know on Monday the company has the lowest number of sales. At the same time, the other days of the week went to the negative side of the impact, because they are more or less all equal in the number of sales and the price is not so much affected.

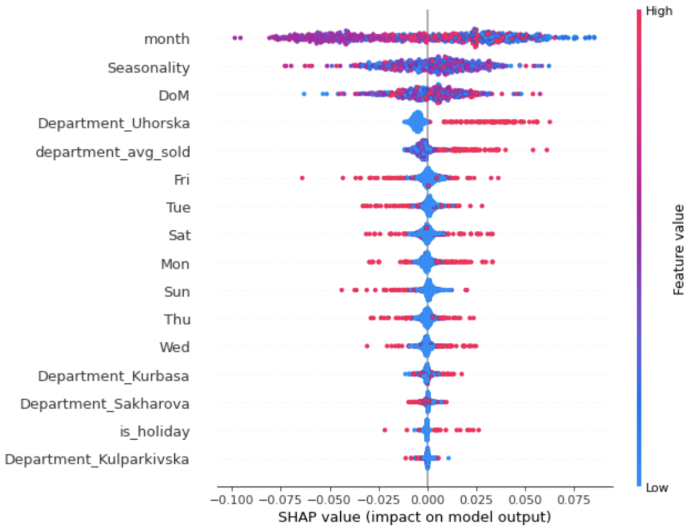


FIGURE 7.5: SHAP Value for each Feature (RF on Price)

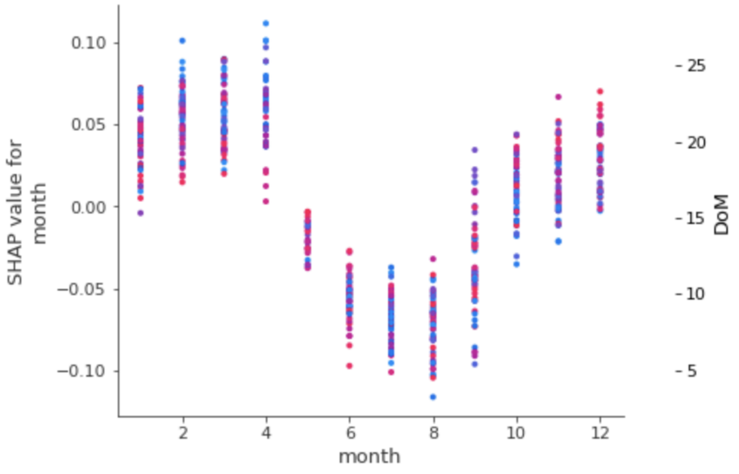
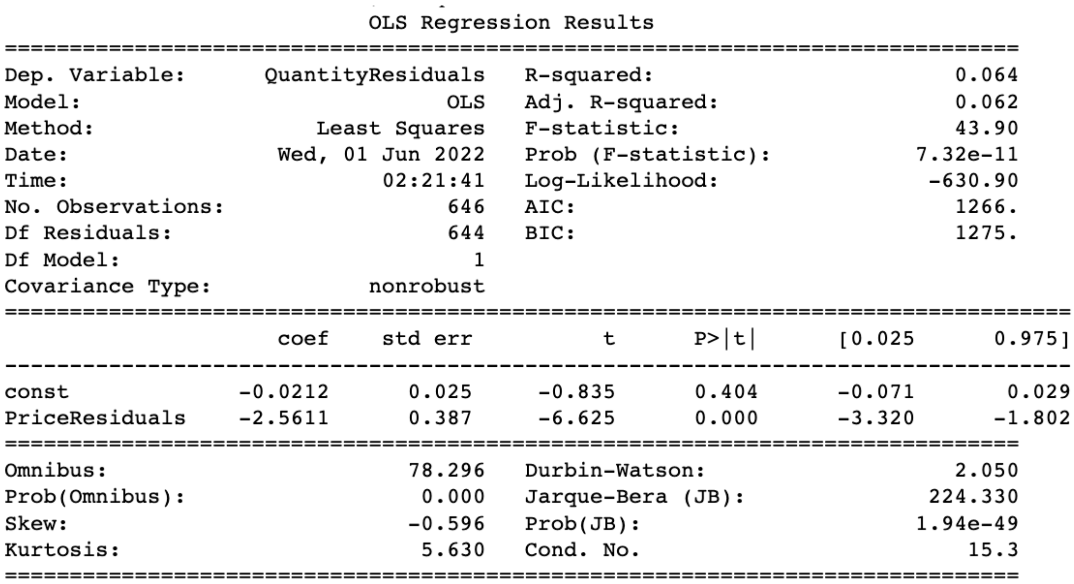


FIGURE 7.6: SHAP Dependence Plot of "Month" Feature (RF on Price)

7.4 OLS on Residuals

The last step in the Doble ML approach is regressing residuals obtained from quantity Random Forest on residuals from Random Forest. For this regression we use the OLS method. The least squares principle provides a way of choosing the coefficients effectively by minimizing the sum of the squared errors. We perform OLS on every goodId separately to get as accurate an estimation as possible. For coconut eclair the results of OLS regression are presented in the Figure 7.7 and the distribution of its residuals in illustrated on QQ-plot in Figure 7.8.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:      QuantityResiduals   R-squared:                       0.064
Model:                            OLS   Adj. R-squared:                  0.062
Method:                 Least Squares   F-statistic:                     43.90
Date:                Wed, 01 Jun 2022   Prob (F-statistic):           7.32e-11
Time:                        02:21:41   Log-Likelihood:                -630.90
No. Observations:                 646   AIC:                             1266.
Df Residuals:                     644   BIC:                             1275.
Df Model:                           1
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const             -0.0212      0.025     -0.835      0.404      -0.071       0.029
PriceResiduals    -2.5611      0.387     -6.625      0.000      -3.320      -1.802
==============================================================================
Omnibus:                       78.296   Durbin-Watson:                   2.050
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              224.330
Skew:                          -0.596   Prob(JB):                     1.94e-49
Kurtosis:                       5.630   Cond. No.                         15.3
==============================================================================
```

FIGURE 7.7: OLS regression results for coconut eclair

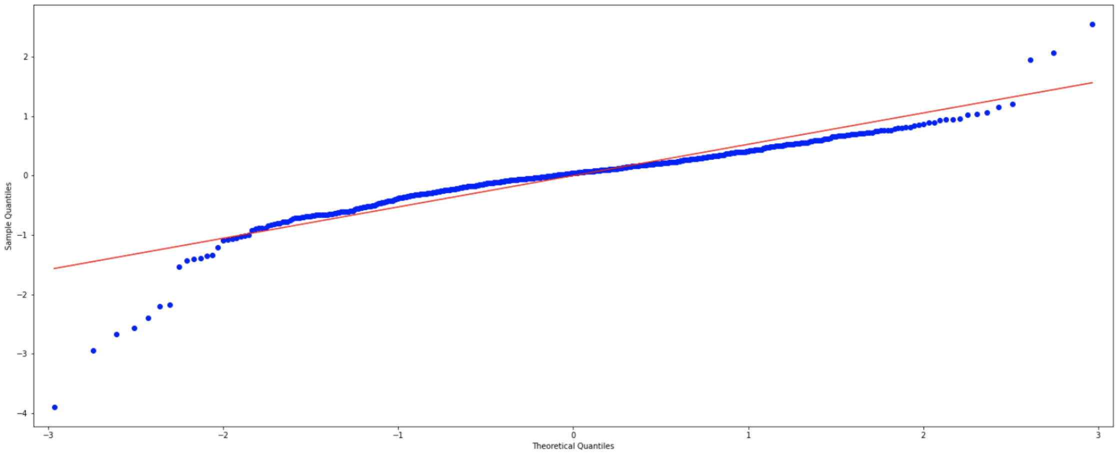


FIGURE 7.8: QQ-plot for OLS regression residuals

7.5 Results of elasticity estimates

In our case, the elasticity estimate is the exact $\hat{\beta}_1$ coefficient obtained by OLS regression. To be more precise, the estimate is equal to the coefficient determined for PriceResiduals parameter. To compare with naive OLS results in section 7.2 (Figure

7.1) we have efficient and more accurate esimate of elasticity. Naive OLS elasticity estimate for coconut eclair equals -1.2209, whether efficient Double ML estimator is -2.627.

Chapter 8

Conclusions and Further Work

8.1 Results

A big advantage of this approach is that we managed to get individual elasticity estimates for all the selected products.

As a consequence of Double ML we construct a convenient table with all values of price elasticity of demand for each product (Figure 8.1).

Knowing the estimate of elasticity, moreover, knowing not the general elasticity of the business, but for each product separately, we get this great privilege to adjust prices in out favor. This number tells us how much we can change the price to, ideally, meet all the existing demand at that price and, furthermore, sell all the products.

As a result of this project we managed to meet the preliminary set challenges. We reached the goal of determining the actual price elasticity of various products offered by a café, as it could have been if the cafe experimented with their prices, because we simulated the price changes correctly adjusting the original demand, thus studying how the demand changes if the price moved by an arbitrary percentage rate.

The result of this work gave a huge opportunity to improve current pricing strategy and enhance business performance.

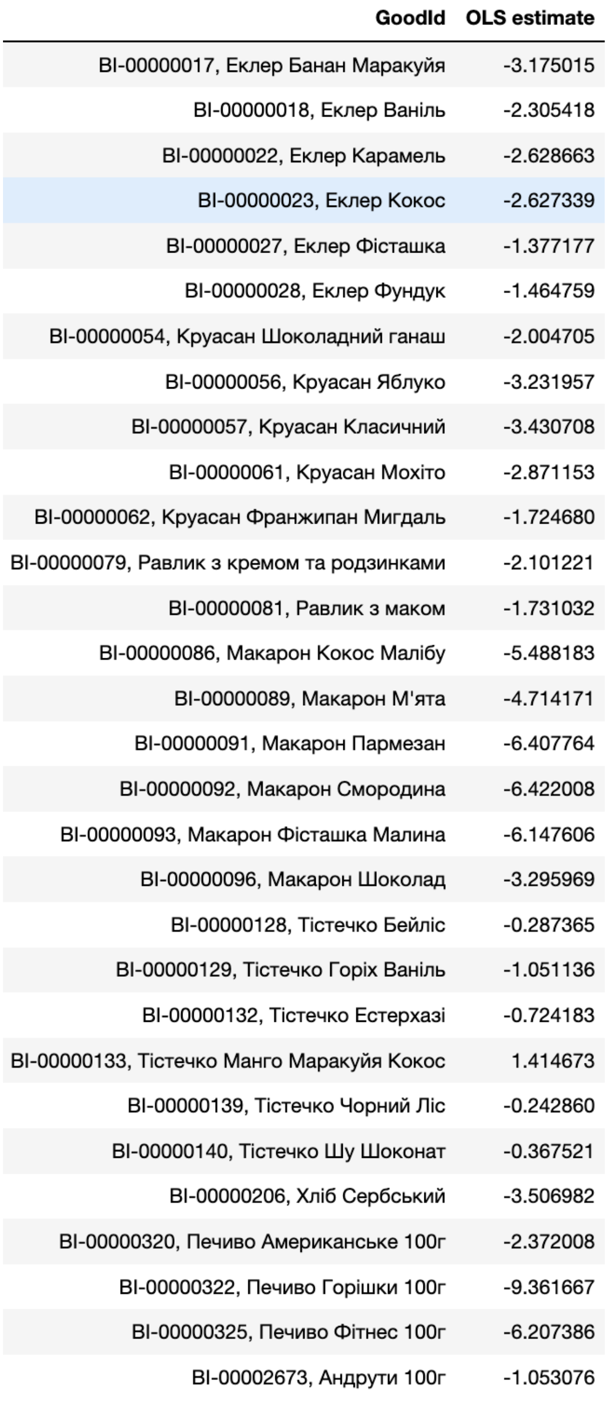
8.2 Further Work

Moreover, the Double ML method we used in this work was enough for us to determine the correct and efficient elasticity for selected products, there is still space to improve the model to make it even more accurate.

Firstly, to continue this discovery it is worth trying to tune the hyper parameters and study the outcomes.

Secondly, to get more interesting interconnections and results, one can work deeper with accounting system to extract more volumes of data on cost, marketing costs, logistics and etc.

The last but not least, is making this idea work in practice and try apply ML modeling as a branch of sales analysis on the regular basis. In addition, the collected empirical results of this model's performance and efficiency in application will open even more horizons for improvements.



GoodId	OLS estimate
BI-00000017, Еклер Банан Маракуйя	-3.175015
BI-00000018, Еклер Ваніль	-2.305418
BI-00000022, Еклер Карамель	-2.628663
BI-00000023, Еклер Кокос	-2.627339
BI-00000027, Еклер Фісташка	-1.377177
BI-00000028, Еклер Фундук	-1.464759
BI-00000054, Круасан Шоколадний ганаш	-2.004705
BI-00000056, Круасан Яблуко	-3.231957
BI-00000057, Круасан Класичний	-3.430708
BI-00000061, Круасан Мохіто	-2.871153
BI-00000062, Круасан Франжипан Мигдаль	-1.724680
BI-00000079, Равлик з кремом та родзинками	-2.101221
BI-00000081, Равлик з маком	-1.731032
BI-00000086, Макарон Кокос Малібу	-5.488183
BI-00000089, Макарон М'ята	-4.714171
BI-00000091, Макарон Пармезан	-6.407764
BI-00000092, Макарон Смородина	-6.422008
BI-00000093, Макарон Фісташка Малина	-6.147606
BI-00000096, Макарон Шоколад	-3.295969
BI-00000128, Тістечко Бейліс	-0.287365
BI-00000129, Тістечко Горіх Ваніль	-1.051136
BI-00000132, Тістечко Естерхазі	-0.724183
BI-00000133, Тістечко Манго Маракуйя Кокос	1.414673
BI-00000139, Тістечко Чорний Ліс	-0.242860
BI-00000140, Тістечко Шу Шоконат	-0.367521
BI-00000206, Хліб Сербський	-3.506982
BI-00000320, Печиво Американське 100г	-2.372008
BI-00000322, Печиво Горішки 100г	-9.361667
BI-00000325, Печиво Фітнес 100г	-6.207386
BI-00002673, Андрути 100г	-1.053076

FIGURE 8.1: Elasticity estimates obtained by OLS regression for every product

Bibliography

[1] Manole, L. (2019). *How can AI and Machine learning impact healthcare industry?*
https://data-science-blog.com/blog/2019/10/08/how-can-ai-and-machine-learning-impact-healthcare-industry/

[2] *The Value of Artificial Intelligence for Retail in 2022*
https://spd.group/artificial-intelligence/ai-for-retail/

[3] Campbell P. (2022). *Importance of Pricing: Why pricing is important for SaaS and beyond* The landmark study was published in a 1992 Harvard Business Review by Michael Marn and Robert Rosiello, both senior pricing folks at McKinsey and Company
https://www.priceintelligently.com/blog/bid/157964/two-reasons-why-pricing-is-the-most-important-aspect-of-your-business

[4] Chernozhukov V., Chetverikov D., Demirer M., Duflo E., Hansen C., and Newey W. (2016) *Double Machine Learning for Treatment and Causal Parameters* Original work about Double ML method presented by aforementioned authors at Massachusetts Institute of Technology
https://arxiv.org/pdf/1608.00060.pdf

[5] Published by OTexts: *Time series decompositions*
https://otexts.com/fpp2/decomposition.html

[6] Lebanon G. (2006) *Consistency of Estimators*
http://theanalysisofdata.com/notes/consistency.pdf

[7] Roemheld L. (2021) *Causal Inference in the Wild: Elasticity Pricing*
https://towardsdatascience.com/causal-inference-example-elasticity-de4a3e2e621b

[8] Columbus L. (2018) *10 Ways Machine Learning Is Revolutionizing Marketing*
https://www.forbes.com/sites/louiscolumbus/2018/02/25/10-ways-machine-learning-is-revolutionizing-marketing/#6fd13db95bb6

[9] Velasco B. (2021) *Double Machine Learning for causal inference*
https://towardsdatascience.com/double-machine-learning-for-causal-inference-78e0c6111f

[10] *Double Machine Learning documentation*
https://docs.doubleml.org/stable/guide/basics.html

[11] Lyashenko V. *How to use random forest for regression: notebook, examples and documentation*
https://cnvrg.io/random-forest-regression/