INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

Forecasting Sales for Aesthetic Products Using Machine Learning

Rushayeal Galea Massa

Supervisor: Alan Gatt

June – 2023

A dissertation submitted to the Institute of Information and Communication Technology in partial fulfilment of the requirements for the degree of Bachelor of Science (Hons) in Business Analytics

Authorship Statement

This dissertation is based on the results of research carried out by myself, is my own composition, and has not been previously presented for any other certified or uncertified qualification.

The research was carried out under the supervision of Alan Gatt – Forecasting sales for aesthetic products using Machine Learning, Rushayeal Galea Massa

RG. Galea

5th June 2023

Copyright Statement

In submitting this dissertation to the MCAST Institute of Information and Communication Technology, I understand that I am giving permission for it to be made available for use in accordance with the regulations of MCAST and the Library and Learning Resource Centre. I accept that my dissertation may be made publicly available at MCAST's discretion.

5th June 2023

Acknowledgements

First and foremost, I'd like to offer my deepest appreciation to my supervisor Mr. Alan Gatt for providing guidance and feedback throughout this project. I would also like to express my heartfelt gratitude to my mother, the pillar of our family, for her unwavering support throughout my academic path, which would not have been possible without her. I am also grateful for my boyfriend and my siblings, who have always believed in me and encouraged me to keep pushing forward. Finally, I would like to thank Erika, my best friend, who provided me with invaluable advice and emotional support that will serve me well throughout my life.

Abstract

In this study, aesthetics sales data obtained from a company based in the United Kingdom is used to train and test various machine learning algorithms, to forecast demand for particular products. In this data, the Covid-19 sales period is included, and questions related to the effect caused by the pandemic are answered in the research conducted. This research was carried out to have a detailed analysis of the sales made between the year 2015 and the year 2022 and highlight any patterns and trends which may occur in the data. The problem faced during this research was related to the categorical values which are present in the original dataset so to eliminate this issue, one-hot encoding was used to transform the data into numerical form. As suggested by earlier studies, Random Forest and XGBoost were used to perform the forecasting. While an additional algorithm, Neural Network, was also tested on the data. The algorithm that performed the best of all the experiments tried out was Neural Network, with an RMSE results as low as 0.98, which was obtained in the experiment for the last 3 months, while the category was Mesotherapy. This was followed by XGBoost, which also achieved a low RMSE results of 1.98, which was obtained in the experiment for the summer period, while the category was needles. The data was tested out on different time periods to analyse the patterns and trends further. The time periods used were summer, winter, the whole 2019 year, Covid-19 period and the last 3 months of the data. During each period, each algorithm was tested out for each category one by one. This was done since the sales of the categories vary from one another and the result cannot be predicted correctly if all the sales of the categories are combined. From this study it was found that the Covid-19 pandemic did have a negative effect on the sales of this company as they decreased substantially from the sales of the previous years. It was also found that machine learning algorithms are an efficient solution to be used when predicting supply and demand for aesthetic products. According to the findings of this study, the Covid-19 pandemic had almost the same impact on all regions of the United Kingdom. between the years 2020 and 2021. Based on an analysis of current data collected during the year 2022, it was

discovered that sales began to rebound to the high levels seen before the pandemic affected the United Kingdom and Ireland.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Research Problem

This research is based on the sales data of a UK based aesthetic company. The gathered data includes the sales transactions which occurred from 2015 to 2022. The transactions are available for 26 different products which are sold by the company. The quantity feature will be used later to look for any patterns or trends which might be present in the dataset. In addition to trends and patterns, seasonality and any other possible events will also be investigated. For this type of study, there are various machine learning algorithms which can be evaluated, since algorithms can perform differently on the same dataset. These machine learning algorithms will be used to predict sales data so that if the result is accurately forecasted, these algorithms can be used to predict future data. The Covid-19 sales periods will be analysed in this research to be able to determine if the pandemic had any effect on the sales of this company. This will be conducted so that the company can reduce their costs and meet customer demands in a most efficient manner. Other time period will also be analysed such as summer, winter, one whole year and the last three months of the data.

1.2 Research Question

In this study, the research questions are:

- Did the Covid-19 lockdown affect the sales of aesthetics products?

- Did the Covid-19 pandemic have the same impact on all the regions in the UK?

- Can machine learning algorithms be used to make a prediction for supply based on previous transactions?

The main hypothesis stated for this study is that Covid-19 did in fact affect the sales of aesthetic products. This hypothesis was formulated due to the negative effects other non-essential businesses experiences during the Covid-19 periods. The hypothesis for the second research question is that the effect was approximately the same for all the regions in the UK since the lockdown was enforced on all the UK regions. The hypothesis formulated for the third research question is that machine learning can be used to make predictions for a supply based on previous transactions, given that enough useful information in the correct format is provided.

1.3 Aims and Objectives

Sales forecasting is used in many businesses nowadays as this is very helpful when it comes to making orders and planning production for certain products. This is especially helpful when it comes to planning supply and demand for products which have a short expiration date. The aim of this research is to find out how the Covid-19 pandemic affected the company's sales and in what way were they affected. Another aim of this research is to find out the Covid-19's effect on the regions in the United Kingdom. The company will also have a forecasting solution which is already tried and tested against their data to help them finalise any future orders. They will also be provided with a detailed analysis of the product quantities which were sold during different months for different categories. The risk of over-ordering too many products or not meeting the customer demands will be reduced as they can easily generate an estimate of the supply and demand. The company will know more regarding which seasons are more successful and which are the most sold products from their ranges, together with the most popular categories. This study will also be very helpful to the company when it comes to planning and providing training seminar related to their products since they can choose on what to focus them more about. Since the findings will also be compared between different regions across the UK, this will determine how regions were impacted.

1.4 Purpose Statement

The reason why this research was chosen was that data analysis is a very interesting subject to explore in a more realistic scenario. The data was gathered through a personal contact who is working in a company that distributes aesthetic products mainly across the United Kingdom and a few other countries such as USA and Brazil. Data analysis techniques will be used to identify the patterns which are present in this data. Machine learning forecasting techniques will also be used to help them formulate a better marketing and supply strategy depending on the current and future market demand and find out which is the optimal algorithm to be used on this dataset. The reason for choosing to conduct this research using this approach is that it saves more time when trying out experiment using computerised methods rather than statistical model. This is especially suitable since the dataset is also larger in size.

Chapter 2: Literature Review

Forecasting is the process of making predictions based on past and present data. Forecasting is used by industries and business entities, including aesthetic companies to research the supply and demand for their products. This assists industries in controlling the market chains.

2.1 Sales Forecasting:

Forecasting entails making predictions about the expected future demand and supplies. This is useful so that the customer demands, and the supply requirements are met.

2.1.1 Meeting customer demand

In businesses, past sales information can be used to forecast the future sales for the immediate months. Forecasting in companies is used, so that the business can have a general grasp of the market demand and help to set goals and plan resources and logistics for this demand. Due to the increase in the use of AI technology, competition in the market is growing at a more rapid pace [1]. The company can also have an indication of what type of services or products are more popular with clients and in what ways are the trends and patterns are changing and evolving. If the company does not satisfy the market demand for a product, customers will have to opt to purchase from competitors, which will then result in loss of sales and possibly even losing the customers' loyalty. A business can get all the demand information by applying their sales data to any of the desired forecasting models and increasing the chances of improving the satisfaction levels of its customers.

2.1.2 Cosmeceutical products supply and demand

Cosmeceuticals are the products which fit the niche between drugs and cosmetics. This term is used in the professional skin care industry to describe a product that has measurable biological effect in the skin but is regulated as a cosmetic since it claims to affect appearance. These professional skin care products come in the form of sunscreen, antiaging creams, foundation, facemasks, derma fillers and more. Cosmeceuticals are the fastest growing segment of the personal care industry and their formulations have expanded from skin to body to hair and several cosmeceutical treatments for conditions. Studies focusing on the cosmeceutical products highlight that there will be strong growth perspective for this industry in the coming years [2].

2.1.2.1 Seasonality

Seasonality stands for regular patterns that are determined by different seasons over a given number of periods. Seasonality in terms of sales refers to the oscillations in total amount of sales that occur throughout one year and then repeats in the following years. Seasonality is not determined by the volume of sales of the whole year but in volumes during specific periods. This periodic seasonality can be based on short volumes such as weeks or months [3]. In an industry, seasonality in sales is highly influenced by the different seasons of the year, holiday periods such as summer breaks or Christmas holidays, and other notable dates such as Mother’s Day, Father’s Day, or Valentine’s Day. When a forecast is predicted based on seasonality, the accuracy must be very high because if the forecast is inaccurate, this may cause major issues in marketing, production, investment, and expenses. If the marketing and advertisement is not scheduled in the correct seasonality, the targets will not be achieved and the money for campaigns would not have the desired effect. Over production of products cause by incorrect seasonality can also be an issue as certain products have to be sold in a specific period of time. If these products are not sold in this period of time, they may have to be either thrown away or sold at a very cheap price that does not cover expenses.

2.1.2.2 Shelf life

Once a product is produced, this product is labelled with an estimated shelf life. This shelf-life duration depends on the product itself which can be either a small number of days or it may still be good for several years. Products which are natural with less chemicals or additives tend to have a shorter shelf life than products with preservatives and additives. Shelf life is a very important feature in a product as it determines if the product is still safe to consume or use and guarantees the quality of the product. When a customer purchases a product which usually has a shelf life of more than one year, they would not expect the product to expire in a couple of days or months.

2.1.3 Extraordinary events affecting supply and demand

When an extraordinary event happens around the world, this may influence the sales of businesses which can lead to either a decrease or an increase of sales of certain products.

2.1.3.1 Covid-19

The global pandemic in 2020 lead to a worldwide lockdown which affected the worldwide economy and stability [4]. To slow the spread of COVID-19, governments enforced social distancing restrictions and lockdowns on businesses deemed nonessential. The essential businesses were also enforced by restrictions, yet they were less drastic. According to research done by Fairlie and Fossen [5], from February 2020 to April 2020, the number of active business owners dropped by 22%. Year-over-year sales usually increase by 3% to 4%. However, in the second quarter of 2020, sales decreased by 17% whereas online sales increased by 180%. It was found that sales losses were largest in businesses affected by mandatory lockdowns such as accommodations, drinking places, and arts, entertainment, and recreation [5].

In the United Kingdom, a 19.8% decline in Gross Domestic Product (GDP) was caused by public health measures such as social distancing, travel restrictions, and the closure of non-essential businesses between April and June 2020.The biggest quarterly recession in household expenditure ever occurred over this period, when spending on dining out, lodging, transport, and recreation decreased by over 20%.

The 11.6 million jobs affected by the furlough plan considerably lessened the labour market's effects, causing the unemployment rate to increase from 3.8% at the end of 2019 to 5.2% at the end of 2020.

Following the removal of restrictions, the GDP increased by 17.6% in the third quarter of 2020, from July to September. In the third quarter of 2020, household expenditure increased by 19.6%, including increases in dining, lodging, and transportation costs. Average home prices increased by 13.5% in the year leading up to June 2021.

Despite the Delta variant's emergence and the following lockdown causing a 1.2% GDP decline over the first three months of 2021, the remainder of the year saw incremental growth. In spring and summer of 2021, household spending increased once more, returning steadily to pre-coronavirus pandemic levels by 8.5% and 2.6%, respectively.

By the first quarter of 2022, GDP had restored to pre-coronavirus pandemic levels[1] [6].

2.1.3.2 Brexit

As the United Kingdom officially exited the single market and customs union at the end of the transition period on December 31, 2020, which was the result of the June 2016 referendum, trade with the EU initially dropped.

According to research found, the Brexit process is estimated to have reduced the level of UK productivity by between 2% to 5% over the three years since the referendum [7].

It is challenging to separate the economic effects of Brexit from the COVID-19 pandemic, the disruption of the global supply chain, and the increases in energy and food prices since they all overlapped [6].

---

1 https://www.ons.gov.uk/economy/grossdomesticproductgdp/articles/gdpandeventsinhistoryhowthecovid19pandemicshockedtheukeconomy/2022-05-24#:~:text=The%20COVID%2D19%20pandemic%20prompted,country%20reopened%20over%20the%20summer.

2.1.3.3 Change in Government

Since 2015, the UK has had five different prime ministers in the government. David Cameron was the prime minister between 2015 and 2016. Between 2016 and 2019, Theresa May was elected prime minister. Boris Johnson acted as a prime minister between 2019 and 2022 followed by Liz Truss in the same year. Rishi Sunak is the current prime minister of the UK since 2022.

2.2 Machine learning:

Machine learning is a field devoted to understanding and developing ways that allow machines to learn approaches that use data to enhance computer performance on a set of tasks.

2.2.1 Introduction to machine learning

Machine learning is a set of algorithms found in the Artificial Intelligence (AI) discipline, which allows a computer to predict outcomes without being specifically programmed to do so [8]. This type of AI is becoming more popular and widely used in fields such as banking and finance, real estate, healthcare, retail, education, insurance, and pharmaceuticals [9]. The main objective of machine learning is to identify patterns based on predictors and then be able to use these patterns to predict an outcome on unseen data. These models can be trained more than once, given related-context data to be able to predict future data and make necessary decisions [10].

If an algorithm is well trained, it will be able to learn very complex scenarios and predict the expected data with high precision. The choice of the correct predictors is very important when training a machine learning model [11].

Besides forecasting data, due to its versatile nature, machine learning can also be used for facial recognition, car identification, detecting credit fraud, detecting spam, providing personalised recommendations and services, and virtual customer support amongst others [12]. Machine learning is made up of different algorithms, each using different mathematical models to fit data and be able to conduct predictions. Some commonly used algorithms are Linear Regression, Decision Trees, Random Forest, KNN and K-means.

2.2.2 Machine learning approaches

There are different approaches for machine learning; most common are supervised learning, unsupervised learning, and reinforcement learning. Each of these techniques works by implementing different algorithms which process and learn data in different ways.

The main difference between supervised and unsupervised techniques is that supervised training requires the programmer to label data beforehand in order for the machine learning algorithm to predict an outcome (focuses more on classifying the data using labels), while unsupervised focuses more on clustering the data, reducing dimensions and identifying sequences by association [13].

Supervised Learning is commonly used in sentiment analysis, predictive analysis based on regression or categorical classification, natural language processing, detecting email spam and image classification. Some algorithms used in this approach are Decision Tree, Random Forest, Support Vector Machines and Linear Regression. Unsupervised learning is mostly used in scenarios of speech processing, object categorisation, audio classification, and automatic labelling. Commonly used algorithms for this approach are K-Means, Hierarchical Clustering, Density-Based Clustering and A-priori. Algorithms which can be implemented in both supervised and unsupervised learning are Naïve Bayes, Auto Regressive Integrated Moving Average (ARIMA) and Deep Learning.

Reinforced learning includes techniques like deep Q-network, post-decision state, Dyna-Q and Q-learning. These methods assist IoT devices in selecting security protocols and key parameters for various threats through trial and error. As an example, Q-learning is used as a model-free technique to enhance malware detection, offloaded anti-jamming, and authentication performance.

2.3 Forecasting supply and demand:

Nowadays, forecasting using computerised methods has become a very popular process that is used in different fields to be able to make predictions of future data based on past data. As stated by Petropoulos *at al.* [14], in around 15 years, the field of forecasting has seen amazing growth in both theory and practice. Using today's advanced technology, one can immediately get a detailed hour-by-hour weather forecast. In this forecast, the temperature, rain, weather conditions, UV index, wind and much more are predicted. Probability forecasts are used when uncertainty is irreducible, for example it can be used during elections to forecast which party will win the election. This does not necessarily mean that the party forecasted to win will surely win, but there is a great possibility that the forecast is correct. Forecasts can also be utilised in companies when they are selling a service or products to predict future sales based on the past sales made [14].

All these forecasts can be estimated based on using multiple past parameters called the predictors. This prediction is made available using the different algorithms found in machine learning to be able to not only forecast upcoming data, but also help in decision-making and analysing current status such as the company's performance. The information based on the forecasting can help businesses to allocate resources, anticipate expenses and plan their budgets wisely. When it comes to predicting sales, one can also have an idea of how the production schedules need to be set [15]. When performing the forecasting, one must pay attention to the data being used and how it is used in the forecasting, as if the data is not used correctly, the forecast will not output the correct results either. If a company makes their decisions based on the incorrect predictions, the company may suffer many losses in both target audiences and the sales of products or services.

2.3.1 Time Series Modelling

At present, the most commonly used forecasting method of sales forecasting is time series modelling. From the sales data, multiple patterns and trends can be analysed such as trend in sales, seasonality, autocorrelation, patterns caused by the impact of external factors, pricing, and competitors' behaviours [16]. Sales prediction can be a very complex problem, especially if the data includes outliers, and missing data. At present, several time series models have been developed to be able to overcome this issue by using models such as ARIMA, SARIMAX, SARIMA, Random Forest, and SVM. To be able to implement such models, the sales data needs to include historical data for a long period of time to capture the seasonality and patterns, while also outliers must also be removed from the dataset before using one of the time series models.

In the Time Series Modelling, different patterns may be present in the dataset. The data can include trends which occur when the data is either increasing or decreasing persistently for a long period of time which does not have to be linear [17]. Another type of time series pattern is the cyclic pattern. This pattern includes repeated fluctuations which are non-periodic, while the duration of these fluctuations is usually of a minimum of two years. These fluctuations are usually due to the economic conditions. A third pattern is the seasonal pattern which reflects the seasonality that is present in the Time Series data. Seasonality is always of a fixed and known frequency [18]. An example of seasonality is like demand for sunblock which will be highest during the summer season.

Since most machine learning models can understand only numerical values, when data contains categorical values, a different approach must be taken. This issue can be resolved by using multiple encoding techniques such as label encoding or one-hot encoding. In the label encoding technique, each categorical variable is assigned an integer [19]. When a categorical feature has more than two values, the label encoding technique may cause some undesired issues as the integers assigned may be misinterpreted by the algorithms as having some sort of hierarchical order. This issue may be resolved by using the one-hot encoding technique. In this technique, several additional features are created based on the number of unique values in the categorical feature. This mapping can have Q number of possible values, into a vector with Q number of elements, where only the element corresponding to the current feature value is “1”, while the remaining elements are “0’s” [20]. This encoding is applied to all the unique variables in the categorical variables.

2.3.2 Forecasting Models

When predicting data, several models or algorithms can be used. The most used models for time series are Naïve model, Exponential smoothing model, ARIMA or SARIMA, Linear regression method, Multi-Layer Perceptron (MLP), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Decision Trees, XGBoost and AdaBoost. Each model has its own unique algorithm, and the accuracy may vary depending on the data being predicted.

In the study ‘Demand Forecasting Using Coupling Of Machine Learning And Time Series Models For The Automotive After Market Sector’, demand for automotive products is forecasted through different models using historical demand history data along with external factors that influence demand. Machine learning models and Time-Series models are split into two sections and subcategorised to compare the best model [21]. According to the researcher, it was found that the top 3 models of machine learning are ADA-boost, XGB boost and GBM. On the other hand, the top 3 models for time series models are Auto-ARIMA, Naïve, and Naïve-rept. The best model is then selected by calculating the Root Mean Squared Error (RMSE) of each model and choosing the approach with the smallest RMSE.

2.3.3 Related Studies

In a study conducted by Akanksha *at al.* [22],store sales were predicted using Linear Regression, Random Forest and XGBoost models. The Linear Regression model's performance was so poor, that it was not included in the paper. The dataset used contained approximately 2.94 million observations and had 6 variables. In this scenario, the XGBoost gained the best results, obtaining an RMSE of 3.86, while Random Forest obtained an RMSE of 5.13.

As proposed by Huang *at al.* [23], to forecast cosmetics sales, the most appropriate method is LS-SVM. In this paper, the three models which were used are BPNN, LS-SVM and AR model. In terms of Mean absolute percentage error (MAPE), LS-SVM test resulted in 9.21%, BPNN test resulted in 12.56% while the AR model resulted in 17.13%. When evaluating the models using the Pearson Correlation Coefficient, LS-SVM had the highest correlation of 0.91, BPNN had a correlation of 0.86 while the AR model had the least correlation of 0.81.

In a study conducted by Sajawal *at al.* [24], it was stated that sales forecasting is the most challenging task for the inventory management, marketing, customer service and Business financial planning for the retail industry. Therefore, they used different machine learning techniques to perform predictive analysis of retail sales data. The regression models used were Linear regression, Random Forest Regression and Gradient Boosting Regression, while the time series models used were ARIMA and LSTM. The results show that Gradient Boosting Regression performed best with an RMSE of 0.63 followed by Random Forest with an RMSE of 0.69. The model that performed the worst in this study was the ARIMA model.

Vithitsoontorn and Chongstitvatana [25] researched both statistical and deep learning methods when conducting demand forecasting. They concluded that both are reliable and are suitable to be used when predicting demand. The study showed that ARIMAs predictions of the future were made in an average straight line, whereas LSTM predicted the future value based on the seasonality and trend of the data. When the model was trained on the monthly observations of the data, the model provided better error scores. The ARIMA model obtained an RMSE of 21501.66 while the Multivariate LSTM obtained an RMSE of 20693.86.

In a study done by Wang *at al.* [26], they attempted to find a forecasting method to balance their purchasing and sales in retail companies. This study explores ARIMA, SVM, RNN and LSTM in five dimensions which are predictive performance, generalization ability, runtime, cost, and convenience. When it comes to accuracy, SVM and LSTM were the best two models to use both when working with normalised or non-normalized data. It was also concluded that LSTM is the most convenient model to use.

Chapter 3: Research Methodology

In the research methodology, the process and the steps taken to achieve the goals is explained in further detail. The first step in the process was gathering the data and loading it into Python. Next, data cleaning was performed on the data and a new file was created with the cleaned data. PowerBI was used to explore the data using some visuals and to make initial predictions. The data was then prepared for Machine Learning algorithms such as data training and testing, and the first experiments were tried out. Some evaluations were outputted by the experiments and the results were documented.

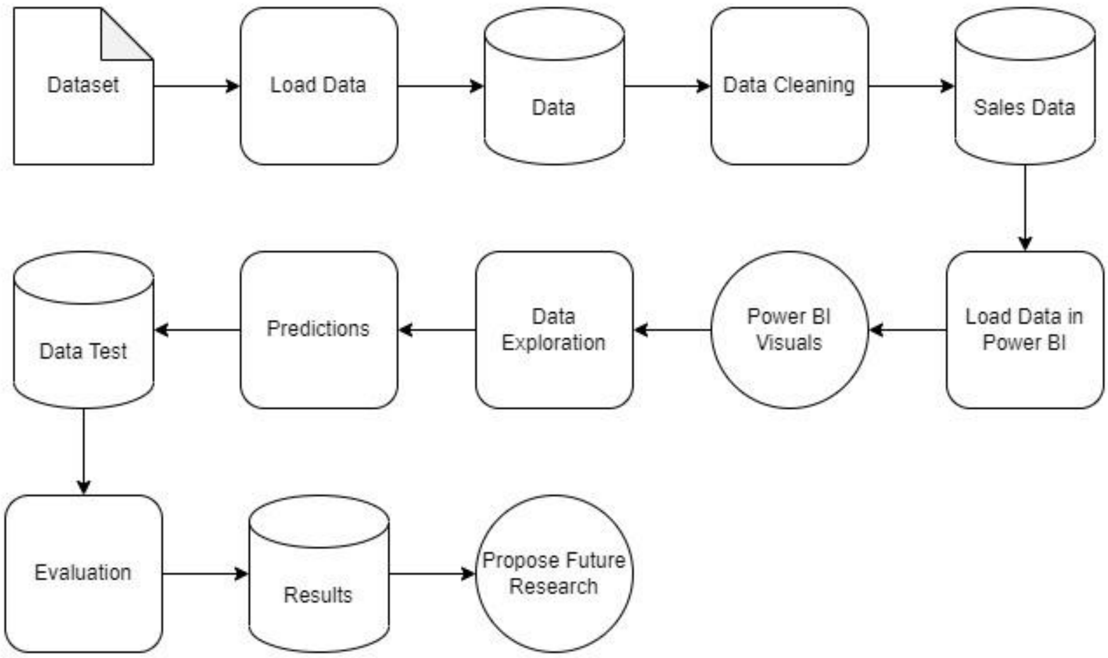

Figure 1: *Pipeline*

3.1 Data Acquisition

The data used in this study was acquired from an aesthetic company based in London which distributes products mainly to the United Kingdom and Ireland.

3.2 Data description

The data acquired is made up of 86,765 unique records which each refer to a unique sale. Each sale is made up of 13 different variables, these are shown in Table 1: Variable Description together with a short description of each variable. The data includes all the sales made between November 2015 and March 2022.

Variable	Description
Data Entry Date	When the data was inputted in the system.
Pharmacy	11 unique pharmacy names.
Year	The year in which the sale was made.
Month	The month and the year in which the sale was made.
Quarter	The quarter and the year in which the sale was made.
Product	26 unique product names, each fall under one of the six categories.
Qty	Quantity sold for each product.
Cust.Name	4,317 unique customer names which identifies who purchased the products.
Comp.Name	4,299 unique company names which identifies which company purchased the products.
Area Code	Full Area code that reflects the regional code in which the sale was made. This is made up of letters and numbers.
Sales Rep	Identifies the sales representative who handled the sale.
Code	The code of the area but inly the letters are listed. Blank is used for countries outside of the UK and Ireland.
Location	10 unique locations in which the sales were made identified by a number. These represented regions

	mainly in the UK and Ireland, and other countries such as the USA.

Table 1: *Variable Description*

3.3 Data Cleaning

To be able to make a time-series forecasting of sales, the data needs to be in a numerical form. To do this, some data cleaning is required so that the data will be tailored according to what is needed and clearly understood by the algorithm.

The first step in the data cleaning process was loading the data from the csv file into a data frame and making sure that the data was being loaded correctly. The data was checked for any null values and the totals for each column were displayed. Null values were found to be present in the ‘Cust.Name’, ‘Comp.Name’ and ‘Code’ columns. The null values were ignored and kept in the data as the mentioned columns were not going to be used when doing the forecasting. The columns which were not considered that important when predicting future sales were the ‘Data Entry Date’, ‘Cust.Name’, ‘Comp.Name’, ’Sales Rep’, ’Area Code’ and ‘Code. These columns were removed so that the forecasting process would be made only on the relevant data. The reason why these were irrelevant where because the sales were not affected by names or when the sale was inputted in the system. Regarding the ‘Area Code’ and ‘Code’ column, the location is enough to apply a forecasting algorithm on this data as only the name of the location will be taken into consideration when doing the analysis.

The next step in the data cleaning process was to make changes to the ‘Month’ column. This column was converted from text values to numeric values to make the data in a numerical form as much as possible. The same process was repeated on the ‘Quarter’ column since the data previously included the letter ‘Q’ in addition to the quarter number. As a result, these two columns were completely changed to numerical forms and could now be used in the forecasting process.

The ’Location’ column contained only a number which represented a specific location. This needed to be changed so that the data could be more easily interpreted by seeing the real location listed in the records. This was done by automatically replacing each number with the specific location value for each record found in the data frame.

The ’Product’ column contained some data inputting errors. Some products were typed down differently in certain records such as written all in capital letters instead of capitalising the first letter of each word. This resulted in having to first extract all the unique values found in the column, and then determining which products were doubled. To make the data more practical, each original product name was changed to an identifiable name such as the type of product and a number added at the end. For the products which were duplicated, instead of giving one value to replace, the variations of the products were listed.

A category column was missing from the data. The different products were first analysed, and then each product was assigned to one category. Six different categories were created, and these were added to the original data frame. The new column named ‘Category’ was used to store these values. The next step was assigning each product to a category, and this was done by using dictionaries found in Python.

Since the data contained thousands of unique customer names, a dictionary could not be used straightaway to generate the practical names. Therefore, the unique names had to be stored in a variable and then an array was created. A function was also created so that each name stored in the array would be changed to the word ‘company’ and a unique number added at the end. A dictionary was then used to replace the customer’s names with the names found in the dictionary created. A copy of the dictionary containing old and new names was then saved in a csv file so that there would be a reference to the real names. The same process was repeated for the company names and the pharmacies.

The final step in the data cleaning process included removing records containing negative quantities which represented the samples which were given to customers. The new data frame was then saved into a new csv file to be used later in the forecasting methods.

3.4 Data Exploration

The data was explored by plotting it against various plot such as Histograms, Boxplots, Pie Charts, and others. The purpose of using a boxplot is to determine if the data contains outliers and if so, in what columns are the outliers present. The purpose of using a histogram is to be able to have a visual representation of how the data is distributed across the different values found in a variable while pie chart obtains the same result but displays them in a pie chart together with percentages.

The distribution of variables can be easily identified when the data is plotted using Histograms. From the histogram shown in Figure 2: Category

Distribution, it can be noted that ‘Profilho’ was the most popular category followed by the ‘Dermafiller’ category. In Figure 3: Product Distribution, it is evident that ‘Injection 7’ was the most sold product across all times with the highest quantities.

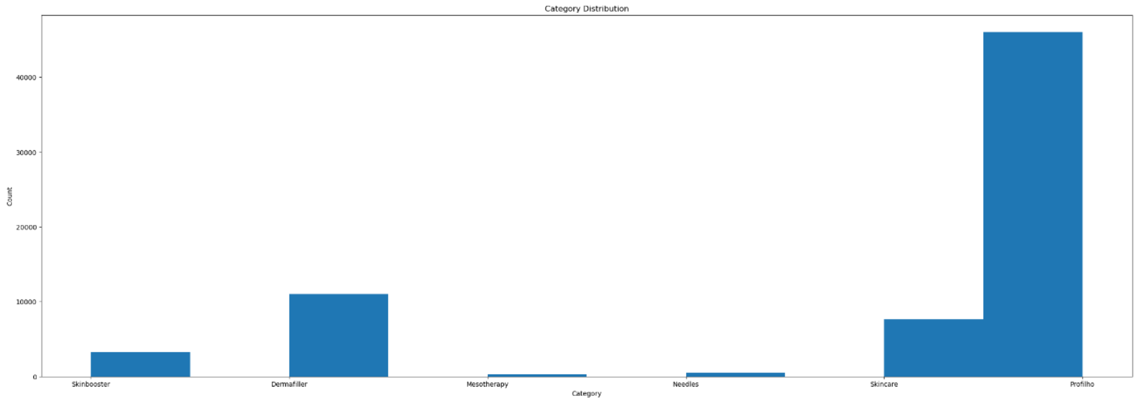

Figure 2: *Category Distribution*

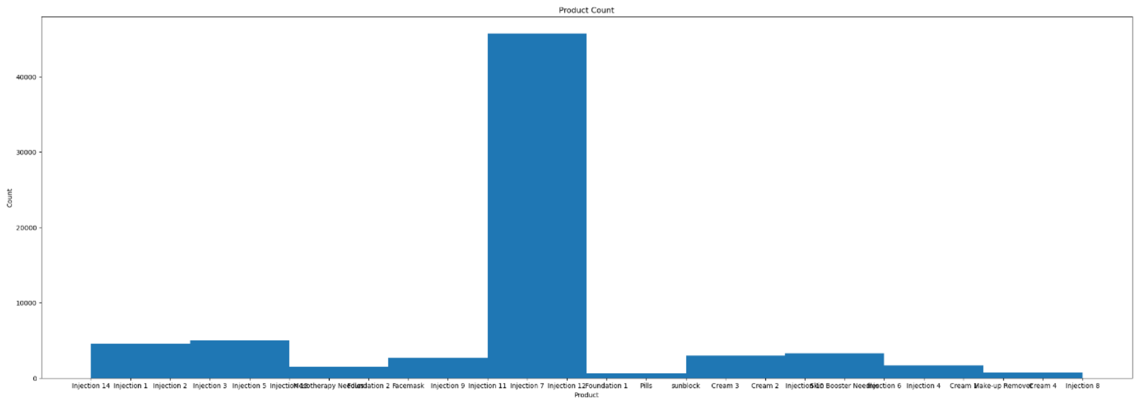

Figure 3: *Product Distribution*

Using boxplots, it was concluded that from all the categories, ‘Profilho’ contained the highest number of anomalies, while from the products ‘Injection 7’ contained the highest number as well. These two variables were filtered out and a boxplot was plotted for each scenario. These are show in in Figure 4: Category Boxplot and Figure 5: Product Boxplot respectively.

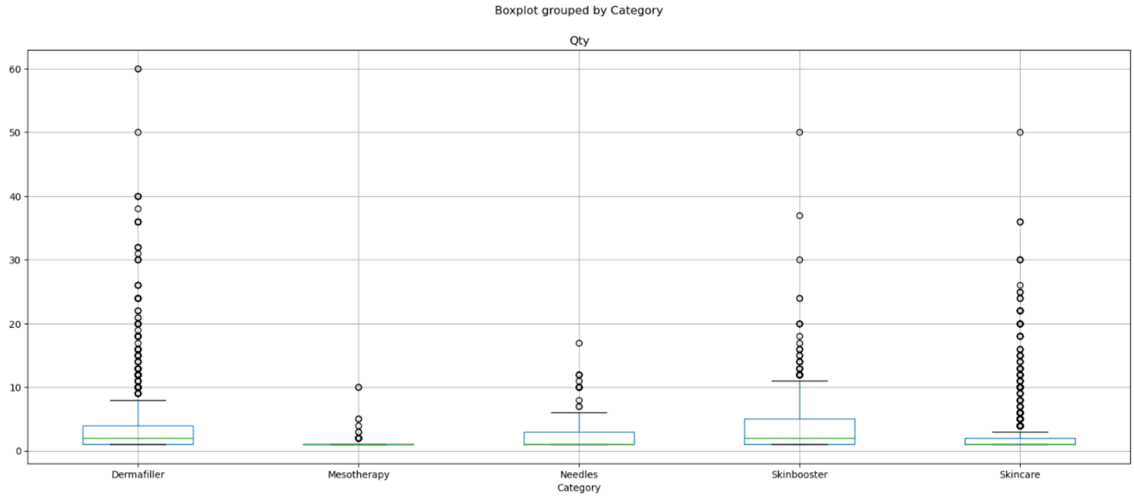

Figure 4: *Category Boxplot*

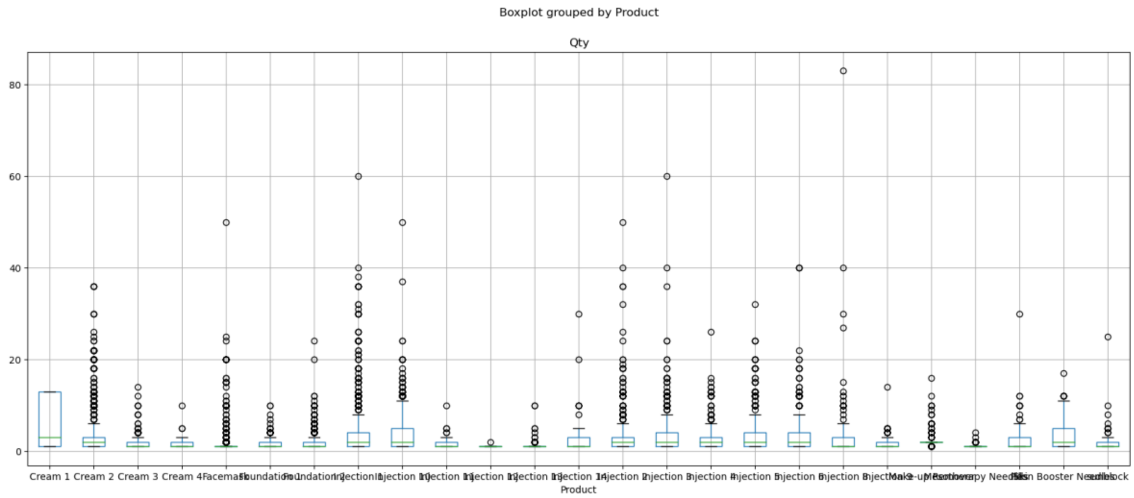

Figure 5: *Product Boxplot*

‘Injection 7’ made up 85% of the products sold as shown in the pie chart in Table 2: Product Percentage Table. As shown in Table 3: Category Percentage Table**Error! Reference source not found.** ‘Profilho’ made up 85% of the categories sold while ‘Dermafiller’ made up 8%. As seen in Table 4: Location Percentage Table 20% of the sales were made in ‘London’ followed by 16% made in ‘NorthEast, Midlands’.

Product	Percentage
Injection 7	66.44%
Injection 10	4.41%
Injection 3	3.89%
Cream 2	3.88%
Facemask	3.77%
Injection 1	3.68%
Injection 5	3.35%
Injection 2	2.74%
Foundation 2	1.53%
Injection 6	1.26%
Injection 4	1.09%
Foundation 1	0.64%
Injection 8	0.56%
Make-up Remover	0.43%
Skin Booster Needles	0.38%
Cream 3	0.35%
Injection 13	0.33%
Mesotherapy Needles	0.29%
Pills	0.27%
Injection 14	0.16%
Injection 9	0.11%
Sunblock	0.10%
Injection 11	0.08%
Cream 4	0.07%
Injection 12	0.05%
Cream 1	0.05%

Table 2: *Product Percentage Table*

Category	Percentage
Profilho	67.01%
Dermafiller	16.03%
Skincare	11.13%
Skinbooster	4.76%
Needles	0.68%
Mesotherapy	0.39%

Table 3: *Category Percentage Table*

Location	Percentage
NorthEast, Midlands	19.57%
NorthWest	15.40%
London	14.28%
SouthEast	10.92%
Scotland, Newcastle	10.42%
Outside London the rest	9.83%
SouthWest	9.56%
Outside London M25	4.78%
Ireland & Norhtern Ireland	4.68%
Brazil, USA, Blank, IE, Zimbabwe, Zurich	0.56%

Table 4: *Location Percentage Table*

The trend of the sales data was plotted against a time graph as seen in Figure 6: Time Graph. While this shows that the sales were increasing and growing by time, it also shows very sharp drop in sales during specific times in 2020 and 2021 which hints that sales were affected by extraordinary events.

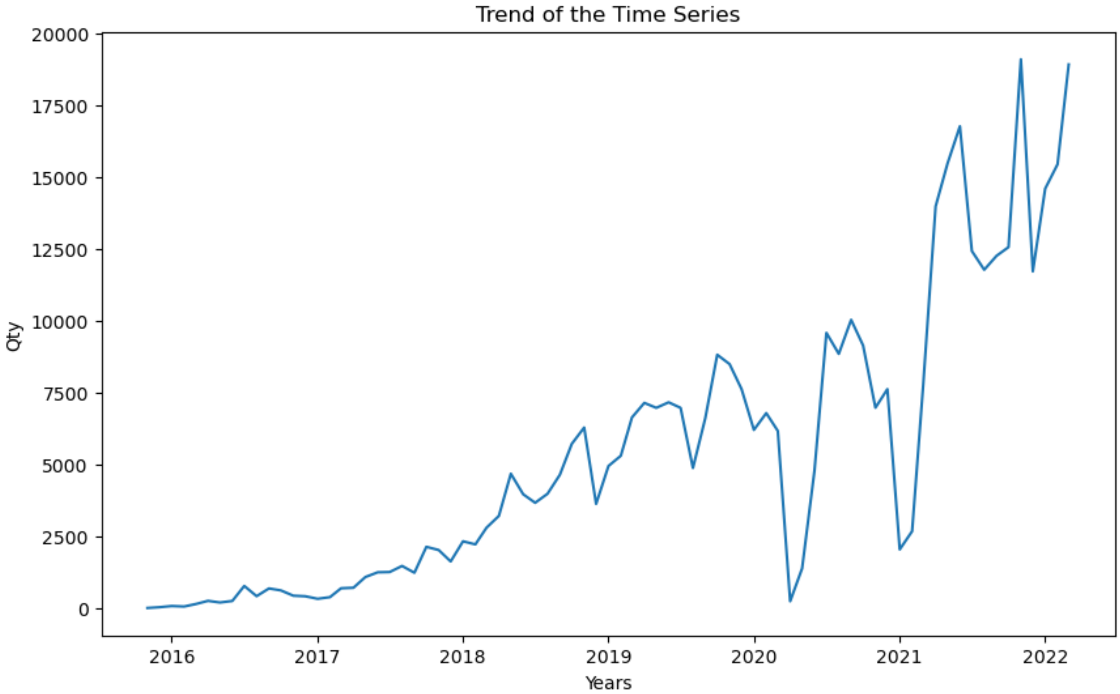

Figure 6: *Time Graph*

In the time graph plotted in Figure 7: Time Graph with legend, we can see that Profilho had very high sales over the years followed by the Dermafiller category. By the end of 2020 the Profilho sales spiked up and went back down during the rest of the data.

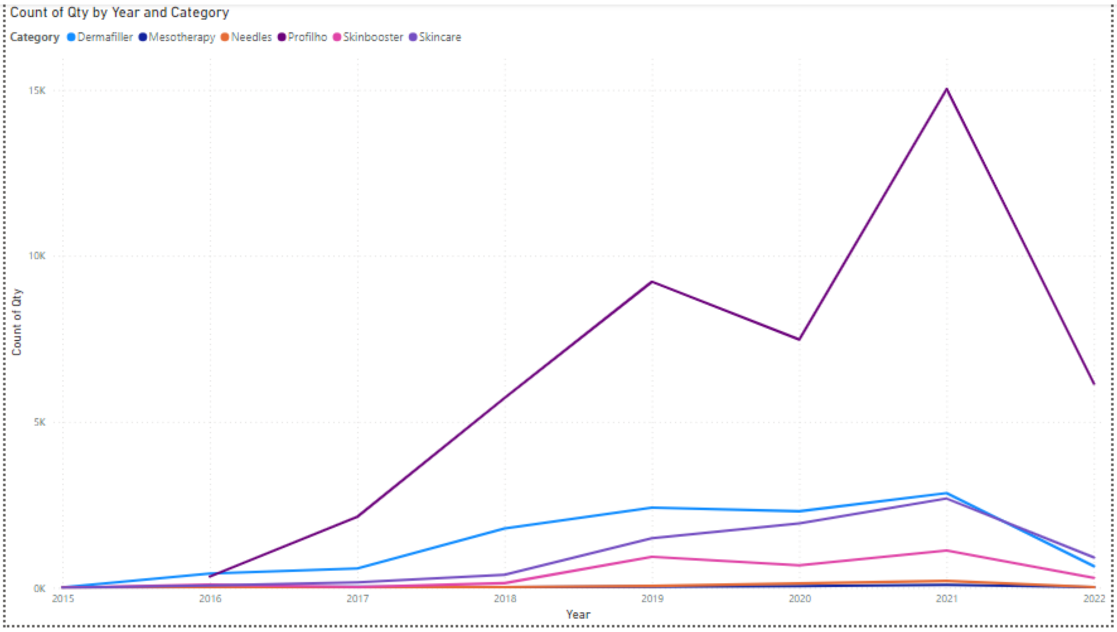

Figure 7: *Time Graph with legend*

Seasonality was also plotted against a time graph while also highlighting the different months and years of the sales. From the visual shown in **Figure 8:** Seasonality Graph**Figure 8:** Seasonality Graph, we can conclude that November seemed to have the highest number of sales as opposed to the month of April.

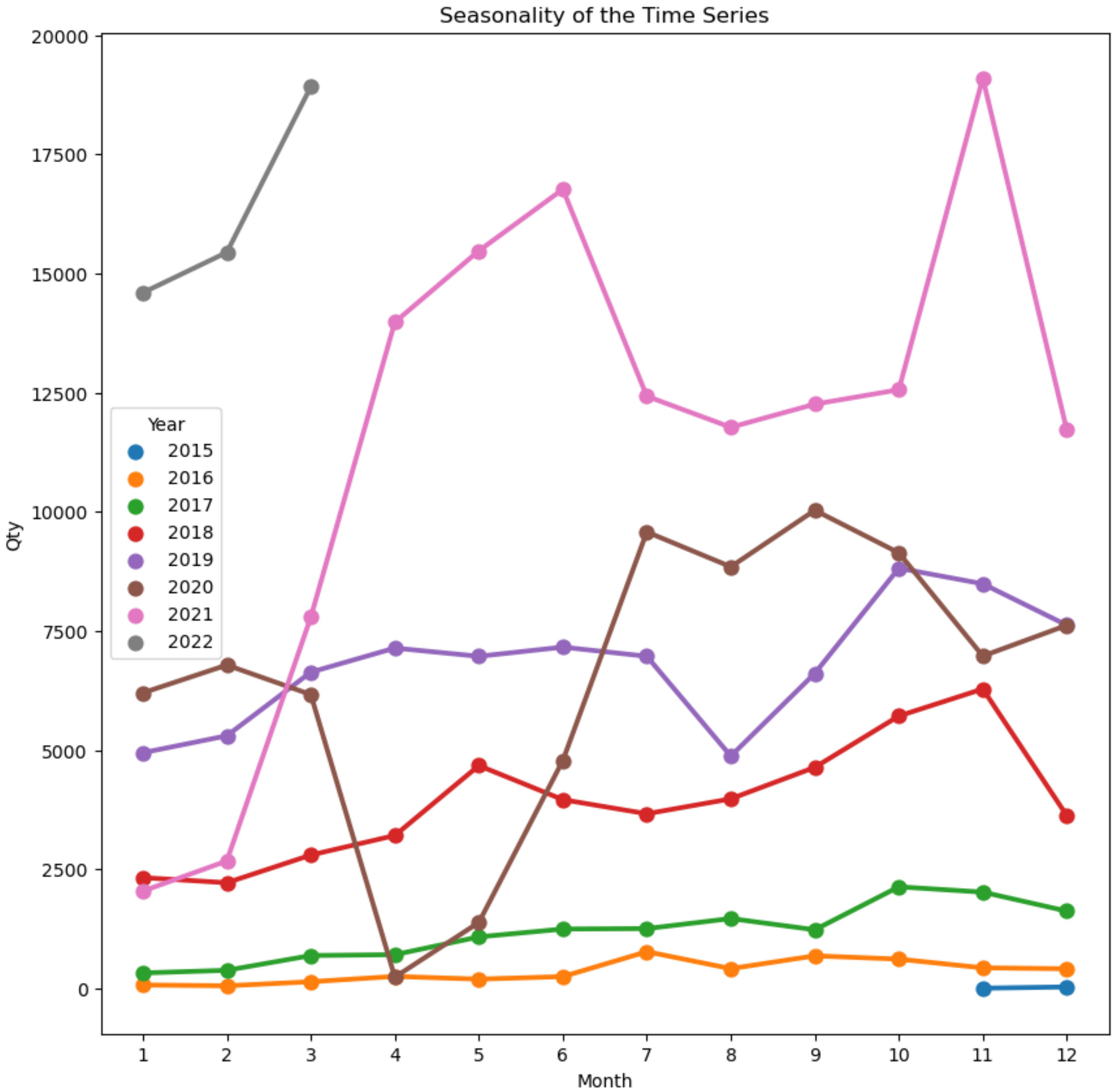

Figure 8: *Seasonality Graph*

In Figure 9: Different Graphs, the trend and seasonality are plotted against a year axis. These visuals help to analyse further if the extraordinary events did in fact have an effect on this sales data.

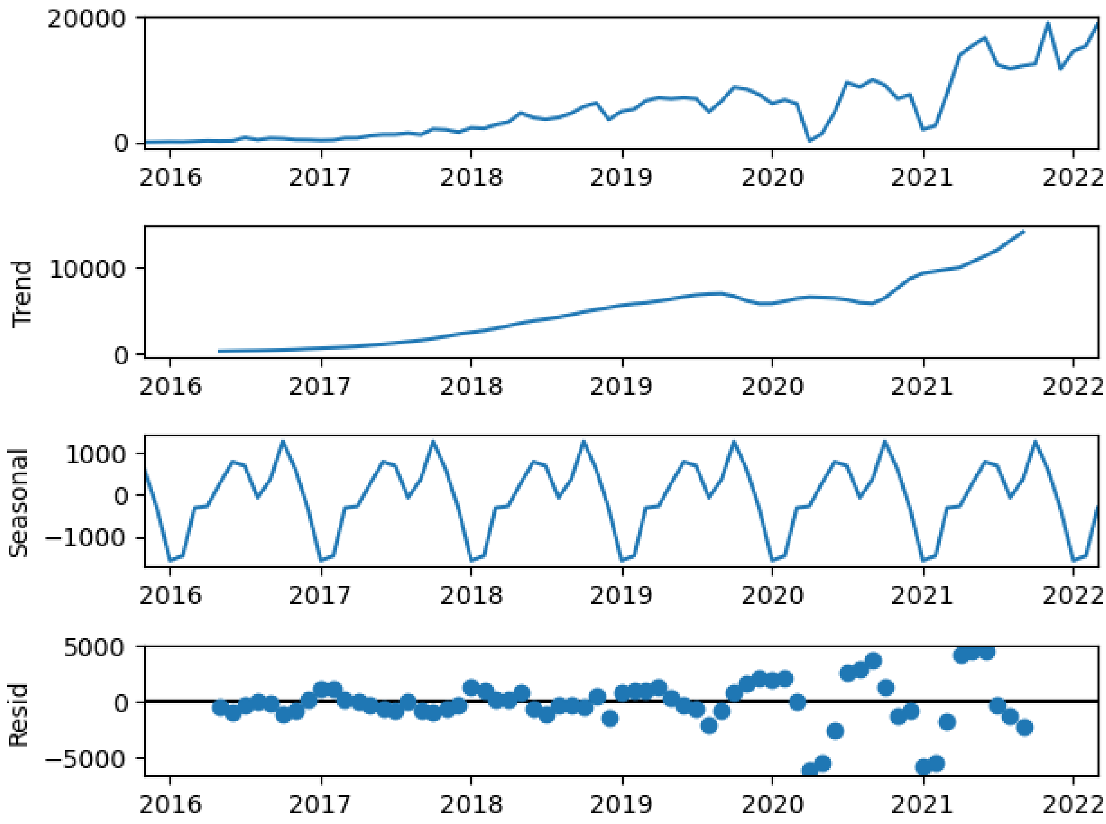

Figure 9: *Different Graphs*

3.5 Data Preparation for machine learning

To choose the forecasting algorithms, the literature review was reviewed, and three algorithms were chosen. The RMSE results from the literature review studies were also noted to be compared later to the results obtained from this study. To be able to forecast the results, the text data used in the forecasting needed to be changed to numerical values. For this reason, as suggested by the previous studies [19] and [20], one-hot-encoding was used on the categorical values in the columns ‘Category’, ‘Pharmacy’, ‘Product’, and ‘Location’. After the one-hot-encoding was performed, the data obtained needed to be grouped by dates. Since the data only includes the month of the sales, all the dates had to be grouped to the first day of the month so that each date is unique and could be used as an index when forecasting. The next step taken in this methodology was to create the lagged features in which the values at prior timesteps that were considered useful were stored. The lagged features go back up to seven time-steps, this means that for each category, the model includes the sales quantity values from the previous seven time-steps as features. The ultimate step taken before the splitting and training of models was to drop any null values which might be present in the data frame.

3.6 Training of Models

To train these models, five different types of data splitting were used. The first experiment was conducted on a summer period from the data. The dates which were selected were between the months of June and September for the year 2019. This type of experiment predicted the four selected months of the data and was made up of seven different experiments for each algorithm. First all the categories were used during the forecasting process, and afterward they were predicted one by one.

A second type of splitting was conducted on a winter period from the data. The testing dates were selected between October 2019 and December 2019. This type of experiment predicted the three selected months of the data and was also made up of seven different experiments for each algorithm like the summer experiment. The reason for choosing the summer and winter periods from this specific year is because 2019 was the last year before the Covid-19 pandemic infected the United Kingdom.

To have an overview of sales made during each month of the year 2019, the third type of splitting was conducted using one whole year as testing data and the dates chosen were between January 2019 and December 2019. This type of experiment predicted 11 months of the data one by one based on the past data from the previous years. The reason for having one less month in the results, was that January had to be removed from the testing data as it contained null values.

A different data splitting type that was used, was conducted on one whole year during the Covid-19 period which also predicted 11 months of data due to the same reason mentioned before. The testing dates selected were between April 2020 and April 2021, during which two enforced lockdowns happened.

The last data splitting was conducted by choosing the last four months present in the data which were between December 2021 and March 2022. January 2021 was also not included in the forecasting due to the null values. This was conducted to analyse whether the sales had returned to the pre-pandemic level of sales or not.

As suggested by the previously mentioned related studies [22] and [24], to test the forecasting of sales, in this study Random Forest and XGBoost algorithms were used. An additional supervised algorithm which was also used was Neural Networks.

3.7 Testing and Evaluation

To test these experiments, all the variations of the models were executed on cleaned test data. A total of 102 unique experiments were tried one at a time. The first test that was tried on each algorithm contained the sales of all of the six categories. The tests that followed, included trying each forecasting algorithm on each category one by one to be able to obtain results that make more sense since not all the categories make the same number of sales.

At the end of each experiment, the RMSE, MSE, R^2, Explained Variance and Max Error were calculated together with an 'Actual vs Predicted Sales Data' matrix to evaluate how the tests performed. The root-mean-square error (RMSE) is a commonly used measure of the differences between values predicted by a model and observed values. The mean-square-error (MSE) is the value of RMSE before finding the square root while the R^2 is the coefficient of determination. R^2 is a statistical measure that indicates the proportion of the variance explained by an independent variable in a regression model for a dependent variable. The Explained Variance score is comparable to the R^2 score, but it does not account for systematic offsets in the prediction while the Max Error metric calculates the maximum residual error in the regression model. From the results obtained from each experiment, an evaluation and conclusion were formulated.

Chapter 4: Analysis of Results and Discussion

This chapter reflects on the results obtained from the experiments that were tried out. The best parameters for each algorithm during each testing period can be found in the Appendix section. In the Appendix, all the results obtained for each category including the differences between the actual and predicted values can also be found. As shown in table 5, the Neural Network algorithm performed best in 14 of the 30 experiments, followed by XGBoost, which performed best in 11 experiments, and Random Forest performed best in just five.

	Summer	**Winter**	**Whole Year**	**Covid**	**Last 3 Months**
Dermafiller	27.96 Random Forest	24.26 XGBoost	97.84 XGBoost	183.05 XGBoost	67.94 Neural Network
Mesotherapy	1.67 Random Forest	4.58 XGBoost	3.93 XGBoost	1.85 Neural Network	0.98 Neural Network
Needles	1.98 XGBoost	17.72 Neural Network	15.65 XGBoost	22.03 XGBoost	1.90 Neural Network
Profilho	318.10 Neural Network	1320.98 XGBoost	1103.03 Neural Network	818.93 Neural Network	814.83 Neural Network
Skinbooster	33.95 Random Forest	51.87 Neural Network	61.91 Neural Network	21.42 Random Forest	23.83 XGBoost
Skincare	33.99 Neural Network	97.87 XGBoost	127.95 Random Forest	65.80 Neural Network	19.83 Neural Network

Table 5: *Periods-Category RMSE best results*

4.1 Summer Period

The summer period is based on the sales made during June 2019 to September 2019, which were all pre-pandemic sales. The results obtained during these times were quite good considering that only four months of the data were predicted. For this time period, the algorithm that obtained the best result was Random Forest with an RMSE as low as 1.67, obtained by the Mesotherapy category. This algorithm also performed best for three out of the six categories. For the Dermafiller category, the RMSE result obtained was 27.96 and the algorithm that performed the best was Random Forest Regression. Even though this category has the second highest number of sales in the data, the RMSE obtained is still not as high as some other categories. For the Needles category, the algorithm that performed the best was XGBoost with an RMSE of 1.98. The Profilho category obtained very high RMSE results compared to other categories, with an RMSE of 318.10, due to the extremely high number of sales. This category also tends to increase the growth of sales constantly, so the algorithm may not have performed the best due to this reason. Random Forest was the best algorithm for the Skinbooster category during the summer period with an RMSE of 33.95 which is neither the best, nor the worst. For the Skincare category, the best algorithm was Neural Networks with an RMSE of 33.99 which is very similar to the results obtained for the Skinbooster category.

4.2 Winter Period

The winter period is based on the sales made during October 2019 to December 2019, which were also all pre-pandemic sales. The results obtained during these times were also good considering that only three months of the data were predicted. For this time period, the algorithm that obtained the best result was XGBoost with an RMSE of 4.58, obtained by the Mesotherapy category. This algorithm also performed the best for four out of the six categories during the winter months. For the Dermafiller category during the winter period, the algorithm that performed the best was XGBoost with an RMSE of 24.26 which is very similar to the result obtained during the summer period. Neural Networks performed the best for the Needles category with an RMSE of 17.72, which is very high when compared to the result obtained during the summer period. For the Profilho category the RMSE obtained was 1320.98 by the XGBoost algorithm. This result was the highest RMSE from all the categories and also from all the time periods. This result may have been obtained poorly due to the reason that the pandemic having already started affecting some countries around the world and the purchasing of aesthetic products was not being considered a necessity. The algorithm that performed the best on the Skinbooster category was Neural Network with an RMSE of 51.87, which is considered satisfactory when compared to how the other algorithms performed on this category during this time period. XGBoost was the algorithm that performed the best on the Skincare category in this test data with an RMSE of 97.87.

4.3 Whole Year Period

The whole year period is based on the sales made between January 2019 and December 2019, which was the last pre-pandemic year. The predictions made during this period were also satisfactory since there were not any lockdowns present at that time yet. The lowest RMSE obtained for this data splitting was 3.93 obtained by the XGBoost algorithm for the Mesotherapy category. This algorithm performed the best for three out of the six categories during the 2019 year. For the Dermafiller category, the algorithm that performed the best was XGBoost with an RMSE of 97.84. This result is quite high when compared to the results of the other time periods obtained for this category excluding the covid year. The best RMSE result obtained by the Needles category was 15.65 and the algorithm that obtained this result was also XGBoost. This result was better than the result obtained during the winter period but not as good as the result of the summer period. For the Profilho category, the lowest RMSE obtained was 1103.03 by the Neural Network algorithm. This result was the second highest obtained result from the Profilho category following the winter period. This result shows that when the year 2019 was trained on the data between 2015 and the end of 2018, the sales did not reflect the past sales. The same reasoning applies for the Skinbooster and Skincare category as the RMSE is also considered a little high. For the Skinbooster category the best algorithm was Neural Network with an RMSE of 61.91 while for the Skincare category, the best algorithm was Random Forest with an RMSE of 127.95.

4.4 Covid-19 Period

The Covid-19 period is based on the sales made between April 2020 and April 2021, during which two lockdowns happened in the United Kingdom. The first lockdown took place between April 2020 and June 2020, while the second lockdown took place between January 2021 and March 2021.

4.4.1 Dermafiller

Dermafiller – Covid					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	206.86	42791.74	0.67	0.79	502.46
XGBoost	**183.05**	**33507.03**	**0.74**	**0.80**	**464.39**
Neural Network	188.81	35648.21	0.73	0.82	384.63

Table 6: *Algorithm Results – Dermafiller – Covid*

	Apr 2020	**May 2020**	**Jun 2020**	**Jul 2020**	**Aug 2020**	**Sep 2020**
Actual	16	80	392	1002	861	879
Predicted	89.17	116.35	372.24	726.48	685.13	633.03
Difference	-73.17	-36.35	19.76	275.52	175.87	245.97

	Oct 2020	**Nov 2020**	**Dec 2020**	**Feb 2021**	**Mar 2021**	**Apr 2021**
Actual	630	521	737	122	521	1182
Predicted	686.40	527.04	653.40	161.14	504.25	717.61
Difference	-56.4	-6.04	83.6	-39.14	16.75	464.39

Table 7: *Actual vs Predicted - Dermafiller - Covid*

As stated in Table 6: Algorithm Results – Dermafiller – Covid, for the Dermafiller category, XGBoost was the algorithm that performed the best with an RMSE of 183.05. Neural Network obtained a very close result with an RMSE of 188.81 while Random Forest performed the worst with an RMSE of 206.86.

In Table 7: Actual vs Predicted - Dermafiller - Covid, the actual sales and the predicted sales are shown together with the difference between them. In certain months, the algorithm predicted more sales than the actual sales, while in other months, the algorithm predicted less sales. For instance, in the month of April 2020 the algorithm predicted 73.17 more sales, while in the month of

April 2021 the algorithm predicted 464.39 less sales. For the months of April 2020 and May 2020, the algorithm predicted way more sales due to the test data being trained on the pre-pandemic data, while the test data occurred during the first lockdown. This shows that the sales for these two months decreased from sales made during the pre-pandemic months.

4.4.2 Mesotherapy

Mesotherapy – Covid					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	3.01	9.06	0.23	0.27	7.93
XGBoost	3.28	10.76	0.08	0.10	6.70
Neural Network	**1.85**	**3.44**	**0.71**	**0.77**	**3.64**

Table 8: *Algorithm Results – Mesotherapy – Covid*

	Apr 2020	**May 2020**	**Jun 2020**	**Jul 2020**	**Aug 2020**	**Sep 2020**
Actual	1	1	2	4	7	4
Predicted	1.18	2.80	3.55	6.05	7.49	6.92
Difference	-0.18	-1.8	-1.55	-2.05	-0.49	-2.92

	Oct 2020	**Nov 2020**	**Dec 2020**	**Feb 2021**	**Mar 2021**	**Apr 2021**
Actual	9	9	5	5	6	13
Predicted	7.25	8.88	8.64	6.92	5.29	11.56
Difference	1.75	0.12	-3.64	-1.92	0.71	1.44

Table 9: *Actual vs Predicted - Mesotherapy – Covid*

As stated in Table 8: Algorithm Results – Mesotherapy – Covid, for the Mesotherapy category, Neural Network was the algorithm that performed the best with a low RMSE of 1.85. Random Forest and XGBoost obtained a very close result with an RMSE of 3.01 and an RMSE of 3.28 respectively.

For this category, since the sales are very low, the predicted results in Table 9: Actual vs Predicted - Mesotherapy – Covid did not differ by much from the actual sales. This is the reason why the RMSE results obtained in Table 8: Algorithm Results – Mesotherapy – Covid was very small. The largest difference happened in the month of December 2020 with a difference of 3.64 more sales, while the best predicted result was in the month of November 2020 when there was a difference of only 0.12 less sales than the actual sales. After the first lockdown, the sales were as low as 1 while during the second lockdown, the sales were also quite low compared to the post-pandemic sales such as 13 in the month of April 2021.

4.4.3 Needles

Needles – Covid					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	22.69	514.94	0.24	0.43	71.09
XGBoost	**22.03**	**485.42**	**0.29**	**0.42**	**68.86**
Neural Network	30.75	945.70	-0.39	0.24	86.79

Table 10: *Algorithm Results – Needles – Covid*

	Apr 2020	**May 2020**	**Jun 2020**	**Jul 2020**	**Aug 2020**	**Sep 2020**
Actual	1	28	8	17	22	105
Predicted	2.12	9.72	11.82	15.03	31.04	36.14
Difference	-1.12	18.28	-3.82	1.97	-9.04	68.86

	Oct 2020	**Nov 2020**	**Dec 2020**	**Feb 2021**	**Mar 2021**	**Apr 2021**
Actual	45	17	36	9	32	41
Predicted	30.09	21.67	17.11	9.21	30.22	34.43
Difference	14.91	-4.67	18.89	-0.21	1.78	6.57

Table 11: *Actual vs Predicted - Needles – Covid*

As stated in Table 10: Algorithm Results – Needles – Covid, XGBoost was the best performing algorithm in the Needles category, with an RMSE of 22.03. Random Forest achieved an RMSE of 22.69, very identical to the XGBoost method, whereas Neural Network achieved an RMSE of 30.75.
During the first lockdown in the United Kingdom, the Needles category sales did not drop considerably. In Table 11: Actual vs Predicted - Needles – Covid there were hardly any high differences; the only one resulted a 68.86 difference for the Month of September 2020. Aesthetic product sales grew from prior year data, as seen by the fact that forecasted sales are substantially lower in value than actual sales. As an example, during the month of September 2020, the forecasted sales totalled 36.14, whereas the actual sales totalled 105. On the other hand, the algorithm predicted some correct outcomes, such as a forecasted outcome of 2.12 in April 2020 versus the actual sales of 1.

4.4.4 Profilho

Profilho – Covid					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	2501.59	6257969.26	0.39	0.72	6616.18
XGBoost	2228.09	4964391.25	0.51	0.75	5937.97
Neural Network	**818.93**	**670646.48**	**0.93**	**0.94**	**1498.09**

Table 12: *Algorithm Results – Profilho – Covid*

	Apr 2020	May 2020	Jun 2020	Jul 2020	Aug 2020	Sep 2020
Actual	168	1121	4008	7895	7215	8289
Predicted	1085.21	1781.72	4374.84	6396.91	6045.02	7643.40
Difference	-917.21	-660.72	-366.84	1498.09	1169.98	645.6

	Oct 2020	Nov 2020	Dec 2020	Feb 2021	Mar 2021	Apr 2021
Actual	7634	5786	6120	2372	6920	11870
Predicted	6684.58	5709.61	6484.80	2846.99	6240.24	10933.28
Difference	949.42	76.39	-364.8	-474.99	679.76	936.72

Table 13: *Actual vs Predicted - Profilho – Covid*

As stated in Table 12: Algorithm Results – Profilho – Covid, for the Profilho category, Neural Network was the algorithm that performed the best with an RMSE of 818.93 which is quite u high result. Random Forest and XGBoost algorithms performed very poorly with an RMSE result of 2501.59 and 2228.09 respectively.

The largest difference in sales happened in July 2020, as seen in Table 13: Actual vs Predicted - Profilho – Covid. This resulted in a 1498.09 difference. The Neural Network algorithm predicted 917.21 more sales than actual sales during the first month of testing. Profilho used to have a very high number of sales before the pandemic infected the United Kingdom, but when the lockdowns were implemented, sales for this category dropped dramatically. This effect primarily impacted April 2020, since sales were not as low in the remaining months. The smallest sales difference occurred in November 2020, with a difference as small as 76.39, followed by a difference of 364.8 in December 2020.

4.4.5 Skinbooster

Skinbooster – Covid					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	**21.42**	**458.81**	**0.96**	**0.96**	**45.95**
XGBoost	27.98	783.14	0.93	0.96	50.11
Neural Network	110.95	12309.57	-0.10	0.50	194.54

Table 14: *Algorithm Results – Skinbooster – Covid*

	Apr 2020	**May 2020**	**Jun 2020**	**Jul 2020**	**Aug 2020**	**Sep 2020**
Actual	14	39	65	217	267	223
Predicted	32.51	20.39	42.25	245.82	283.11	268.95
Difference	-18.51	18.61	22.75	-28.82	-16.11	-45.95

	Oct 2020	**Nov 2020**	**Dec 2020**	**Feb 2021**	**Mar 2021**	**Apr 2021**
Actual	322	195	271	54	136	310
Predicted	317.39	194.04	256.33	41.8	161.01	319.27
Difference	4.61	0.96	14.67	12.2	-25.01	-9.27

Table 15: *Actual vs Predicted - Skinbooster – Covid*

As stated in Table 14: Algorithm Results – Skinbooster – Covid, for the Skinbooster category, Random Forest was the algorithm that performed the best with an RMSE of 21.42. XGBoost obtained a very close RMSE result to Random Forest, as it obtained an RMSE of 27.98. In this type of experiment, Neural Network obtained the worst result with an RMSE of 110.95. The Random Forest RMSE obtained by this experiment is quite a good result considering the quantity of sales made in this category.

As stated in Table 15: Actual vs Predicted - Skinbooster – Covid, the difference between predicted and actual sales in this category was not significant. The greatest difference occurred in September 2020, when the algorithm predicted 45.95 more sales than actual sales. This outcome could possibly be related to the pandemic taking place in the United Kingdom at that time. In this experiment, the Random Forest algorithm predicted a fairly close outcome with only 0.96 difference in November 2020, when actual sales were 195 and forecasted sales were 194.04.

4.4.6 Skincare

Skincare – Covid					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	95.19	9061.22	0.72	0.79	193.77
XGBoost	67.16	4509.87	0.86	0.88	137.10
Neural Network	**65.80**	**4330.27**	**0.86**	**0.91**	**177.75**

Table 16: *Algorithm Results – Skincare – Covid*

	Apr 2020	**May 2020**	**Jun 2020**	**Jul 2020**	**Aug 2020**	**Sep 2020**
Actual	41	118	309	445	477	534
Predicted	45.43	89.57	321.67	622.75	569.31	582.48
Difference	-4.43	28.43	-12.67	-177.75	-92.31	-48.48

	Oct 2020	**Nov 2020**	**Dec 2020**	**Feb 2021**	**Mar 2021**	**Apr 2021**
Actual	502	448	453	116	178	564
Predicted	546.42	497.21	511.19	132.28	160.05	575.83
Difference	-44.42	-49.21	-58.19	-16.28	17.95	-11.83

Table 17: *Actual vs Predicted - Skincare – Covid*

As stated in Table 16: Algorithm Results – Skincare – Covid, for the Skincare category, Neural Network was the algorithm that performed the best with an RMSE of 65.80. XGBoost obtained a very close RMSE result to Neural

Network, as it obtained an RMSE of 67.16. Random Forest produced the worst results in this type of experiment, with an RMSE of 95.19.

The lockdown months did not produce a large number of differences, as shown in Table 17: Actual vs Predicted - Skincare – Covid. This does not necessarily imply that the Covid-19 epidemic had no effect on this category's sales, because in a number of instances, actual sales were lower than predicted. There was a substantial difference in between the lockdown periods which lasted from July through December 2020. For instance, the biggest difference occurred in July 2020, when the algorithm forecasted 177.75 more sales than actual sales. On the other hand, the smallest difference occurred in April 2020, when the algorithm forecasted only 4.43 more sales than actual sales.

4.5 Last 3 Months Period

The last 3 months period is based on the sales made between December 2021 and March 2022, which were the last few months of the data. The results obtained during this time period were quite good and even contained the lowest RMSE obtained from all the experiments.

4.5.1 Dermafiller

Dermafiller – Last 3 Months					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	130.93	17142.74	-0.47	0.27	218.95
XGBoost	126.66	16042.45	-0.38	0.52	207.96
Neural Network	**67.94**	**4616.01**	**0.60**	**0.75**	**117.54**

Table 18: *Algorithm Results - Dermafiller - Last 3 Months*

	Dec 2021	Feb 2022	Mar 2022
Actual	623	711	883
Predicted	617.29	711.06	765.46
Difference	5.71	-0.06	117.54

Table 19: *Actual vs Predicted - Dermafiller – Last 3 Months*

According to Table 18: Algorithm Results - Dermafiller - Last 3 Months, Neural Network was the best performing algorithm in the Dermafiller category, with an RMSE of 67.94. XGBoost came in second with an RMSE of 126.66, while Random Forest came in last with an RMSE of 130.93.

According to Table 19: Actual vs Predicted - Dermafiller – Last 3 Months, the experiment's smallest difference occurred in February 2022, with a difference of only 0.06. The algorithm forecasted 117.54 less sales in March 2022 than actual sales, indicating that sales began to increase again after the pandemic.

4.5.2 Mesotherapy

Mesotherapy – Last 3 Months					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	0.99	0.98	-0.46	-0.43	1.50
XGBoost	1.74	3.02	-3.53	-3.33	2.30
Neural Network	**0.98**	**0.96**	**-0.44**	**0.41**	**1.48**

Table 20: *Algorithm Results - Mesotherapy - Last 3 Months*

	Dec 2021	Feb 2022	Mar 2022
Actual	8	7	9
Predicted	6.52	7.25	9.80
Difference	1.48	-0.25	-0.8

Table 21: *Actual vs Predicted - Mesotherapy– Last 3 Months*

According to Table 20: Algorithm Results - Mesotherapy - Last 3 Months, the Neural Network algorithm performed the best in the Mesotherapy category, with a very low RMSE of 0.98. Random Forest came in second to Neural Network with an RMSE of 0.99, while XGBoost came in last with an RMSE of 1.74.

According to Table 21: Actual vs Predicted - Mesotherapy– Last 3 Months, the predicted results were quite close to the actual values, with a difference as small as 0.25. Therefore, in conclusion, the Mesotherapy category sales did not rise or decrease following the Covid-19 lockdowns.

4.5.3 Needles

Needles – Last 3 Months					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	3.17	10.05	0.38	0.82	4.52
XGBoost	5.35	28.65	-0.77	0.72	6.78
Neural Network	**1.90**	**3.62**	**0.78**	**0.99**	**2.34**

Table 22: *Algorithm Results - Needles - Last 3 Months*

	Dec 2021	**Feb 2022**	**Mar 2022**
Actual	8	9	17
Predicted	6.18	6.66	15.56
Difference	1.82	2.34	1.44

Table 23: *Actual vs Predicted - Needles – Last 3 Months*

The Neural Network algorithm performed the best in the Needles category, according to Table 22: Algorithm Results - Needles - Last 3 Months, with a low RMSE of 1.90. Random Forest finished second to Neural Network with an RMSE of 3.17, and XGBoost finished last with an RMSE of 5.35.

According to Table 23: Actual vs Predicted - Needles – Last 3 Months, the predicted values were slightly lower than the actual ones. As a result, sales in the Needles category began to slightly rise at the end of the sales data.

4.5.4 Profilho

Profilho – Last 3 Months					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	1751.73	3068542.82	0.61	0.62	2444.21
XGBoost	1990.93	3963801.21	0.49	0.70	2644.16
Neural Network	**814.83**	**663947.75**	**0.92**	**1.00**	**963.92**

Table 24: *Algorithm Results - Profilho - Last 3 Months*

	Dec 2021	**Feb 2022**	**Mar 2022**
Actual	10237	13809	17092
Predicted	10893.38	14603.91	18055.92
Difference	-656.38	-794.91	-963.92

Table 25: *Actual vs Predicted - Profilho – Last 3 Months*

According to Table 24: Algorithm Results - Profilho - Last 3 Months, the Neural Network method performed the best in the Profilho category, with an RMSE of 814.83, which is still extremely high. Random Forest came in second place to Neural Network with an RMSE of 1751.73, while XGBoost came in last with an RMSE of 1990.93.

According to Table 25: Actual vs Predicted - Profilho – Last 3 Months, the Neural Network's forecasted values were not close to the actual sales. As a result, one may claim that Profilho sales have not yet returned to pre-pandemic levels, although this may not be accurate given the category's enormous number of sales.

4.5.5 Skinbooster

Skinbooster – Last 3 Months					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	41.11	1690.34	0.60	0.68	55.77
XGBoost	**23.83**	**568.03**	**0.86**	**0.91**	**40.42**
Neural Network	50.13	2513.23	0.40	0.54	77.29

Table 26: *Algorithm Results - Skinbooster - Last 3 Months*

	Dec 2021	**Feb 2022**	**Mar 2022**
Actual	205	327	354
Predicted	209.49	286.58	346.94
Difference	-4.49	40.42	7.06

Table 27: *Actual vs Predicted - Skinbooster – Last 3 Months*

According to Table 26: Algorithm Results - Skinbooster - Last 3 Months, the XGBoost algorithm performed the best in the Skinbooster category, with an RMSE of 23.83. Random Forest and Neural Network obtained similar results in this experiment, with RMSEs of 41.11 and 50.13, respectively.

According to Table 27: Actual vs Predicted - Skinbooster – Last 3 Months, the experiment's smallest difference occurred in December 2021, with a difference of only 4.49, while the algorithm predicted 40.42 fewer sales than actual sales in February 2022. This indicates that the Skinbooster sales began to rise again following the pandemic.

4.5.6 Skincare

Skincare – Last 3 Months					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	26.52	703.51	0.29	0.29	34.86
XGBoost	43.01	1849.59	-0.86	0.53	61.08
Neural Network	**19.83**	**393.31**	**0.60**	**0.61**	**25.42**

Table 28: *Algorithm Results - Skincare - Last 3 Months*

	Dec 2021	**Feb 2022**	**Mar 2022**
Actual	638	582	564
Predicted	639.85	607.42	540.97
Difference	-1.85	-25.42	23.03

Table 29: *Actual vs Predicted - Skincare – Last 3 Months*

According to Table 28: Algorithm Results - Skincare - Last 3 Months, Neural Network was the best performing algorithm in the Skincare category, with an RMSE of 19.83. Random Forest came close to Neural Network in terms of RMSE, with a score of 26.52. XGBoost produced the worst results in this experiment, with an RMSE of 43.01.

According to Table 29: Actual vs Predicted - Skincare – Last 3 Months, the Neural Network's forecasted values were not all close to actual sales. This indicates that the Skincare category has not yet recovered its pre-pandemic sales levels.

4.6 Research questions results

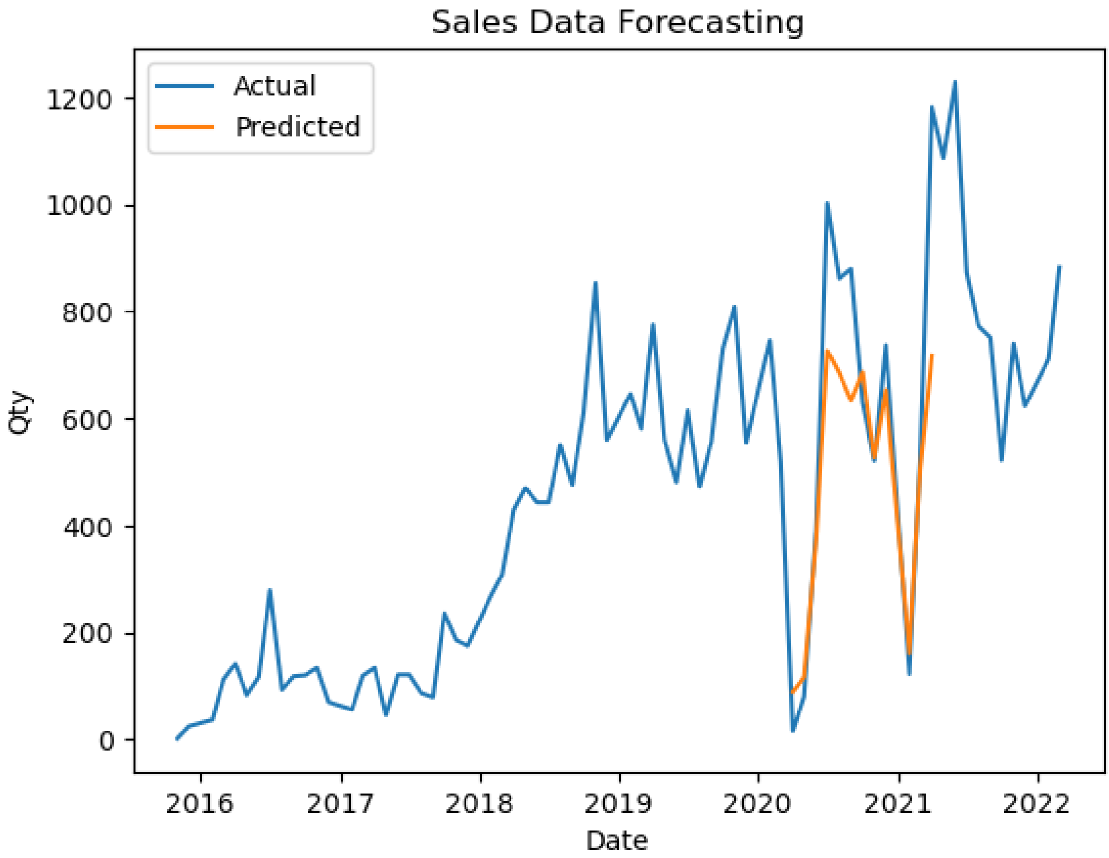

Figure 10: *Covid Forecast*

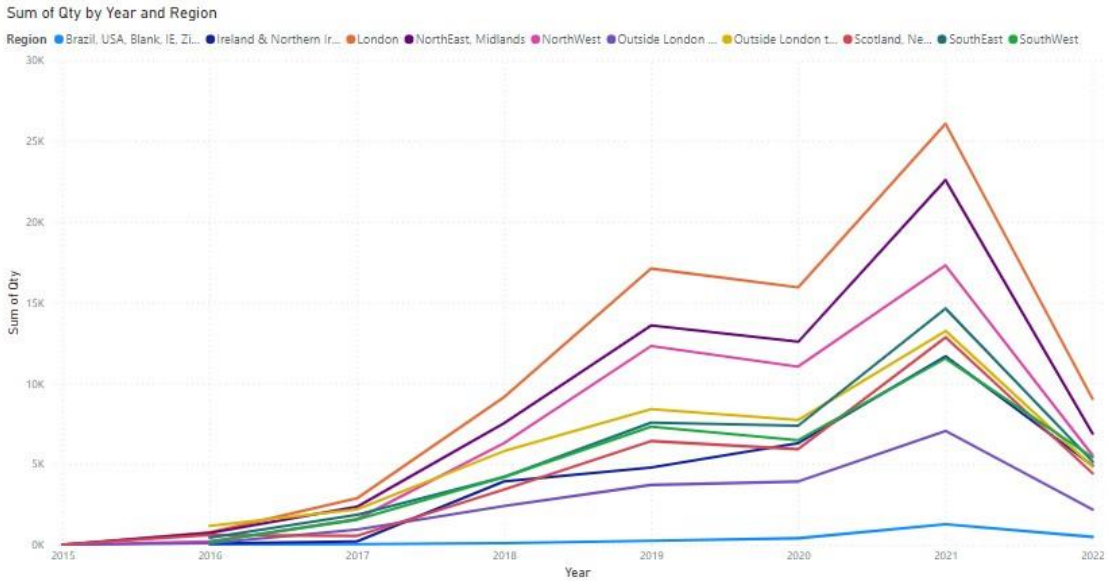

Figure 11: *Sales Time Graph by Regions*

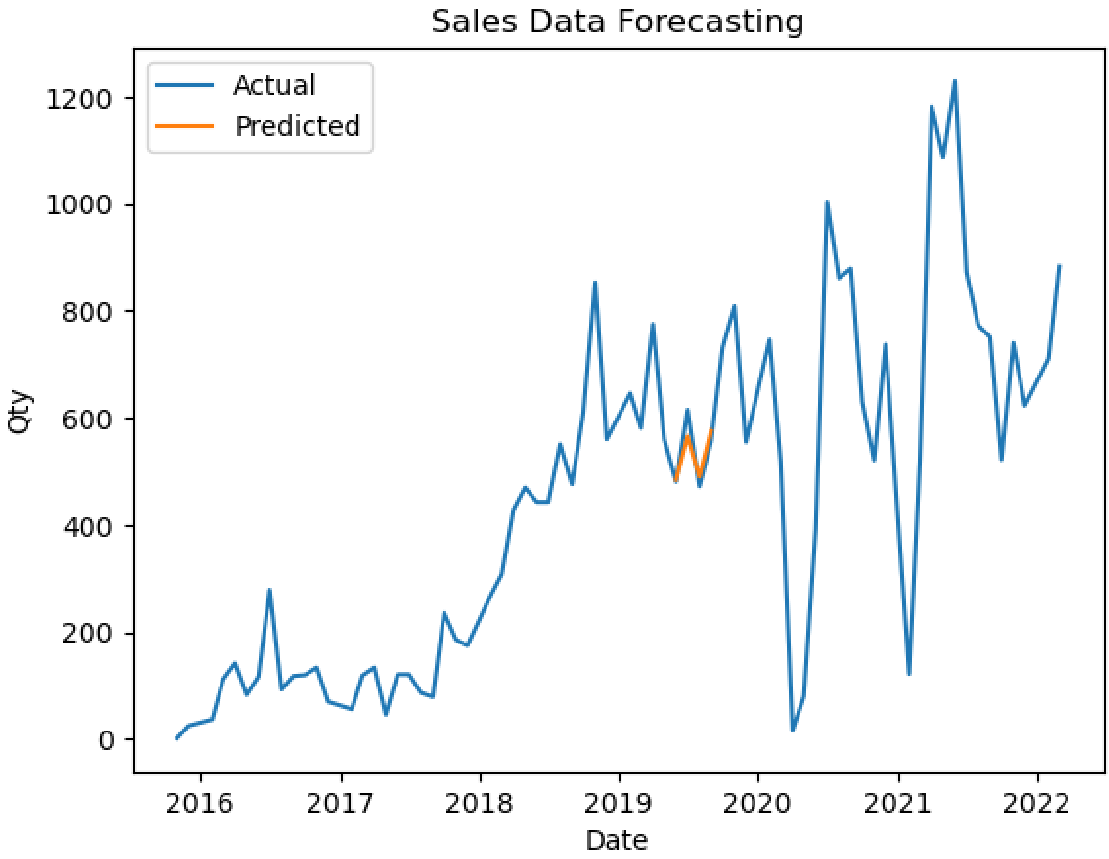

Figure 12: *Summer Prediction*

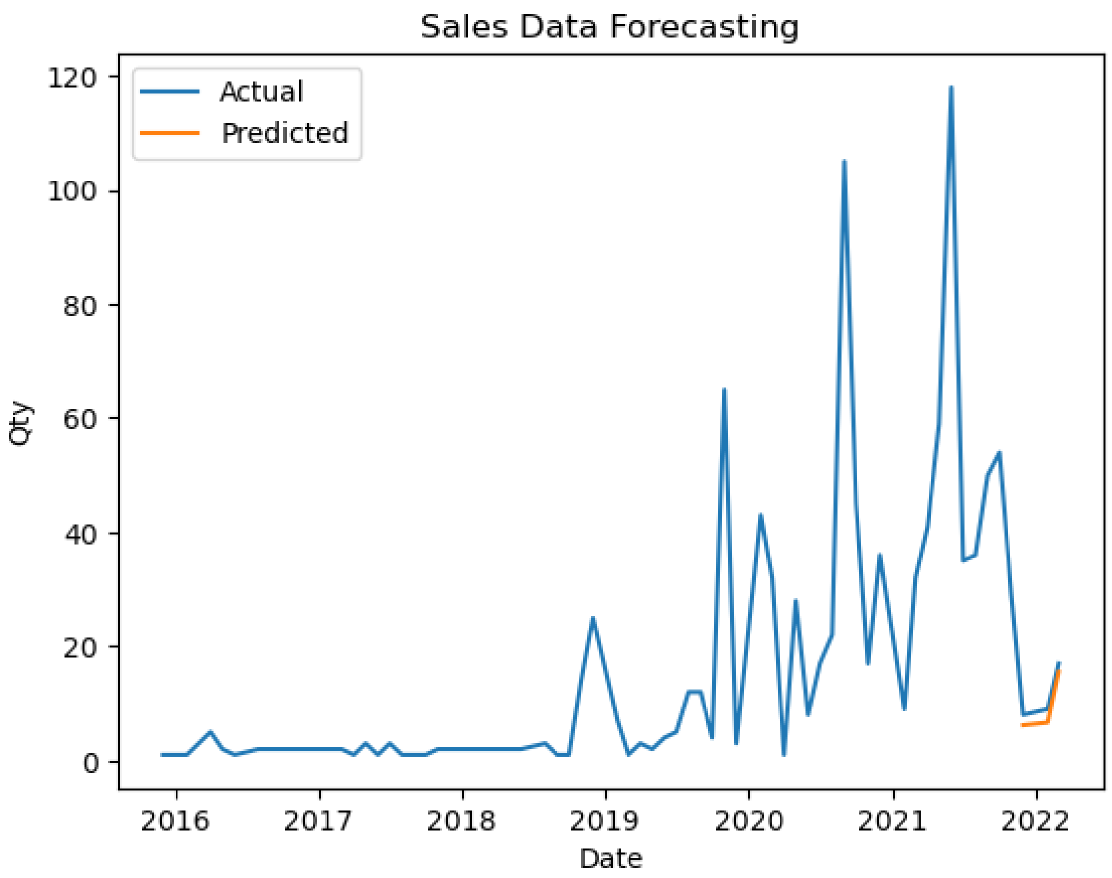

Figure 13: *Last 3 Months Prediction*

To answer research questions about the pandemic period, the year in which two lockdowns occurred in the United Kingdom was used as testing data. According to the findings, Covid-19 did have an effect on the sales of aesthetic products. Figure 10: Covid Forecast, depicts this result, which illustrates that the forecast was incorrectly predicted due to an unusual event which does not follow a trend or seasonality.

According to the findings of this research, Covid-19 had approximately the same impact on all regions in the United Kingdom, as shown in Figure 11: Sales Time Graph by Regions between the years 2020 and 2021.

Finally, it was discovered that machine learning algorithms can be used to predict supply based on previous transactions, as illustrated in Figure 12:

Summer Prediction and Figure 13: Last 3 Months Prediction, where the prediction line is very close to the actual sales line.

4.7 Comparing findings to previous studies

When the results of the study were compared to previous studies, the algorithms were relatively comparable. The RMSEs values differed because the amount of data gathered for this investigation was different from the amount of data used in the related studies. Akanksha at al. [22] and Sajawal at al. [24], both indicated that XGBoost outperformed Random Forest, which is in line with our findings.

4.8 Additional findings

According to the data, the biggest increase of 6187 sales occurred between March 2021 and April 2021, when there was a total of 7793 and 13980 sales, respectively, and the second United Kingdom lockdown had just ended. According to the findings, sales began to gradually grow again at the end of the sales data. Finally, after analysing the results, it was also found that November had the most total sales at 43,261, which was 69.73% greater than April, which had the lowest total sales at 25488.

Chapter 5: Conclusion and Recommendations

This study aimed to identify the effects of Covid-19 on products and regions in the United Kingdom. It also aimed to investigate whether machine learning algorithms can be used to predict supply based on previous transactions. Based on a quantitative study of the sales data provided, it is possible to conclude that Covid-19 had an unfavourable effect on aesthetic product sales. This resulted in a drop in sales between 2020 and 2021. As a result of the difference in pre-pandemic and the sales made during the pandemic, the research methods used during this time period were not appropriately predicted. On the other hand, most of the experiments conducted on pre-pandemic sales were accurate, particularly in small categories such as Mesotherapy.

This analysis also concluded that the Covid-19 pandemic had roughly the same impact on all regions in the UK, since if one region's sales declined, so did the others, and vice versa. This discovery was made through using visuals to view the sales in a more effective way.

The last conclusion made from this research was about the use of machine learning algorithms when it comes to predicting supply based on previous transactions. Since the forecasting model correctly predicted the outcome in circumstances where no unusual events happened, we may conclude that machine learning algorithms are superior to statistical models for predicting supply. This can therefore be used to predict the supply of upcoming seasons to avoid over-ordering of products or not reaching customer demands. As a result, this can be used to forecast supply for upcoming months with a high level of confidence. Since for this study three models were used, these could be compared between each other and evaluate which model works best on this type of data. For this study, Neural Network algorithms performed best, followed by XGBoost and Random Forest Respectively. The RMSEs obtained by the Neural Network algorithm were as low as 0.98 for the Mesotherapy category, while for XGBoost, the lowest RMSE was 1.98 for the Needles category. For the Random Forest, the lowest RMSE obtained was of 1.67 for the Mesotherapy category.

The research methodology chosen was suitable for this study since it provided important insights that aided in answering the research questions proposed and confirmed the hypothesis formulated. The goal of this study, which used machine learning, was to determine how close the predicted outcomes were to the actual sales. The results obtained were as expected, as throughout the Covid-19 period, the predictions were not accurate in some tests while they were fairly accurate for others. Data cleaning was a very important step in the methodology since the data needed to be altered in a way that could be used in the prediction algorithms. One-hot-encoding was also a very important technique to use to change data from categorical to numerical without causing undesired issues such as altering the value of the data. This was critical to the study because the experiments were based on both sales amounts and category predictors. While random splitting methods do provide a sales prediction, it limits the ability to forecast a specific time period. As a result, the data experiments for this study were broken down into testing periods based on specific dates and the forecast results were between the specified dates. The testing experiments tried in this research were also sufficient to answer the research questions stated at the beginning of the study. There was no need to explore more prediction variations at this stage, as certain questions were also answered by the visuals created. During this study, additional findings were discovered that might be utilised by the company in terms of the most popular months, categories, and products. In the most recent experiment, it was also discovered that sales levels began to revert to pre-pandemic levels, and they are increasing again.

The limitations of this research in terms of sales data obtained includes a lack of recent data. Since the data's latest month was March 2022, this did not include for example the summer period after the pandemic. Therefore, experiments could not be conducted to compare if the seasonal sales have returned to their original state. The only post-pandemic sales in the data are between April 2021 and March 2022, which are not enough to train the models again. To train the post-pandemic data and be able to make evaluations, at least two full years after April 2022 are needed.

Another limitation of the data is that when the data was inputted into the system, the date did not include the day of when the transaction was made. To be able to perform forecasting using machine learning algorithms, the dates need to be unique. As a result, instead of grouping the dates based on the standard format of date variable, the data had to be grouped by month to produce the unique date needed.

Another obstacle encountered throughout this investigation was the short time available to accomplish the research. This study, like every other dissertation, was bound by a deadline, thus the research had to be well planned and organised to finish and reach the intended results on time. There was not enough time to expand and analyse the research beyond than what was intended. If the research timeline was based on a longer period, more areas of the study could have been evaluated. For example, other suggested regression models could have been tried out while also exploring other time periods such as choosing seasons from other years. On the other hand, the methodology process taken in this study was enough to answer the research questions and some other additional findings.

One last limitation of this study includes the lack of research available on predicting supply and demand for aesthetic products using machine learning algorithms. In this research field, multiple studies on forecasting supply using machine learning can be found, but very little research is directly conducted on the ranges of aesthetic products. Most of the studies focused on forecasting the supply of aesthetic products mostly use statistical models such as ARIMA. There was no study that chose to attempt Neural Network methods, but in this situation, it was the best performing model.

Based on these conclusions, if future research is to be attempted, the researcher could try out a variation of experiments while also gathering more data. The researcher can attempt to collect the most recent data from the company to generate more real-time predictions and forecast sales numbers for the following months. With the data provided, predictions can still be made but they will be even more accurate if more recent data is used. Initial experiment can be made using the best performing algorithm which for this data was Neural Network, and then try other algorithms.

The best performing algorithm for this data, Neural Network, can be used for the initial experiments, and subsequently alternative algorithms can be tried. In future research, other algorithms suggested by the studies reviewed can also be tried out such as ARIMA, SARIMA, SVM and LSTM. Int the other studies, when it came to estimating sales data, these models also produced accurate outcomes, therefore they should be tried out to obtain an RMSE as close to zero as possible.

A variety of different experiments can also be tried out, such as splitting the data by products instead of splitting by the categories. Forecasts will be more focused on specific products, and forecasts could be modified to each, and every product sold by the company. With the results obtained from such experiments, the ordering of products can be more easily made.

On the other hand, if the experiments are to be tried out on a different dataset, the best working algorithm for that data should be found out first, then further predictions could be made, since there is no one optimal algorithm for all the series.

List of References

[1] H. Wei and Q. Zeng, “Research on sales Forecast based on XGBoost-LSTM algorithm Model,” *Journal of Physics: Conference Series,* vol. 1754, no. 012191, 2021.

[2] A. Lohani, A. Verma, H. Joshi, N. Yadav and N. Karki, “Nanotechnology-Based Cosmeceuticals,” *ISRN Dermatology,* vol. 2014, no. 843687, pp. 1-14, 2014.

[3] J. M. C. d. Mesquita and H. C. Martins, “Retail industry: seasonality in sales, and financial results,” *Brazilian Business Review,* vol. 8, no. 3, pp. 64-82, 2011.

[4] S. A. Al-Thaqeb, B. G. Algharabali and K. T. Alabdulghafour, “The pandemic and economic policy uncertainty,” *Int J Fin Econ,* vol. 27, no. 3, pp. 2784-2794, 2022.

[5] R. Fairlie and F. M. Fossen, “The early impacts of the COVID-19 pandemic on business sales,” *Nature Public Health Emergency Collection* , vol. 58, no. 4, 2021.

[6] “GDP and events in history: how the COVID-19 pandemic shocked the UK economy,” Office for National Statistics, 23 May 2022. [Online]. [Accessed 4 February 2023].

[7] N. Bloom, P. Bunn, S. Chen, P. Mizen, P. Smietanka and G. Thwaites, “THE IMPACT OF BREXIT ON UK FIRMS,” NATIONAL BUREAU OF ECONOMIC RESEARCH, Cambridge, 2019.

[8] J. Li, "A Feature Engineering Approach for Tree-based Machine Learning Sales Forecast, Optimized by a Genetic Algorithm Based Sales Feature Framework," IEEE, Chengdu, China, 2022.

[9] R. Brown, "Becoming Human: Artificial Intelligence Magazine," Medium, 4 December 2019. [Online]. Available: https://becominghuman.ai/where-is-artificial-intelligence-used-today-3fd076d15b68. [Accessed 10 January 2023].

[10] G. Bonaccorso, Machine Learning Algorithms, Birmingham: Packt Publishing Ltd, 2017.

[11] P. Popovski, G. Veljanovski, M. Kostov and M. Atanasovski, "Optimizing Short Term Load Forecast: A study on Machine Learning Model Accuracy and Predictor Selection," IEEE, Ohrid, North Macedonia , 2022.

[12] R. Deb, "What are Machine Learning Applications? Top 10 Industry and Real-World Use Cases," Emeritus Online Courses, 19 December 2022. [Online]. Available: https://emeritus.org/blog/machine-learning-what-are-machine-learning-applications/. [Accessed 10 January 2023].

[13] K. R. Dalal, "Analysing the Role of Supervised and Unsupervised Machine Learning in IoT," IEEE, Coimbatore, India, 2020.

[14] F. Petropoulos, D. Apiletti, V. Assimakopoulos and M. Z. Babai, "Forecasting: theory and practice," *International Journal of Forecasting,* vol. 38, no. 3, 2022.

[15] D. J. Dalrymple, "Sales forecasting methods and accuracy," *Business Horizons,* vol. 18, no. 6, pp. 69-73, 1975.

[16] B. M. Pavlyshenko, “Machine-Learning Models for Sales Time Series Forecasting,” *Data,* vol. 4, no. 15, 2019.

[17] R. Khandelwal, “Step by Step Time Series Analysis,” Medium, 11 September 2019. [Online]. Available: https://medium.datadriveninvestor.com/step-by-step-time-series-analysis-d2f117554d7e. [Accessed 1 February 2023].

[18] R. J. Hyndman and G. Athanasopoulos, “Forecasting: Principles and Practice (2nd ed),” OTexts, April 2018. [Online]. Available: https://otexts.com/fpp2/tspatterns.html . [Accessed 1 February 2023].

[19] T. Al-Shehari and R. A. Alsowail, “An Insider Data Leakage Detection Using One-Hot Encoding, Synthetic Minority Oversampling and Machine Learning Techniques,” *Entropy ,* vol. 23, no. 10, 2021.

[20] M. S. Hossain and H. Mahmood, “Short-Term Load Forecasting Using an LSTM Neural Network,” IEE, Champaign, IL, USA, 2020.

[21] A. K. Sharma, M. Kiran, P. P. S. Jeba, P. Maheshwari and V. Divakar, “Demand Forecasting Using Coupling Of Machine Learning And Time Series Models For The Automotive After Market Sector,” IEEE, Mysuru, India , 2022.

[22] A. Akanksha, D. Yadav, D. Jaiswal, A. Ashwani and A. Mishra, “Store-sales Forecasting Model to Determine Inventory Stock Levels using Machine Learning,” IEEE, Nepal, 2022.

[23] J.-C. Huang, M.-H. Shu, B.-M. Hsu and T.-J. Wu, “A Novel Revenue Development and Forecasting Model using Machine Learning,” ijssst, China, 2016.

[24] M. Sajawal, S. Usman, H. S. Alshaikh, A. Hayat and M. U. Ashraf, “A Predictive Analysis of Retail Sales Forecasting using Machine Learning Techniques,” LGURJCSIT, 2022.

[25] C. Vithitsoontorn and P. Chongstitvatana, “Demand Forecasting in Production Planning for Dairy Products Using Machine Learning and Statistical Method,” IEEE, Thailand, 2022.

[26] J. Wang, G. Q. Liu and L. Liu, “A Selection of Advanced Technologies for Demand Forecasting in the Retail Industry,” IEEE, China, 2019.

Appendices

1. Summer Period

1.1 All Categories Summer

All Categories – Summer					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	699.14	488800.48	0.40	0.53	950.69
XGBoost	594.42	353333.23	0.57	0.61	811.06
Neural Network	**452.28**	**204555.26**	**0.75**	**0.79**	**830.28**

	Jun 2019	**Jul 2019**	**Aug 2019**	**Sept 2019**
Actual	7164	6971	4878	6613
Predicted	6333.72	6875.05	5199.93	6486.43
Difference	830.28	95.95	-321.93	126.57

Best Parameters:

- Random Forest:
  Criterion: poisson
  Max_depth: 10
  Max_features: sqrt
  Min_samples_leaf: 1

- XGBoost:
  Criterion: friedman_mse
  Loss: squared_error
  Max_depth: 5
  Max_features: log2
  Min_samples_leaf: 2

- Neural Network:
  Activation: identity
  Solver: lbfgs

1.2 Dermafiller Summer

Dermafiller – Summer					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	**27.96**	**781.67**	**0.77**	**0.77**	**49.17**
XGBoost	42.16	1777.74	0.48	0.49	70.54
Neural Network	53.79	2893.14	0.16	0.20	84.72

	Jun 2019	**Jul 2019**	**Aug 2019**	**Sept 2019**
Actual	487	615	472	557
Predicted	484.82	565.83	490.07	576.18
Difference	2.18	49.17	-18.07	-19.18

Best Parameters:

- Random Forest:
  Criterion: friedman_mse
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 1

- XGBoost:
  Criterion: friedman_mse
  Loss: squared_error
  Max_depth: 5
  Max_features: sqrt
  Min_samples_leaf: 2

- Neural Network:
  Activation: identity
  Solver: lbfgs

1.3 Mesotherapy Summer

Mesotherapy – Summer					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	**1.67**	**2.79**	**-0.28**	**-0.25**	**2.48**
XGBoost	2.21	4.88	-1.23	-1.20	3.17
Neural Network	3.14	9.83	-3.49	-1.57	6.14

	Jun 2019	Jul 2019	Aug 2019	Sept 2019
Actual	4	3	6	2
Predicted	5.37	3.43	3.52	3.72
Difference	-1.37	-0.43	2.48	-1.72

Best Parameters:

- Random Forest:
  Criterion: squared_error
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 1

- XGBoost:
  Criterion: friedman_mse
  Loss: squared_error
  Max_depth: 10
  Max_features: sqrt
  Min_samples_leaf: 1

- Neural Network:
  Activation: logistic
  Solver: lbfgs

1.4 Needles Summer

Needles – Summer					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	2.00	3.99	0.72	0.86	3.41
XGBoost	**1.98**	**3.93**	**0.72**	**0.83**	**3.51**
Neural Network	4.80	23.05	-0.62	0.66	6.71

	Jun 2019	**Jul 2019**	**Aug 2019**	**Sept 2019**
Actual	4	5	12	12
Predicted	3.32	3.51	12.84	8.49
Difference	0.68	1.49	-0.84	3.51

Best Parameters:

- Random Forest:
  Criterion: friedman_mse
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 1

- XGBoost:
  Criterion: friedman_mse
  Loss: squared_error
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 1

- Neural Network:
  Activation: tanh
  Solver: adam

1.5 Profilho Summer

Profilho – Summer					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	517.28	267576.27	0.59	0.77	741.03
XGBoost	423.05	178973.35	0.73	0.75	565.15
Neural Network	**318.10**	**101185.83**	**0.84**	**0.86**	**610.61**

	Jun 2019	**Jul 2019**	**Aug 2019**	**Sept 2019**
Actual	6046	5717	3940	5465
Predicted	5435.39	5813.70	4085.75	5428.94
Difference	610.61	-96.7	-145.75	36.06

Best Parameters:

- Random Forest:
  Criterion: poisson
  Max_depth: 5
  Max_features: sqrt
  Min_samples_leaf: 1

- XGBoost:
  Criterion: squared_error
  Loss: huber
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 1

- Neural Network:
  Activation: identity
  Solver: lbfgs

1.6 Skinbooster Summer

Skinbooster – Summer					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	**33.95**	**1152.39**	**0.50**	**0.53**	**44.26**
XGBoost	36.26	1315.07	0.43	0.46	51.16
Neural Network	58.23	3391.10	-0.48	0.54	82.92

	Jun 2019	**Jul 2019**	**Aug 2019**	**Sept 2019**
Actual	342	245	212	272
Predicted	297.74	288.65	237.39	282.01
Difference	44.26	-43.65	-25.39	-10.01

Best Parameters:

- Random Forest:
  Criterion: friedman_mse
  Max_depth: 5
  Max_features: sqrt
  Min_samples_leaf: 1

- XGBoost:
  Criterion: friedman_mse
  Loss: squared_error
  Max_depth: 10
  Max_features: sqrt
  Min_samples_leaf: 2

- Neural Network:
  Activation: identity
  Solver: adam

1.7 Skincare Summer

Skincare – Summer					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	51.55	2657.21	0.09	0.85	73.15
XGBoost	76.29	5819.75	-1.00	0.86	99.33
Neural Network	**33.99**	**1155.26**	**0.60**	**0.74**	**57.96**

	Jun 2019	**Jul 2019**	**Aug 2019**	**Sept 2019**
Actual	287	386	236	305
Predicted	252.95	328.04	238.27	314.88
Difference	34.05	57.96	-2.27	-9.88

Best Parameters:

- Random Forest:
  Criterion: poisson
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 1

- XGBoost:
  Criterion: friedman_mse
  Loss: quantile
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 2

- Neural Network:
  Activation: identity
  Solver: adam

2. Winter Period

2.1 All Categories Winter

All Categories – Winter					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	1792.80	3214142.17	-11.64	0.46	2136.78
XGBoost	1610.31	2593089.41	-9.20	0.49	1899.85
Neural Network	**457.41**	**209220.28**	**0.18**	**0.46**	**711.11**

	Oct 2019	**Nov 2019**	**Dec 2019**
Actual	8819	8492	7624
Predicted	8107.89	8684.05	7332.28
Difference	711.11	-192.05	291.72

Best Parameters:

- Random Forest:
  Criterion: squared_error
  Max_depth: 10
  Max_features: sqrt
  Min_samples_leaf: 1

- XGBoost:
  Criterion: friedman_mse
  Loss: huber
  Max_depth: 10
  Max_features: sqrt
  Min_samples_leaf: 2

- Neural Network:
  Activation: identity
  Solver: lbfgs

2.2 Dermafiller Winter

Dermafiller – Winter					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	93.22	8690.51	0.23	0.77	142.54
XGBoost	**24.26**	**588.44**	**0.95**	**0.95**	**34.59**
Neural Network	97.11	9430.14	0.16	0.16	135.06

	Oct 2019	**Nov 2019**	**Dec 2019**
Actual	731	809	555
Predicted	722.10	774.41	577.12
Difference	8.9	34.59	-22.12

Best Parameters:

- Random Forest:
  Criterion: squared_error
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 1

- XGBoost:
  Criterion: squared_error
  Loss: squared_error
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 2

- Neural Network:
  Activation: identity
  Solver: lbfgs

2.3 Mesotherapy Winter

Mesotherapy – Winter					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	5.60	31.33	-0.20	0.19	9.65
XGBoost	**4.58**	**20.93**	**0.19**	**0.47**	**7.91**
Neural Network	5.81	33.76	-0.30	0.15	9.90

	Oct 2019	Nov 2019	Dec 2019
Actual	14	5	2
Predicted	6.09	5.32	1.57
Difference	7.91	-0.32	0.43

Best Parameters:

- Random Forest:
  Criterion: poisson
  Max_depth: 5
  Max_features: sqrt
  Min_samples_leaf: 2

- XGBoost:
  Criterion: squared_error
  Loss: huber
  Max_depth: 5
  Max_features: sqrt
  Min_samples_leaf: 1

- Neural Network:
  Activation: logistic
  Solver: adam

2.4 Needles Winter

Needles – Winter					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	30.96	958.77	-0.14	0.23	53.63
XGBoost	32.36	1047.10	-0.25	0.15	56.03
Neural Network	**17.72**	**314.10**	**0.63**	**0.71**	**30.43**

	Oct 2019	**Nov 2019**	**Dec 2019**
Actual	4	65	3
Predicted	7.91	34.57	4.00
Difference	-3.91	30.43	-1

Best Parameters:

- Random Forest:
  Criterion: squared_error
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 1

- XGBoost:
  Criterion: friedman_mse
  Loss: squared_error
  Max_depth: 10
  Max_features: sqrt
  Min_samples_leaf: 2

- Neural Network:
  Activation: identity
  Solver': adam

2.5 Profilho Winter

Profilho – Winter					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	1443.81	2084586.86	-10.25	0.76	1639.44
XGBoost	**1320.98**	**1744994.71**	**-8.42**	**0.57**	**1630.13**
Neural Network	1376.40	1894477.23	-9.23	-0.28	1959.08

	Oct 2019	**Nov 2019**	**Dec 2019**
Actual	7337	6611	6312
Predicted	5706.87	5668.65	5012.15
Difference	1630.13	942.35	1299.85

Best Parameters:

- Random Forest:
  Criterion: squared_error
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 1

- XGBoost:
  Criterion: friedman_mse
  Loss: squared_error
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 1

- Neural Network:
  Activation: identity
  Solver: lbfgs

2.6 Skinbooster Winter

Skinbooster – Winter					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	95.15	9053.49	-4.07	0.85	115.47
XGBoost	105.61	11154.25	-5.25	0.77	128.05
Neural Network	**51.87**	**2690.48**	**-0.51**	**0.85**	**65.52**

	Oct 2019	**Nov 2019**	**Dec 2019**
Actual	369	405	303
Predicted	313.50	339.48	276.56
Difference	55.5	65.52	26.44

Best Parameters:

- Random Forest:
  Criterion: poisson
  Max_depth: 5
  Max_features: sqrt
  Min_samples_leaf: 1

- XGBoost:
  Criterion: friedman_mse
  Loss: squared_error
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 2

- Neural Network:
  Activation: identity
  Solver: adam

2.7 Skincare Winter

Skincare – Winter					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	146.18	21369.34	-1.31	0.08	236.12
XGBoost	**97.87**	**9578.92**	**-0.03**	**0.16**	**160.44**
Neural Network	249.76	62380.59	-5.73	-0.05	363.33

	Oct 2019	**Nov 2019**	**Dec 2019**
Actual	364	597	449
Predicted	415.65	436.56	430.85
Difference	-51.65	160.44	18.15

Best Parameters:

- Random Forest:
  Criterion: poisson
  Max_depth: 5
  Max_features: log2
  Min_samples_leaf: 2

- XGBoost:
  Criterion: friedman_mse
  Loss: quantile
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 2

- Neural Network:
  Activation: identity
  Solver: lbfgs

3. Whole Year

3.1 All Categories Whole Year

All Categories – Whole Year					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	1876.24	3520264.55	-1.84	0.22	3170.99
XGBoost	1458.24	2126469.87	-0.72	0.19	2799.59
Neural Network	**462.02**	**213458.06**	**0.83**	**0.86**	**1101.31**

	Feb 2019	**Mar 2019**	**Apr 2019**	**May 2019**	**Jun 2019**	**Jul 2019**
Actual	5304	6638	7144	6971	7164	6971
Predicted	5555.62	6727.33	7535.84	8072.31	7039.37	7167.34
Difference	-251.62	-89.33	-391.84	-1101.31	124.63	-196.34

	Aug 2019	**Sep 2019**	**Oct 2019**	**Nov 2019**	**Dec 2019**
Actual	4878	6613	8819	8492	7624
Predicted	5388.61	6410.45	8311.02	9012.97	7465.73
Difference	-510.61	202.55	507.98	-520.97	158.27

Best Parameters:

- Random Forest:
  Criterion: friedman_mse
  Max_depth: 10
  Max_features: sqrt
  Min_samples_leaf: 1

- XGBoost:
  Criterion: squared_error
  Loss: squared_error
  Max_depth: 10
  Max_features: sqrt
  Min_samples_leaf: 2

- Neural Network:
  Activation: identity
  Solver: lbfgs

3.2 Dermafiller Whole Year

Dermafiller – Whole Year					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	116.41	13552.42	-0.17	0.48	224.22
XGBoost	**97.84**	**9571.77**	**0.17**	**0.41**	**195.24**
Neural Network	120.33	14480.47	-0.25	0.26	216.33

	Feb 2019	**Mar 2019**	**Apr 2019**	**May 2019**	**Jun 2019**	**Jul 2019**
Actual	646	581	775	560	481	615
Predicted	534.01	609.19	593.25	572.76	494.07	541.93
Difference	111.99	-28.19	181.75	-12.76	-13.07	73.07

	Aug 2019	**Sep 2019**	**Oct 2019**	**Nov 2019**	**Dec 2019**
Actual	472	557	731	809	555
Predicted	527.54	539.90	622.60	613.76	555.01
Difference	-55.54	17.1	108.4	195.24	-0.01

Best Parameters:

- Random Forest:
  Criterion: poisson
  Max_depth: 10
  Max_features: sqrt
  Min_samples_leaf: 2

- XGBoost:
  Criterion: squared_error
  Loss: huber
  Max_depth: 5
  Max_features: sqrt
  Min_samples_leaf: 2

- Neural Network:
  Activation: identity
  Solver: lbfgs

3.3 Mesotherapy Whole Year

Mesotherapy – Whole Year					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	4.48	20.11	-0.21	0.06	10.72
XGBoost	**3.93**	**15.41**	**0.07**	**0.26**	**9.37**
Neural Network	4.70	22.11	-0.33	-0.02	11.63

	Feb 2019	Mar 2019	Apr 2019	May 2019	Jun 2019	Jul 2019
Actual	1	2	3	12	4	3
Predicted	2.32	2.09	3.31	4.28	2.50	2.70
Difference	-1.32	-0.09	-0.31	7.72	1.5	0.3

	Aug 2019	Sep 2019	Oct 2019	Nov 2019	Dec 2019
Actual	6	2	14	5	2
Predicted	3.32	4.65	4.63	3.07	1.70
Difference	2.68	-2.65	9.37	1.93	0.3

Best Parameters:

- Random Forest:
  Criterion: friedman_mse
  Max_depth: 5
  Max_features: sqrt
  Min_samples_leaf: 1

- XGBoost:
  Criterion: squared_error
  Loss: huber
  Max_depth: 5
  Max_features: sqrt
  Min_samples_leaf: 1

- Neural Network:
  Activation: tanh
  Solver: adam

3.4 Needles Whole Year

Needles – Whole Year					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	15.71	246.86	0.20	0.28	51.33
XGBoost	**15.65**	**244.82**	**0.20**	**0.28**	**50.56**
Neural Network	17.50	306.10	0.00	0.12	57.16

	Feb 2019	Mar 2019	Apr 2019	May 2019	Jun 2019	Jul 2019
Actual	7	1	3	2	4	5
Predicted	1.43	1.12	2.12	2.30	3.00	2.03
Difference	5.57	-0.12	0.88	-0.3	1	2.97

	Aug 2019	Sep 2019	Oct 2019	Nov 2019	Dec 2019
Actual	12	12	4	65	3
Predicted	13.39	10.04	13.37	14.44	1.73
Difference	-1.39	1.96	-9.37	50.56	1.27

Best Parameters:

- Random Forest:
  Criterion: friedman_mse
  Max_depth: 5
  Max_features: sqrt
  Min_samples_leaf: 1

- XGBoost:
  Criterion: squared_error
  Loss: squared_error
  Max_depth: 10
  Max_features: sqrt
  Min_samples_leaf: 1

- Neural Network:
  Activation: logistic
  Solver: adam

3.5 Profilho Whole Year

Profhilo – Whole Year					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	1675.11	2805995.72	-2.19	0.27	3001.09
XGBoost	1477.46	2182875.99	-1.48	0.16	2750.87
Neural Network	**1103.03**	**1216675.12**	**-0.38**	**0.29**	**2346.13**

	Feb 2019	Mar 2019	Apr 2019	May 2019	Jun 2019	Jul 2019
Actual	4145	5359	5625	5749	6046	5717
Predicted	5141.51	6576.04	6941.26	8095.13	7141.58	6901.03
Difference	-996.51	-1217.04	-1316.26	-2346.13	-1095.58	-1184.03

	Aug 2019	Sep 2019	Oct 2019	Nov 2019	Dec 2019
Actual	3940	5465	7337	6611	6312
Predicted	4678.68	5682.38	7490.43	6500.00	56347.66
Difference	-738.68	-217.38	-153.43	111	-50035.7

Best Parameters:

- Random Forest:
  Criterion: poisson
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 1

- XGBoost:
  Criterion: squared_error
  Loss: squared_error
  Max_depth: 5
  Max_features: sqrt
  Min_samples_leaf: 2

- Neural Network:
  Activation: identity
  Solver: lbfgs

3.6 Skinbooster Whole Year

Skinbooster – Whole Year					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	204.98	42015.89	-8.91	-0.04	277.52
XGBoost	204.83	41955.24	-8.90	0.09	278.94
Neural Network	**61.91**	**3832.73**	**0.10**	**0.54**	**102.10**

	Feb 2019	Mar 2019	Apr 2019	May 2019	Jun 2019	Jul 2019
Actual	268	400	320	407	342	245
Predicted	249.81	335.97	278.00	304.90	289.60	276.99
Difference	18.19	64.03	42	102.1	52.4	-31.99

	Aug 2019	Sep 2019	Oct 2019	Nov 2019	Dec 2019
Actual	212	272	369	405	303
Predicted	234.65	269.53	290.91	305.06	231.12
Difference	-22.65	2.47	78.09	99.94	71.88

Best Parameters:

- Random Forest:
  Criterion: friedman_mse
  Max_depth: 10
  Max_features: sqrt
  Min_samples_leaf: 1

- XGBoost:
  Criterion: squared_error
  Loss: huber
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 1

- Neural Network:
  Activation: identity
  Solver: adam

3.7 Skincare Whole Year

Skincare – Whole Year					
Algorithm	**RMSE**	**MSE**	**R^2**	**Explained Variance**	**Max Error**
Random Forest	**127.95**	**6370.32**	**-0.46**	**0.43**	**295.96**
XGBoost	136.68	18682.36	-0.66	0.58	276.44
Neural Network	349.29	122000.11	-9.87	0.00	581.08

	Feb 2019	Mar 2019	Apr 2019	May 2019	Jun 2019	Jul 2019
Actual	237	295	418	241	287	386
Predicted	213.63	253.56	257.18	241.87	235.15	261.72
Difference	23.37	41.44	160.82	-0.87	51.85	124.28

	Aug 2019	Sep 2019	Oct 2019	Nov 2019	Dec 2019
Actual	236	305	364	597	449
Predicted	172.91	200.43	291.05	301.04	287.79
Difference	63.09	104.57	72.95	295.96	161.21

Best Parameters:

- Random Forest:
  Criterion: poisson
  Max_depth: 5
  Max_features: sqrt
  Min_samples_leaf: 1

- XGBoost:
  Criterion: friedman_mse
  Loss: huber
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 1

- Neural Network:
  Activation: tanh
  Solver: adam

4. Covid Period

4.1 All Categories Covid

Best Parameters:

- Random Forest:
  Criterion: poisson
  Max_depth: 10
  Max_features: sqrt
  Min_samples_leaf: 1

- XGBoost:
  Criterion: squared_error
  Loss: squared_error
  Max_depth: 10
  Max_features: sqrt
  Min_samples_leaf: 2

- Neural Network:
  Activation: identity
  Solver: lbfgs

4.2 Dermafiller Covid

Best Parameters:

- Random Forest:
  Criterion: poisson
  Max_depth: 5
  Max_features: sqrt
  Min_samples_leaf: 1

- XGBoost:
  Criterion: friedman_mse
  Loss: squared_error
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 2

- Neural Network:
  Activation: identity
  Solver: lbfgs

4.3 Mesotherapy Covid

Best Parameters:

- Random Forest:
  Criterion: poisson
  Max_depth: 10
  Max_features: sqrt
  Min_samples_leaf: 2

- XGBoost:
  Criterion: friedman_mse
  Loss: quantile
  Max_depth: 10
  Max_features: sqrt
  Min_samples_leaf: 2

- Neural Network:
  Activation: tanh
  Solver: sgd

4.4 Needles Covid

Best Parameters:

- Random Forest:
  Criterion: friedman_mse
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 1

- XGBoost:
  Criterion: friedman_mse
  Loss: huber
  Max_depth: 5
  Max_features: log2
  Min_samples_leaf: 1

- Neural Network:
  Activation: tanh
  Solver: adam

4.5 Profilho Covid

Best Parameters:

- Random Forest:
  Criterion: squared_error
  Max_depth: 10
  Max_features: sqrt
  Min_samples_leaf: 2

- XGBoost:
  Criterion: squared_error
  Loss: squared_error
  Max_depth: 5
  Max_features: sqrt
  Min_samples_leaf: 2

- Neural Network:
  Activation: identity
  Solver: lbfgs

4.6 Skinbooster Covid

Best Parameters:

- Random Forest:
  Criterion: squared_error
  Max_depth: 10
  Max_features: sqrt
  Min_samples_leaf: 1

- XGBoost:
  Criterion: friedman_mse
  Loss: quantile
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 1

- Neural Network:
  Activation: relu
  Solver: lbfgs

4.7 Skincare Covid

Best Parameters:

- Random Forest:
  Criterion: poisson
  Max_depth: 10
  Max_features: sqrt
  Min_samples_leaf: 1

- XGBoost:
  Criterion: squared_error
  Loss: squared_error
  Max_depth: 5
  Max_features: sqrt
  Min_samples_leaf: 1

- Neural Network:
  Activation: identity
  Solver: adam

5. Last 3 Months

5.1 Dermafiller Last 3 Months

Best Parameters:

- Random Forest:
  Criterion: friedman_mse
  Max_depth: 5
  Max_features: sqrt
  Min_samples_leaf: 1

- XGBoost:
  Criterion: squared_error
  Loss: squared_error
  Max_depth: 5
  Max_features: log2
  Min_samples_leaf: 2

- Neural Network:
  Activation: identity
  Solver: lbfgs

5.2 Mesotherapy Last 3 Months

Best Parameters:

- Random Forest:
  Criterion: friedman_mse
  Max_depth: 5
  Max_features: sqrt
  Min_samples_leaf: 1

- XGBoost:
  Criterion: friedman_mse
  Loss: squared_error
  Max_depth: 5
  Max_features: sqrt
  Min_samples_leaf: 2

- Neural Network:
  Activation: tanh
  Solver: adam

5.3 Needles Last 3 Months

Best Parameters:

- Random Forest:
  Criterion: squared_error
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 1

- XGBoost:
  Criterion: squared_error
  Loss: squared_error
  Max_depth: 10
  Max_features: sqrt
  Min_samples_leaf: 1

- Neural Network:
  Activation: identity
  Solver: adam

5.4 Profilho Last 3 Months

Best Parameters:

- Random Forest:
  Criterion: poisson
  Max_depth: 5
  Max_features: sqrt
  Min_samples_leaf: 1

- XGBoost:
  Criterion: squared_error
  Loss: squared_error
  Max_depth: 10
  Max_features: sqrt
  Min_samples_leaf: 2

- Neural Network:
  Activation: identity
  Solver: lbfgs

5.5 Skinbooster Last 3 Months

Best Parameters:

- Random Forest:
  Criterion: friedman_mse
  Max_depth: 10
  Max_features: log2
  Min_samples_leaf: 1

- XGBoost:
  Criterion: squared_error
  Loss: squared_error
  Max_depth: 5
  Max_features: sqrt
  Min_samples_leaf: 2

- Neural Network:
  Activation: identity
  Solver: adam

5.6 Skincare Last 3 Months

Best Parameters:

- Random Forest:
  Criterion: squared_error
  Max_depth: 5
  Max_features: sqrt
  Min_samples_leaf: 1

- XGBoost:
  Criterion: squared_error
  Loss: squared_error
  Max_depth: 5
  Max_features: log2
  Min_samples_leaf: 2

- Neural Network:
  Activation: identity
  Solver: adam