Contents

[Chapter 1: Introduction 2](#_Toc135745444)

[1.1 Research Problem 2](#_Toc135745445)

[1.2 Research Question 2](#_Toc135745446)

[1.3 Aims and Objectives 2](#_Toc135745447)

[1.4 Purpose Statement 2](#_Toc135745448)

[Chapter 2: Literature Review 2](#_Toc135745449)

[2.1 Sales Forecasting: 2](#_Toc135745450)

[2.1.1 Meeting customer demand 2](#_Toc135745451)

[2.1.2 Aesthetics products supply and demand 2](#_Toc135745452)

[2.1.2.1 Seasonality 3](#_Toc135745453)

[2.1.2.2 Shelf life 3](#_Toc135745454)

[2.1.3 Extraordinary events affecting supply and demand 3](#_Toc135745455)

[2.1.3.1 Covid-19 4](#_Toc135745456)

[2.1.3.2 Brexit 5](#_Toc135745457)

[2.1.3.3 Change in Government 5](#_Toc135745458)

[2.2 Machine Learning: 5](#_Toc135745459)

[2.2.1 Introduction to Machine Learning 5](#_Toc135745460)

[2.2.2 Machine Learning approaches 6](#_Toc135745461)

[2.3 Forecasting supply and demand: 6](#_Toc135745462)

[2.3.1 Time Series Modelling 7](#_Toc135745463)

[2.3.2 Forecasting Models 8](#_Toc135745464)

[2.3.3 Related Studies 8](#_Toc135745465)

[Chapter 3: Research Methodology 9](#_Toc135745466)

[3.1 Data Acquisition 10](#_Toc135745467)

[3.2 Data description 10](#_Toc135745468)

[3.3 Data Cleaning 10](#_Toc135745469)

[3.4 Data Exploration 12](#_Toc135745470)

[3.5 Data Preparation for Machine Learning 14](#_Toc135745471)

[3.6 Training of Models 14](#_Toc135745472)

[3.7 Testing and Evaluation 14](#_Toc135745473)

[Chapter 4: Analysis of Results and Discussion 15](#_Toc135745474)

[Chapter 5: Conclusion and Recommendations 15](#_Toc135745475)

[Bibliography 15](#_Toc135745476)

# Chapter 1: Introduction

## 1.1 Research Problem

## 1.2 Research Question

## 1.3 Aims and Objectives

## 1.4 Purpose Statement

# Chapter 2: Literature Review

Forecasting is the process of making predictions of the future based on past and present data. This function is used by industries and brands such as cosmeceuticals to research the supply and demand for a specific product. This assists industries in controlling the market chains.

## 2.1 Sales Forecasting:

Forecasting entails making predictions about the expected future sales revenues. 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 upcoming months. Forecasting in companies is used, so that the business can have a general grasp of the market demand and to help formulate a more suitable marketing strategy. 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 a customer is searching for 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 Aesthetics 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, like a drug, 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 a number of tropical 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 be spent for nothing. 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 a number of 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 have an effect on 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 UK, 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 time 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]](#footnote-1) [6].

### 2.1.3.2 Brexit

As the UK 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].

### 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 present 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]. 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.

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.

## 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 [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 also 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 [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 and an accuracy of 68%, while Random Forest obtained an RMSE of 5.13371 and an accuracy of 43%.

As proposed by [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 0.9084, BPNN had a correlation of 0.8628 while the AR model had the least correlation of 0.8076.

In a study conducted by [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.

[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 21501.657 while the Multivariate LSTM obtained an RMSE of 20693.862.

In a study done by [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



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 UK 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. The first variable is a date which shows when the records was inputted into the file called the ‘Data Entry Date’ while the ‘Pharmacy’ column contains 11 unique pharmacy names. The data includes all the sales made between November 2015 and March 2022. The date of the sale is split into 3 different columns, ‘Year’ which refers to the year when the sale was made, ‘Month’ which includes both the Month and the year of the sale, and the ‘Quarter’ column which also includes the year and the quarter in which the sale was made.

The ‘Product’ column includes 26 unique products sold by the company which fall under 6 categories which are not included defined in the data. For each sale of a product, the quantity sold is listed in the ‘Qty’ column. The customer’s name as well as the company’s name for each sale is listed in the different columns called ‘Cust.Name’ and ‘Comp.Name’ respectively. The data includes 4,317 unique customers while having 4,299 unique companies. In the ‘Sales Rep’ column, the name of the sales representative of the company is also listed.

The sales made were spread across 10 different locations including different regions in the UK, Ireland and other countries such as the USA. The different countries and regions were represented by a unique number in the ‘Location’ column. To specify the locations further, 2 other columns are listed as ‘Code’ and ‘Area Code’ which correspond to the real are codes found in the UK and Ireland. For countries such as the USA, the area codes were listed as ‘blank’ since they are not part of the UK or Ireland.

By using this dataset, a prediction of the next following months was made and evaluated.

## 3.3 Data Cleaning

To be able to make a time-series forecasting of sales, the data needs to be in a numerical form. In order 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 which taken in the data cleaning process was loading the data from the csv file into a Pandas 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 ’Sales Rep’ column was deleted as it was not going to be used either.

The next step was creating a single date column on which the forecasting can be made. The month value from the ‘Month’ column was extracted and placed in a new column called ‘New-Month’. Each month value in the new column was then changed from text to numeric while the old ‘Month’ column was deleted, and the column containing the month as a numeric value was named ‘Month’. The same process was then repeated on the ‘Quarter’ column to extract the quarter as a number.

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 keep the data anonymous, each original product name was changed to an identifiable name such as the type of product and a number added at the end. This was done using the replace function found in Pandas. 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 this needed to be added in the data frame. In order to do this task, the ‘Product’ column needed to be duplicated, and then a dictionary had to be used. The dictionary was created by writing the names of the products in the first value and declaring the category in the second value. The dictionary was then used to find the first values in the column and replace them with the category names.

Since the data contained thousands of unique customer names, a dictionary could not be used straightaway to make the names anonymous. 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 columns which were not considered that important when predicting future sales were the ‘Data Entry Date’, ‘Cust.Name’, ‘Comp.Name’, ’Area Code’ and ‘Code. These columns were removed so that the forecasting process would be made only on the relevant data.

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

To save some time during the research, the data was first explored using PowerBI visuals, and then the most important visuals were plotted using Python. The data was explored by plotting it against various plot such as Histograms, Boxplots, Pie Charts, and others.

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.

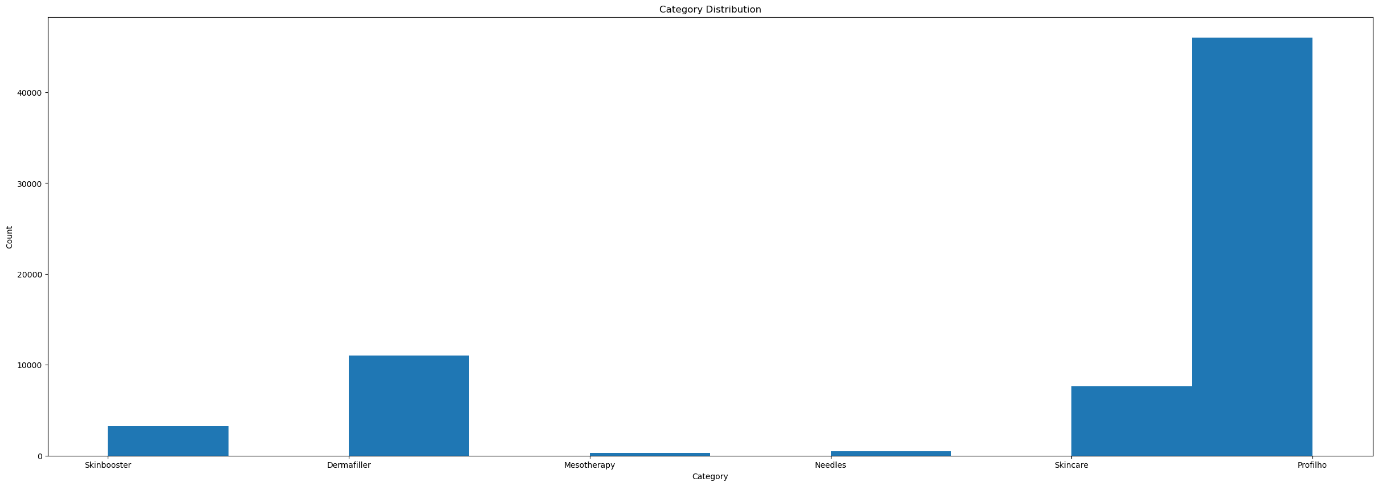


Figure 2: Category Distribution

A picture containing screenshot, diagram, plot, line

Description automatically generated

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 are show in in Figure 4: Category Boxplot and Figure 5: Product Boxplot respectively.

A picture containing line, diagram, plot, parallel

Description automatically generated

Figure 4: Category Boxplot

A picture containing line, plot, screenshot, text

Description automatically generated

Figure 5: Product Boxplot

‘Injection 7’ made up 85% of the products sold as shown in the pie chart in Figure 6: Product Pie Chart. As shown in Figure 7: Category Pie Chart, ‘Profilho’ made up 85% of the categories sold while ‘Dermafiller’ made up 8%. As seen in Figure 8: Location Pie Chart, 20% of the sales were made in ‘London’ followed by 16% made in ‘NorthEast, Midlands’.

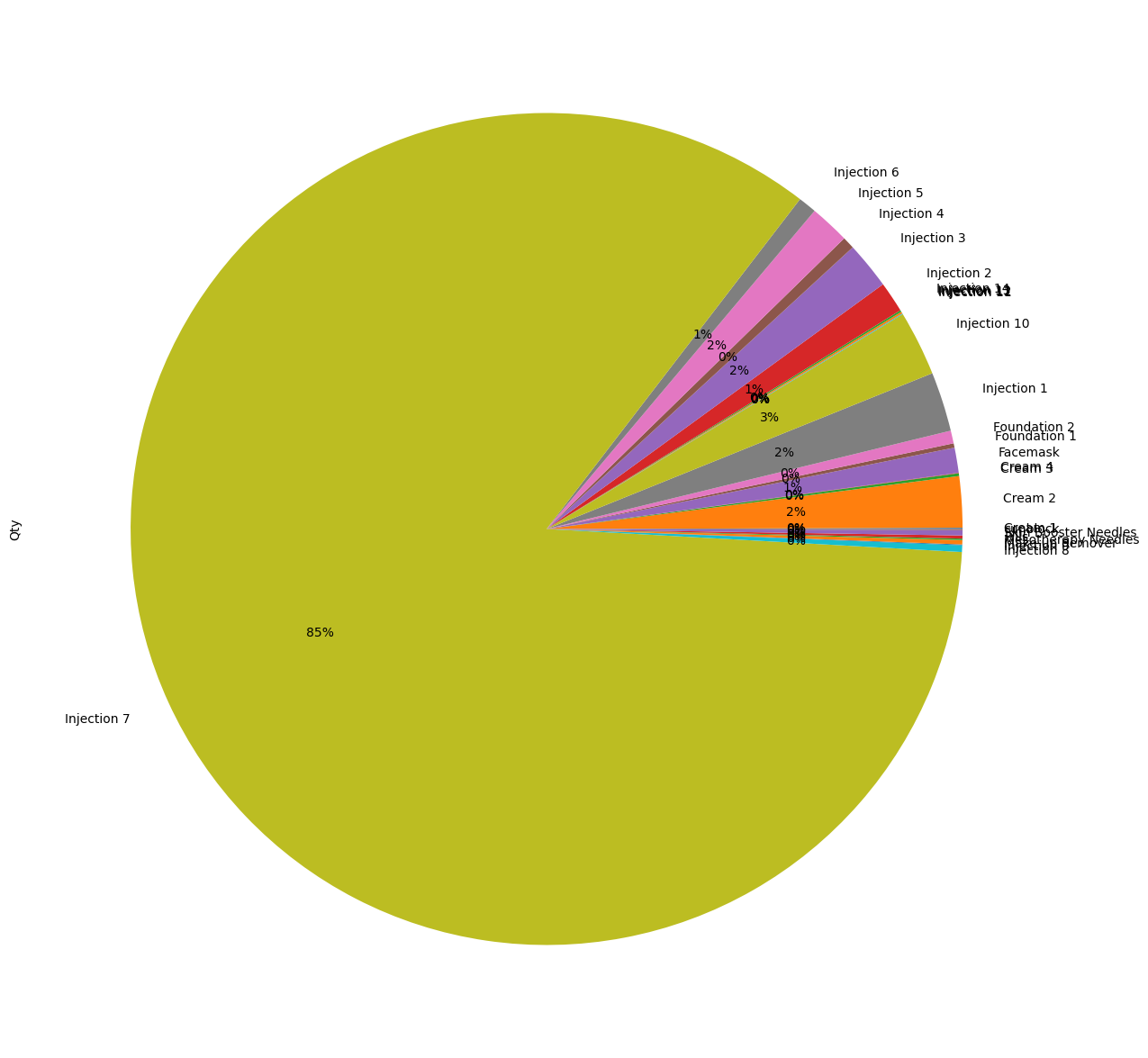


Figure 6: Product Pie Chart

A picture containing circle, diagram, screenshot, graphics

Description automatically generated

Figure 7: Category Pie Chart

A picture containing diagram, colorfulness, circle, screenshot

Description automatically generated

Figure 8: Location Pie Chart

The trend of the sales data was plotted against a time graph as seen in Figure 9: 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.

A picture containing text, plot, screenshot, line

Description automatically generated

Figure 9: Time Graph

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 10: Seasonality Graph, we can conclude that November seemed to have the highest number of sales as opposed to the month of April.

A picture containing diagram, line, plot, text

Description automatically generated

Figure 10: Seasonality Graph

In Figure 11: 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.

A picture containing text, line, screenshot, font

Description automatically generated

Figure 11: Different Graphs

## 3.5 Data Preparation for Machine Learning

Before doing the final forecasting, various tutorials using different algorithms were attempted. From these tutorials, predictions could already be made, and the best performing algorithms were noted down and compared to the related studies. 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 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 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, six different types of splitting the data were used. The first experiment was conducted by splitting the data by percentages, having a value if 80% as training data and a value of 20% as testing data. This splitting type was made up of seven experiments for each algorithm. The testing data was done on a total of 14 months of the data but since in this experiment the ‘Date’ was not converted into the ‘datetime’ datatype, the dates that were predicted varied between different months and years.

The second type of splitting that was used in this study was the random splitting by using the train\_test\_split training set. This splitting type was made up of only one experiment for each algorithm using all the categories at once. This type of splitting also predicted a total of 14 months of the data, but these testing dates were selected at random by the algorithm.

Another type of splitting was conducted on a summer period from the data. The dates were selected between June 2019 and September 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.

A different 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 hit.

To have an overview of sales made during each month of the year 2019, the fifth 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 the 12 selected months of the data one by one based on the past data from the previous years.

The last splitting type that was used was conducted on one whole year during the Covid-19 period which predicted 12 months of data. The testing dates selected were between April 2020 and April 2021, during which two enforced lockdowns happened.

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 108 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. From the results obtained from each experiment, an evaluation and conclusion were formulated.

# Chapter 4: Analysis of Results and Discussion

# Chapter 5: Conclusion and Recommendations

# Bibliography

|  |  |
| --- | --- |
| [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. |

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