A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Alexandra Munik Silva |
| *Student Number* | 2024268 |
| *Module Title* | Strategic Thinking |
| *Assessment Title* | CA2 Strategic Thinking |
| *Assessment Due Date* | Sunday, 15 December 2024 |
| *Date of Submission* | Sunday, 5 December 2024 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Contents

[Strategic overview of the business problem 1](#_Toc184327014)

[Project plan 1](#_Toc184327015)

[Business understanding 2](#_Toc184327016)

[Data understanding 2](#_Toc184327017)

[Data preparation 2](#_Toc184327018)

[Data Cleaning 3](#_Toc184327019)

[Outliers 3](#_Toc184327020)

[Statistic 4](#_Toc184327021)

[Insights 5](#_Toc184327022)

[Machine Learning 6](#_Toc184327023)

[Feature Importances 6](#_Toc184327024)

[Predictive Inference 6](#_Toc184327025)

[Insigth 7](#_Toc184327026)

[Heatmaps 7](#_Toc184327027)

[Insights 8](#_Toc184327029)

[Conclusion 9](#_Toc184327030)

[References 9](#_Toc184327031)

# Introduction

Data and relevant analytics was supporting the business strategies in the increasing need to make better decisions and to operate more effectively and profitably in this very high-competitive and fast-developing market.

Surely, the analysis of data has turned into a core asset for the maintenance of the competitive edge of any organization through right choices and the adaption of strategies to changing market requirements.

According to a well-known data scientist, DJ Patil, "Data is the new oil."

This strong analogy speaks to the valuable power data has in transforming sectors and redesigning the business landscape.

In the e-commerce sector, where every day a large volume of customer data is generated, acumen to convert the data into actionable insights has taken the lead and has become even more important than ever.

Companies can use the more advanced tools like machine learning, artificial intelligence, and statistical models for studying and analyzing consumer behavior as well as future trends; and further enhancement and optimization of varied business determinants in the e-commerce sector, including inventory management, pricing, and promotions.

The application of machine learning and predictive analytics will shift e-commerce companies from working in a mode that is driven by instincts and hunches to a model driven by accurate data, dynamic at scale, and applicable to the market.

This project focuses on the use of data analysis and machine learning to enhance sales forecast and determine the leading factors of consumer behavior for an online store.

This will in turn enable the firm to apply several models like ARIMA, Random Forest, Gradient Boosting, and Decision Trees in forecasting its future sales as well as the prima facie factors for any associated business plans.

Having extended these cutting-edge methodologies to the online store data, the main takeaway to be disclosed is the ensuing impact in the fine-tuning of sales strategies and the attestation it provides in molding a path through which a business maximizes its profit by making data-driven choices.

As stated by Thomas H. Davenport, a well-known Professor in Management and Analytics, "Competing on analytics has become a core competency for successful organizations."

Therefore, this project is intended to prove this statement by showing how the business in the digital era can better keep itself competitive, make better-informed decisions, and in the end achieve better financial outcomes.

We shall look deep into these models, comparing several machine learning algorithms to see which best one for e-commerce.

It will help not just to get the right price and manage inventory well but also to know how the promotion should be ran, giving actionable insights to drive business growth.

This project has the ultimate underpin objective of emphasis-if emphasis be on stressing the fact-of the role that data analysis and machine learning can play a tool as a tool of revolutionizing revolutionize, no less the strategy of business, its adeptness in operations and return on investment.

# Strategic overview of the business problem

Some companies do not use data analysis as a competitive advantage.

Failing to use data analytics can result in missed opportunities for growth, optimization, and innovation; that is the core problem in the modern competitive landscape.

Therefore, in this paper’s second part, we take the problems as to how a company can leverage data analysis to optimally adopt a business strategy and improve decision-making processes, making them agile and ensuring more accurate decisions.

That data-driven strategies will move businesses from decisions based on mere intuitions toward embracing the available prescriptive and predictive practices based on concrete data insights.

As Davenport and Harris (2017) have said, “Perhaps your operational business processes aren't much different from anybody else's, but you feel you compete on making the best decisions. Maybe you can choose the best locations for your stores-if so, you're probably doing it analytically. You may build scale through mergers and acquisitions and select only the best candidates for such combinations.

Most don't work out well, according to widely publicized research, but yours do if so you're probably not making those decisions primarily on intuition.

Good decisions usually have systematically assembled data and analysis behind them.”

This project applies the ARIMA, Random Forest, Gradient Boosting, and Decision Trees machine learning models to the forecasting of future sales and the discerning of effects from different business factors on sales so that optimization strategies can be constructed.

These were selected modeling approaches on the basis that they are well-equipped to manage complex, nonlinear relationships across data when it comes to making predictions about sales and key business drivers.

The machine learning approach is considered to add to the traditional statistical method because it has better accuracy and has the ability to model complex interactions.

Advanced predictive modeling is contended to be the means through which data-driven decision-making is shown to be so critical to gaining a competitive edge in the e-commerce market.

As has been put forward by Brynjolfsson and McAfee (2014), “The use of big data and predictive analytics is not just a competitive advantage but has become a critical factor in superior performance.”

This implies that those companies that are data-driven not only remain competitive but run themselves for long-term success by running their operational efficiency and decision-making optimization.

The different machine learning models which were applied in this project were used to show how data analysis can transform businesses by helping them take much more informed, data-driven decisions and acquire a clear competitive edge over those who do not apply such an approach.

# General Goal

The main goal of the project is to forecast future sales for an e-commerce business using only the historical sales data with several machine learning models that can be implemented not just to forecast sales for the business but also to forecast other important aspects that can help the business to optimize pricing strategies, inventory management, and promotional planning.

This project will use the ARIMA, Random Forest, Gradient Boosting, and Decision Trees to give a guide that will help product reviews, pricing, and seasonality on the behavior of consumers in their purchase decisions.

“The wide adoption of customer relationship management, or CRM, and supply chain management software has allowed enterprises to fully interface and integrate their demand and supply chains. Based on this integration, they are better able to capture up-tothe-minute data about demand for a particular product, as well as data of similar granularity about the supply of corresponding data. Analyzing these two data streams, organizations optimize the price of a particular product along several dimensions so demand meets available supply; for example, the price of a product may be different through one channel (such as the Web) than through another (such as a retail store). Price optimization allows any type of organization to maximize profit margins for each item sold while reducing inventory.” ((Kohavi, Rothleder and Simoudis, 2002)

# 

# Business understanding

An E-commerce that deals with the sales of diverse products was chosen for this study, and it supplied a dataset pertaining to the sales recorded of these items for 12 months.

Along with sales volume, the file includes critical variables like price and review scores and review counts for the products for the year in question; it will help us look into an array of factors that could influence the performance of sales such as the impact of customer reviews and price seasonality.

The main focus of the following analysis will be on the best-selling products throughout the 12 months and whether their reviews and prices sparked any influence on the sales performance.

Other factors that should be taken into consideration would be an examination of how much seasonality or the specific months had impacted the sales- this pattern is important for any e-commerce company that needs to have a grip on the inventory management, on the marketing efforts, and on the pricing strategy for the peak and off-peak periods.

The areas that the sales data analysis would be able to shed some light on are very sensitive and important like:

Future Sales: Using machine learning models to make a precise forecast of what the future sales for the next coming months will look like based on the historical data, thus providing important decisions to the company pertaining to inventory, promotions, and pricing strategies.

Best-Selling Products: Knowing the products that make the highest sales, the firm will focus on best-selling products and enhance its marketing and promotional activities on these specific products.

Influence of Reviews and Price: Considering a close look at how reviews (incorporating the number and score) and price affects product sales, this is what will help the company adjust its pricing and review strategies to enhance sales performance.

Monthly Sales Trends: Since there is fluctuation in the volume of sales in different months, it is important for the firm to begin forecasting seasonal trends and thereafter adjusting their inventory and marketing efforts to meet the new trend.

This is also important for them in planning when to have high-demand-period promotions.

A deeper relationship between price, reviews, seasonality, and sales, and their effect on the company's business model, will be the subject of this project.

Actional insights that could change and implement an improved version of the business model for the company will be aimed to develop through this project.

This, in turn, will lead to higher sales, better customer satisfaction, and an overall enhancement of the effectiveness of the business strategies on the e-commerce market.

# Data understanding

Basically, the data frame describes the ID of the products sold, product name, category, price, product reviews and units sold per month.

This data frame is a file with 1000 rows and 18 columns as described below:

* Product id: numbers that identify products
* Product name: name the products (numbers)
* Category: This case we have 7 categories which is Books, toys, Home & Kitchen, clothing, Health, Sport, electronics and clothing.
* Price: Price Review score: Review provided from customers
* Review count: number of reviews
* Sales month X: Sales month 1 until Month 12

# Data preparation and Data Cleaning

A key part of the data analysis pipeline is data preparation. The libraries used were: Pandas, seaborn, numpy, sklearn.decomposition, matplotlib.pyplot, seaborn. To understand the data frame, the following formulas were used with their following interpretations:

* pd.read\_csv: Used to read the data frame.
* df.head: Used to show the first 5 rows of the dataframe for better visualization.
* print(df.columns): Used to print the column names.
* df.shape: Used to understand the size of the Data frame in this case consisting of 1000 rows and 18 columns.
* df.dtypes: Shows the data types contained in the Dataframe, file size, number of rows and columns.

After analyzing the data frame, the cleaning process began with the following functions:

* df.drop: Function used to discard some columns, in this case the “product\_name” and “product\_id” columns were discarded because according to the proposed analysis, they would not make a big difference in the data frame as they only deal with product names and ids.
* df.isnull: Function for identifying missing values, in this case we have the result of “false” indicating that there are no missing values ​​in the data frame.

For the new analyses, columns such as total\_sales were included, with the intention of understanding the general sales of the year and the dataset was adapted according to the requirements of each Machine Learning model.

**Feature Engineering**:

* + **Creating New Columns**: We added new features like **“total\_sales”** to capture the total sales for each product across all months. This was important for understanding the overall performance of products across the year and providing a more holistic view of sales trends.
  + **Column Adaptation for Machine Learning Models**: To adapt the data for machine learning, columns were further manipulated to fit the requirements of the different models. For example, categorical variables like **product category** were encoded numerically to be used in algorithms like **Random Forest** and **Gradient Boosting**.

After these steps, the dataset was ready for exploration using **data visualization** techniques and for training the **machine learning models**.

After cleaning and preparing the data, including necessary columns such as total\_sales and necessary adaptation of the data to process machine learning, we can proceed with analysis and predictions through Machine learning models.

# EDA and Statistical Analysis

“The visual nature of heatmaps allows for immediate recognition of patterns, such as clusters, trends, and anomalies. This makes heatmaps an effective tool for exploratory data analysis” (Bothma, 2024)

# We can notice that most of the correlations are weak, close to 0, we can also verify in this case that the review variable has a positive correlation in relation to the number of sales. In other words, the higher the review score, the higher the units sold. Which proves our initial idea that reviews impact sales, in this case we can gain some insight into this.

# Insights

* This company in question could invest in good marketing or simply adopt post-sales research as a strategy to increase the number of reviews for this e-commerce.
* Another idea would be to understand reviews through other analyses to understand and guarantee customer satisfaction since through the graph we can see that the higher the Review score, the greater the number of sales.

# Statistical analysis

# Machine Learning

For machine learning, some techniques were applied to understand and make predictions, tools such as Feature Importances, Predictive Inference and Heatmaps were used to make predictions and understand better the data.

# Feature Importances

The model.feature\_importances\_ formula returns the relative importance of each feature in a tree-based machine learning model, such as Random Forest or Gradient Boosting. It helps identify which variables have the greatest influence on the model’s predictions.

Feature importance refers to a class of techniques for assigning scores to input features to a predictive model that indicates the relative importance of each feature when making a prediction. (Brownlee, 2020)

Texto

Descrição gerada automaticamente

In this case we had greater relevance in the features on the number of products sold per month, in which we can see that they present results above 0.1 indicating the importance of these values in the dataframe.

# Predictive Inference

Predictive Inference is the process of using a statistical or machine learning model to predict future or unknown values based on available data. The focus is on accurately estimating responses to new cases, regardless of understanding the underlying mechanisms. (Sanders, 2019)

After that, another tool was applied to predict what the introduction of a new product would be like in the company. We then had the following result through the Predictive Inference technique applied below:

Texto

Descrição gerada automaticamente

We can predict that a new product would have an acceptable performance in the first 6 months, as it would represent 37% of sales compared to the current sales of the best-selling category (books).

# Insigth

* The introduction of a new product would be a good strategy for the company; however, it would be interesting in this case to consider the best-selling products and relate this new product to the others so that the company has more possibilities to apply strategies based on the best-selling products (upselling).

# Conclusion

After the analyzes carried out above, we can conclude that data analysis is a very important and relevant tool for a company, without this tool, a company would be left to make assumptions, considering poor information without much relevance for the company.

Many companies underestimate the power of data analysis, often leading the company to bankruptcy due to lack of information and often due to applying bias in business strategies.

Through this data frame we can extract several insights such as:

* Improvements in marketing campaigns more focused on specific products. Pay attention to customers in some specific categories with the intention of understanding the products they are looking for and their level of satisfaction with the product.
* Pay attention to competitors, understand prices, product variety, etc.
* Apply Upselling techniques with the intention of increasing sales of some products through products that have strong selling power in the company.
* We were able to predict how a new product would perform in future years.
* We were also able to verify the impact between variables, that is, the higher the Review Score of a given category, the greater the number of products sold.

In this case, we have an interesting observation about this analysis, if we were to apply Bias in this analysis we would automatically go to the price factor, but in this case the price factor was not as relevant according to the analysis. Different from the Review factor, which can give us more relevant information and consequently more accurate insights, information about sales, we can have access to the behavior of products sold, customers, employees and the company within a general scenario.

Thus proving the problem initially addressed, about companies that do not invest in data analysis, this project aimed to show how easy it is to make decisions based on data and facts as shown above, with the intention of minimizing expenses, opening up possibilities for new strategies, new products and new perspectives in relation to the market and customers.

# References

Bothma, J. (2024). *Seaborn Heatmaps: A Guide to Data Visualization*. [online] Datacamp.com. Available at: https://www.datacamp.com/tutorial/seaborn-heatmaps.

Brownlee, J. (2020). *How to Calculate Feature Importance With Python*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/calculate-feature-importance-with-python/.

Davenport, T.H. and Harris, J.G. (2017). *Competing on analytics : the new science of winning*. Boston, Massachusetts: Harvard Business Review Press.

GeeksforGeeks (2021). *Detect and Remove the Outliers using Python*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/detect-and-remove-the-outliers-using-python/.

GeeksforGeeks. (2018). *Exploratory Data Analysis in Python*. [online] Available at: https://www.geeksforgeeks.org/exploratory-data-analysis-in-python/.

Kohavi, R., Rothleder, N.J. and Simoudis, E. (2002). Emerging trends in business analytics. *Communications of the ACM*, 45(8). doi:https://doi.org/10.1145/545151.545177.

Sanders, N. (2019). A Balanced Perspective on Prediction and Inference for Data Science in Industry. *Issue 1*. doi:https://doi.org/10.1162/99608f92.644ef4a4.

Versloot, C. (2022). *Machine learning articles*. [online] GitHub. Available at: https://github.com/christianversloot/machine-learning-articles/blob/main/how-to-predict-new-samples-with-your-keras-model.md.