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

Para inciar as analises do Dataframe primeiramente começamos entender quais são as relevâncias dos dados do Dataframe, iniciamos primeiramente visualizando as variáveis e entendendo a relação das variantes através da correlation matrix . Seguem gráficos e suas respectivas analises:

Primeiramente para entender como estão distribuídos os dados, seguimos com uma breve analise:

Gráfico, Gráfico de linhas

O conteúdo gerado por IA pode estar incorreto.

Podemos ver de acorco com o gráfico que as vendas possuem grande variação durante o ano como por exemplo entre os meses 4 e 5 há uma queda considerável, enquanto que nos meses 9 e 10 temos um aumento considerável das vendas. Variaveis como datas comemorativas, estações do ano, podem impactar nas vendas, vamos entender mais a frente através de modelos de machine learning.

Após identificação dos top selers podemos explorar os dados de forma mais acertiva. Texto

O conteúdo gerado por IA pode estar incorreto.

Gráfico, Gráfico de barras

O conteúdo gerado por IA pode estar incorreto.

No gráfico abaixo podemos identificar qual foi o comportamento dos produtos mais vendidos divididos em meses, através desses dados já podemos incluir alguns insigths valiosos sobre as vendas futuras desse ecommerce.

Para entender melhor o comportamento do negócio podemos também visualizar mais produtos como plotados abaixo:

Gráfico, Gráfico de barras

O conteúdo gerado por IA pode estar incorreto.

Algumas variáveis importantes também foram consideradas a critério de analise e entendimento dos dados apresentados. Abaixo podemos ver a distribuição dos Review no geral. Podemos perceber que os reviews de forma geral não nos da uma visão clara sobre o quanto essa variável influencia no comportamento de vendas por exemplo.

Gráfico, Histograma

O conteúdo gerado por IA pode estar incorreto.

“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)

Gráfico, Gráfico de mapa de árvore

O conteúdo gerado por IA pode estar incorreto.

Como citado acima, podemos ter alguns insights sobre o gráfico:

The strong positive correlations with Total Sales are in sales\_month\_1, sales\_month\_10, sales\_month\_6, and sales\_month\_5, with coefficients from 0.31 to 0.36.

That is to say, these months have an impact on the general sales performance for the product categories.

For instance, sales\_month\_1 is at 0.36, which means there is a stronghold of the correlation between this month and general sales for the year.

The positive correlation in these months may be due to seasonal factors and/or promotions that drive customer demand, especially times of the year.

The moderate correlations imply that sales are somewhat related to the months patterned as Month 1, Month 10, and Month 6.

The stronger correlations for these months the periods when certain months are in demand or promotions work would be more important for businesses in targeting additional marketing strategies.

A better business plan, particularly in inventory management as well as for promotions and pricing, would be to understand and capitalize on these patterns.

However, the low or very weak correlation with other factors such as price, review score, and review count leads to the suggestion that these are not strong drivers of the total sales that would be expected and indicates that there are other underlying factors.

Organizando ideias:

Primeiro apresentar gráficos de exploração de dados, dar alguns insigths sobre e explicar sobre base dos top sales.

Começar a explicar machine learning (modelos) cada um deles depois qual o meljor modelo e por que (rascunho em python)

Conclusão

Seguir com banner e vídeo.

# Machine Learning

**Machine Learning Models Used**

In this project, three machine learning models were used to forecast sales and analyze key factors influencing sales performance: **Random Forest**, **Gradient Boosting**, and **ARIMA**. Each of these models brings its own strengths to the table, making them suitable for handling different aspects of the data and addressing various business needs. Understanding the workings of these models and how they can be applied in real-world business contexts is critical for deriving actionable insights from the data.

**Random Forest** is a popular ensemble learning method that builds multiple decision trees and combines their results to improve the prediction accuracy. This model works by aggregating the predictions of individual trees, which helps in reducing the variance and overfitting typically associated with decision trees. As noted by **Liaw and Wiener (2002)**, Random Forest is particularly effective in handling large datasets with numerous features and is highly adaptable to various types of data. In the context of sales prediction, Random Forest can effectively model complex, nonlinear relationships between features like price, reviews, and sales performance. Its ability to handle a wide range of predictor variables makes it valuable for optimizing inventory, pricing, and promotional strategies in e-commerce businesses.

**Gradient Boosting** is another ensemble learning technique that focuses on building models sequentially, where each new model corrects the errors of the previous one. According to **Friedman (2001)**, Gradient Boosting improves model performance by focusing on difficult-to-predict cases, making it especially powerful for capturing intricate patterns in data. In the context of predicting sales, Gradient Boosting can capture more complex relationships and interactions between the features, such as the combined effect of price and review score on sales. This makes it an excellent choice for businesses looking to improve forecast accuracy by incorporating multiple factors that influence sales. By leveraging the power of sequentially built trees, Gradient Boosting helps businesses make more precise predictions, thus enhancing decision-making in areas like stock management and promotional efforts.

**ARIMA** (AutoRegressive Integrated Moving Average) is a time series forecasting model that relies on the historical values of a variable to make predictions. This model is particularly effective for datasets with clear trends and seasonal components, which is common in sales data. **Hyndman and Athanasopoulos (2018)** describe ARIMA as a powerful tool for time series analysis because it accounts for both autocorrelation and trend in the data. For businesses, ARIMA is invaluable for predicting future sales based on past performance, allowing companies to forecast demand and optimize inventory, pricing, and staffing accordingly. In e-commerce, where sales can be seasonal, ARIMA helps in adjusting strategies for peak and off-peak times, ensuring that businesses meet demand efficiently without overstocking or understocking products.

These three models—**Random Forest**, **Gradient Boosting**, and **ARIMA**—are widely recognized for their ability to handle complex datasets and provide businesses with actionable insights. By applying these models, businesses can make more informed, data-driven decisions, ultimately enhancing their ability to forecast sales, manage resources, and plan for the future. These models are critical tools for **e-commerce companies** aiming to optimize their operations and stay competitive in an increasingly data-driven market.

# Decision Tree model

The Decision Tree model was chosen because it is a transparent, interpretable model that allows us to visually inspect how decisions are made. This model works well with both categorical and continuous features, making it ideal for our dataset, which contains variables like product price (continuous) and product category (categorical). Additionally, the decision tree can capture non-linear relationships between features, which is important for understanding complex decision boundaries in sales prediction.

Data Preprocessing and Model Training: Before applying the Decision Tree, we performed essential preprocessing steps:

Class Creation: We transformed the total\_sales variable into a binary classification (sales\_class), with 1 representing high sales and 0 representing low sales. This transformation allows us to frame the problem as a classification task.

Feature Selection: We selected several key features that we hypothesized would have an impact on sales performance, including price, review\_score, review\_count, and sales data from the past 12 months.

Hyperparameter Tuning: We optimized the model’s hyperparameters, such as max\_depth, min\_samples\_split, and min\_samples\_leaf, using GridSearchCV to improve the model's performance and avoid overfitting. These hyperparameters help control the complexity of the tree, preventing it from becoming too deep and overfitting the training data.

Key Findings from the Decision Tree:

The Decision Tree splits based on the most important features, such as price and review\_score. For example, products with higher prices or higher review scores tend to fall into the High Sales category, as these features are good indicators of customer satisfaction and demand.

The depth of the tree was set to 15, which allowed it to capture more detailed patterns. While this improved performance, it also made the tree more complex, which could lead to overfitting if not carefully monitored.

The class weights were balanced to address the potential imbalance in sales distribution, ensuring that the model doesn’t bias towards the majority class (Low Sales).

Evaluation of the Model: The performance of the model was evaluated using several key metrics, including:

Accuracy: The Decision Tree achieved an accuracy of 74% on the test set, indicating a decent performance in predicting both low and high sales products.

Precision and Recall: We observed that the precision and recall for Low Sales were higher than for High Sales, which suggests that the model was better at identifying products with low sales.

F1-Score: The F1-Score for both classes was relatively balanced, with a slightly higher score for Low Sales. This suggests that while the model is good at predicting low sales, there is still room for improvement in predicting high sales.

Conclusion:

In summary, the Decision Tree model has provided valuable insights into which features drive sales classification. While the model performed reasonably well with an accuracy of 74%, further optimization and testing of other models like Random Forest or Gradient Boosting could improve the prediction of high sales products. This work highlights the importance of feature selection, model tuning, and evaluation in building an effective predictive model for e-commerce sales.

# **Random Forest Regressor**

The initial model provided reasonably accurate predictions, with a moderate R² value of 0.45, indicating that 45% of the variability in sales was explained by the model. The forecast for the next year showed that the **Books** category is expected to have the highest sales in **Month 7** (551 sales), followed by **Month 2** (534 sales) and **Month 10** (531 sales). For **Toys**, the highest forecasted sales occur in **Month 2** (606 sales), while **Month 1** and **Month 12** are predicted to have lower sales at 364 each. The **Sports** category forecasts the highest sales in **Month 2** (477 sales), followed by **Month 10** (520 sales) and **Month 9** (513 sales). The **Test MAE** for the models are as follows: **Books** (277.68), **Toys** (245.11), and **Sports** (288.34).

After applying **GridSearchCV** for hyperparameter optimization, the best parameters were found:

* **max\_depth**: 10
* **max\_features**: 'sqrt'
* **min\_samples\_leaf**: 2
* **min\_samples\_split**: 5
* **n\_estimators**: 200

However, after hyperparameter tuning, the model's performance degraded with an **R²** value of **-0.17**, indicating that the optimization may have caused overfitting or underfitting. This suggests that the selected hyperparameters might not be optimal for this dataset.

Insights from the forecast suggest that **promotions** have a significant impact on sales, especially in months like **Month 7** for the **Books** and **Sports** categories, which could indicate seasonal promotions or special events driving sales. For **Toys**, the forecast shows lower sales in **Month 1** and **Month 12**, likely due to **seasonality** (e.g., post-holiday sales slumps). This reinforces the importance of adjusting sales forecasts according to seasonal trends and external events.

While the initial model performed well, the negative R² after hyperparameter tuning highlights the need for further optimization. Additional hyperparameter tuning, exploring alternative models like **Gradient Boosting** or **XGBoost**, and incorporating advanced feature engineering—such as **economic factors** or **marketing spend**—could lead to better predictive results. The key takeaway from this analysis is that **promotions** and **seasonality** are crucial factors influencing sales, and these should be carefully integrated into future sales forecasts to improve accuracy.

In conclusion, this project utilized **Random Forest Regressor** to predict sales for the top three categories, forecasting future sales and providing valuable insights into factors driving sales. Despite the initial success, further model refinement, additional features, and alternative algorithms will be necessary to improve forecasting accuracy and generate more robust business insights.

# Predictive Inference

# Predictive Inference

to the current sales of the best-selling category (books).

# Insigth

# 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.