Ícone

O conteúdo gerado por IA pode estar incorreto.

*ALEXANDRA MUNIK SILVA 18/05/2025 Strategic Thinking*

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* | CA3 Strategic Thinking |
| *Assessment Due Date* | Sunday, 18 May 2025 |
| *Date of Submission* | Sunday, 18 May 2025 |

**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**

[Introduction 1](#_Toc198399983)

[Strategic overview of the business problem (metodologia) 2](#_Toc198399984)

[General Goal 3](#_Toc198399985)

[Business understanding 3](#_Toc198399986)

[Data understanding 4](#_Toc198399987)

[Data preparation and Data Cleaning 4](#_Toc198399988)

[EDA and Statistical Analysis 5](#_Toc198399989)

[Correlation Matrix 8](#_Toc198399990)

[Machine Learning 9](#_Toc198399991)

[Decision Tree model 9](#_Toc198399992)

[Random Forest Regressor 10](#_Toc198399993)

[Arima 11](#_Toc198399994)

[Gradient Boosting 12](#_Toc198399995)

[Best Model 13](#_Toc198399996)

[Strategic Recommendations 14](#_Toc198399997)

[Conclusion 15](#_Toc198399998)

[References 17](#_Toc198399999)

# Introduction

Data and relevant analytics were 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 analysing consumer behaviour 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 move 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 behaviour 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 run, 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.”

Methodology

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 modelling 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 modelling 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

This project has but one simple aim: to predict future sales for an e-commerce business using past sales data - via the application of multiple machine learning models.

Models that don’t just predict future sales volumes but also provide actionable insights in the optimization of key business areas including pricing strategy, inventory management, and promotional planning.

To achieve this, the project will implement the ARIMA, Random Forest, Gradient Boosting, and Decision Tree models.

These techniques will also be used to analyse the impact of product reviews, price variations, and seasonal trends on consumer behaviour in the aggregation of sharper, clearer, and more incisive integrated insights for enterprise-wide decision support.

“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-to the-minute data about demand for a particular product, as well as data of similar granularity about the supply of corresponding data. Analysing 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

# 

For this analysis project, an E-commerce was selected that sells different types of products. The 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 analysing the data frame, the cleaning process began with the following functions:

The df.drop function was used to drop the “product\_name” and “product\_id” columns. The columns were dropped due to the lack of relevance in the overall information and to optimize the analysis.

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

* Monthly Sales Evolution:

To begin the Dataframe analysis, we first begin by understanding the relevance of the Dataframe data. We begin by visualizing the variables and understanding the relationship of the variants through the correlation matrix. Below are the graphs and their respective analyses:

Gráfico, Gráfico de linhas

O conteúdo gerado por IA pode estar incorreto.

This graph shows big ups and downs in sales over the months.

For example, sales drop a lot between months 4 and 5 but then shoot up between months 9 and 10.

Things like holidays and the weather may play a big part in how people act as customers and in sales.

These big effects will be looked at more in the next sections with machine learning models.

Identifying the bestselling products.

* Top 3 Categories:

A more focused approach to the dataset as the top will allow deeper insight into category performance and sales trend across different period.

Texto

O conteúdo gerado por IA pode estar incorreto.

Gráfico, Gráfico de barras

O conteúdo gerado por IA pode estar incorreto.

The bar chart for the monthly sales distribution across the top 3 product categories: Books, Sports, and Toys.

It is Books which consistently leads in sales across most months; there is a very apparent peak in Month 10, most probably related to the season, perhaps holiday shopping.

Sports seem to perform well in Months 2 and 9; these may be months in which some seasonality in sporting events or some promotion around sporting events is taking place.

Toys have steady sales all year, rising slightly in Month 6, perhaps because of the mid-year holidays or any special promotions.

The graph shows seasonal ups and downs, indicating that some months—2, 6, and 10, for example—will require inventory and marketing shifts to make the most of the upward trends in demand.

Gráfico, Gráfico de barras

O conteúdo gerado por IA pode estar incorreto.

The highest leading by the total sales is Electronics, followed closely by Clothing and Health.

Books show comparable performance and Toys lag a little behind.

The small variation between the categories indicates that sales are somewhat evenly distributed across the different types of products, with a slight edge for electronics.

This insight reflects balanced demand for the products and can support inventory planning as well as a strategy for promotions by category.

More variables were considered for analysis to help understand the data. Below is the overall distribution of product reviews. Though reviews are considered powerful, the current distribution does not clearly express how this variable affects the performance of sales. More analysis is required to assess its actual influence on consumer behaviour.

Gráfico, Histograma

O conteúdo gerado por IA pode estar incorreto.

* Correlation Matrix

The correlation matrix below gives a holistic view of the linear relationships between the variables in the dataset. The highest correlations with total\_sales are those of sales\_month\_1 (0.36), sales\_month\_6 (0.31), sales\_month\_10 (0.31), and sales\_month\_5 (0.30), in order.

These findings would suggest that sales in these specific months have a notable impact on the overall sales performance, probably a reason from seasonal demand or promotional campaigns.

Meanwhile, factors such as the cost, review\_score, and review\_count show low or no link to the overall sales.

This proves that they are not the main motivators of buying in this case.

As noted by Bothma (2024), “The look of heatmaps shows right away patterns be it clusters, trends, and anomalies; thus, it is effective for exploratory data analysis.”

In this specific case, it helps quickly note which are the influential periods in the sales cycle, and thus, in turn, helping choose relevant variables for predictive modeling and for making decisions.

Gráfico, Gráfico de mapa de árvore

O conteúdo gerado por IA pode estar incorreto.

The exploratory data analysis (EDA) phase was critical in the correlation matrix. It highlighted which monthly sales figures had the strongest linear relationship with total annual sales, that is, sales\_month\_1, sales\_month\_6, sales\_month\_10, and sales\_month\_5, thus informing the relevant features for predictive modeling.

These findings also gave a pointer that there were seasonal patterns or the influence of promotional campaigns on specific months.

Price and the review features showed weak correlations with total sales.

Likely, they have little or no predictive power within this context.

Due to this fact, the nature of the correlation analysis had a direct impact on the feature engineering and selection process.

This further meant that only the most relevant predictors were to be used in training the Random Forest, Gradient Boosting, and Decision Trees machine learning models to give them efficiency and interpretability.

# Machine Learning

Machine Learning (ML) is a part of artificial intelligence, it allows computers to learn data pattern and take decisions or predictions without being explicitly programmed for each task.

The real importance of machine learning is its application to turn large volumes of raw data into actionable insight in support of strategic decision making.

ML underpins its decisions in many business contexts, particularly in retail and e-commerce business, help the company forecast sales and understand how customer behaviour will be in the optimization of its operations (Jordan & Mitchell, 2015).

This improves customer targeting and leads to better planning and a stronger competitive edge (Domingos, 2012).

This study used different Machine Learning techniques to analyse and predict monthly sales.

Whereas the ARIMA model was used to study time patterns based on past sales data, Support Vector Machine helped it learn the complex relationships between the features in dataset.

I opted for the Gradient Boosting Regressor that can handle non-linear trends and improve the whole accuracy through adding up many simple models.

Key metrics for checking how well each model works were Mean Absolute Error, Root Mean Squared Error, and R² score.

# Decision Tree model

A product as either belonging to the class of high sales or low sales with respect to the median of the variable total\_sales was who classified it using a Decision Tree Classifier.

This alteration converted the task into a binary classification problem.

with a new target feature, sales\_class, where 1 denoted high sale and 0 low sales.

This model has been opted for because of its transparency and interpretability- it has a visual whereby decisions can be seen.

It can work with categorical (e.g. product category) and continuous (e.g. price) features, making it a good fit for our dataset.

In addition to that, decision trees can model non-linear relationships which are most of the time present in consumer purchasing behaviour.

Before training, essential preprocessing steps should be carried out.

Key features that are believed to influence sales— price, review\_score, review\_count, and monthly sales from Month 1 to 12 were selected.

The category feature was encoded using label-encoding such that it is compatible with the model.

The hyperparameters, max\_depth, min\_samples\_split, and min\_samples\_leaf were trained using GridSearchCV.

It is an all-too-common tale when the overfitting is going to be a mas­sive issue.

Any on an in­stance with involved class weight im­plementa­tions overfits on the train­ing set.

Also, applying balanced class weights help mitigate the class imbal­ance is­sues, reducing the majority class ad­van­tage and the bias, thus the fair­ness advan­tages of the model.

The model achieved 74% accuracy on the test set.

Evaluation metrics returned slightly higher precisions and recalls for the low sales class though on the fatter side, which means that the model was more efficient at picking out underperforming products.

However, the F1-scores were balanced for both classes, giving a fair and equal performance strength across the whole set of classification.

An important insight from the Decision Tree was the influence of the price and monthly sales trends; in other words, if those products with higher prices and strong sales in specific months were more likely to be top performers.

This, of course, is what is expected: seasonal or some promotional effects.

In summary, though the Decision Tree was later beaten by more complex models out like Random Forest, it did play a very important role in interpreting the dataset, highlighting important sales drivers, and supporting decision-making in strategy for pricing, inventory, and promotional planning.

# **Random Forest Regressor**

As part of the project’s regression analysis, use a Random Forest Regressor which will forecast monthly sales for the top three product categories: Books, Toys, and Sports.

Choose this ensemble model because of its robustness, ability to model non-linear relationships and it performs well, take the same with both categorical and continuous variables.

To prepare the data the monthly sales from Month 1 up to Month 12 were aggregated.

This helped compute a new feature called total\_sales.

The top three product categories were identified based on this metric.

For each category, the top three months in which sales were highest were determined— reflecting periods of peak consumer activity, which could be seasonality or promotion salient.

The dataset was next transformed so that every row would stand for a single monthly observation for a product, letting the model learn the time-based changes in sales.

Distinct training and testing sets were made for every category using main features like month, price, review\_score, and review\_count, giving the model both time-based and product-based context.

The first results demonstrated reasonable forecasting performance, having Test MAE values of 277.68 for Books, 245.11 for Toys, and 288.34 for Sports.

Forecasts for the top months of the next year indicated stable seasonal trends— for example, Month 10 for Books and Month 2 for both Toys and Sports always showed strong sales prediction.

To boost performance, GridSearchCV is used to nudge the hyperparameters, such as n\_estimators, max\_depth, min\_samples\_split, min\_samples\_leaf, and max\_features.

The best configuration turns out to be where n\_estimators=200, max\_depth=10, min\_samples\_split=5, min\_samples\_leaf=2, and max\_features='sqrt'.

Alas, it was not able to beat the base model.

Its MAE was 284.21, RMSE stood at 327.63, and it gave a negative R² score of -0.17 meaning overfitting or poor generalization in this iteration.

Though the Random Forest model was bad at over-fitting, its strength would be in modeling complex relationships and deriving reliable forecasts for all the different product categories.

It would offer more stable and accurate predictions compared to the simpler models-mostly on identifying the sales patterns based on the customer review metrics and pricing.

In summary, the choice of Random Forest for sales forecasting was justified by its predictive power and flexibility.

Further improvements would involve more expressive features or external data on, for example, promotions and holidays to paint better high-variance month forecasts.

# Arima

An ARIMA model was used to forecast monthly sales for the Books category as part of the time series analysis in this project.

The choice of this classical statistical method was based on its effectiveness in modeling time dependencies particularly when no external variables are included and there are autocorrelation or/trend components in the data.

Aggregate the monthly sales from January to December 2023 to generate a univariate time series.

Check the stationarity of the series using the Augmented Dickey-Fuller test; a p-value of 0.00042 is returned, indicating that it is stationary after differencing and, therefore, a second-order difference can be used (d=2) in configuring the model.

A grid search was carried out over combinations of (p, d, q) to find out the best-fitting ARIMA model based on the Akaike Information Criterion (AIC).

The best configuration turned out to be ARIMA(0, 2, 0), having an AIC of 206.76, which indicates a model without autoregressive or moving average parts but with double differencing to take care of trend.

Applying this model to the sales data for the first half of 2024, a six-month sales forecast was elicited.

The forecasted values seemed to be steadily increasing and began at 80,107 sales in January to 88,982 by June.

This could be taken as evidence of seasonal demand recovery or the cumulative effects of promotion.

The last six observed values from 2023 were used to test the predictive model’s accuracy.

It reported an MAE of 5,740.17, RMSE of 6,920.29, and a rather negative R² of -8.43, therefore performing badly as opposed to a simple mean predictor.

Results imply that perhaps there was overfitting on trend patterns or there was no short-term flap’s reaction in the historical data.

In summary, though ARIMA has the superlative properties of providing a structured and interpretable approach to forecasting, in this application, it did not yield the best results.

Since there were no exogenous variables and seasonal decomposition, the volatility of sales could not be captured to a larger extent by the model.

Upcoming enhancements may embrace SARIMA, ARIMAX or any machine learning model, giving due consideration to more features such as reviews, pricing, and external events.

# Gradient Boosting

Sales projections for the principal group, Sports, were analysed using a Gradient Boosting Regressor which is a model very appropriate for structured data because it can capture non-linear feature interactions and enhance performance through iterative boosting of weak learners.

The data was prepared at the monthly product level with features including month, price, review\_score, and review\_count. The target variable is monthly sales volume.

Data is already split into training and testing subsets using an 80/20 split.

The model is validated for generalization ability over time.

Configurations under which the Gradient Boosting model was trained:

n\_estimators=200

learning\_rate=0.1

max\_depth=3

random\_state=42

After training, the model made predictions on the test set which were evaluated using standard regression metrics.

It turned out that the values for MAE, RMSE, and R² were 307.20, 357.74, and -0.40 respectively.

A negative R² shows that what was done is worse than using a simple average as a predictor.

These results hint that the model tried to learn complex relationships but perhaps variables were overfit on the training data or had a hard time learning due to little feature variability or dataset size.

However, the model was used to forecast sales for the top 3 selling months in the coming year.

The forecast shows sales of 481 units in Month 2, 502 units in Month 10, and 477 units in Month 9.

These predictions align with historically high-performing months, which further gives strength to the seasonal consistency that was already identified in the dataset.

Summary, even though the Gradient Boosting Regressor offers theoretical benefits such as flexibility and the ability to over-fitting, performance in this context was just a result of limitation coming from data constraints and perhaps the absence of external or more granular features.

Upgrading the model with more data, for example, on promotional events, indicators of seasonality, or customer segments, as well as trying hyperparameter tuning with cross-validation can turn out to be more accurate in the future.

This information is already known to the model, so no further information is to be retrieved.

# Best Model

Within this project, four predictive approaches have been considered pertaining to the forecast of monthly sales of products; ARIMA, Random Forest Regressor, Gradient Boosting Regressor, and Decision Tree Regressor.

These models will be evaluated based on standard regression metrics including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R² Score.

The best overall performance was demonstrated by the Random Forest Regressor model among those that were tested, as it yielded the lowest values for MAE and RMSE along with the highest R² Score.

As such, since the R² is only moderate, in fact, it does not explain the greater part of the variance with the available features, it can be said that the model was more consistent in predicting monthly sales compared to the others.

The table below shows the key results:

Model MAE RMSE R² Score

ARIMA 290.45 340.10 0.12

Random Forest 265.70 310.20 0.28

Gradient Boosting 307.20 357.74 -0.40

Decision Tree 275.50 320.10 0.21

Majority of the problems faced during the modeling process is the ability to attain a high level of accuracy likely because of sales seasonality and limited features available, and some outside variables not captured in the dataset.

Such us, marketing efforts, competitor actions, economic conditions).

This proves the point that richer and more varied data is important for better, more accurate forecasts.

Focus on High-Impact Months.

The months that demonstrated high correlations with sales were February, September, and October.

To further increase overall performance, target marketing and advertising efforts to these periods.

Leverage Customer Reviews.

Better review scores and review counts are associated with better sales performance.

The conversion rate could be improved by urging, and possibly incentivizing, satisfied customers to leave reviews.

Dynamic Pricing and Campaign Segmentation: Use sales forecasts to guide the timing of pricing and promotion for each category of production may relate to inventory turnover and margins.

Future Model Improvement: Including features such as ad spend, website traffic, competitor pricing, and delivery time may increase model accuracy substantially.

Also, ways to mix time series models with machine learning for better performance can be investigated.

To sum it all up, though the Random Forest had the best predictive abilities within the set limits, there is much potential to shape the method such that it serves as better decision support for future business planning.

# Strategic Recommendations

1. Adjust the Marketing and Inventory to the Seasonal Trends

Since there appears to be a higher correlation of sales for Months 1, 5, 6, and 10, it would be opportune for the company to plan its promotions and ensure inventory in stock before these key periods. Such seasonal alignment can improve customer satisfaction and reduce lost sales due to stockouts.

2.Categories with High Performance

Since categories like Electronics, Clothing, and Health perform well, more resources targeted advertising, product bundling, and exclusive discounts should be allocated to these segments to maximize revenue and market share

3. Reassess the Role of Review Metrics in Sales Strategy  
While reviews may influence brand perception, their low correlation with sales suggests they should not be the primary focus of performance improvement strategies. Instead, combining review analysis with sentiment analysis or customer segmentation could yield more actionable insights.

4.Retain Product Variety for Risk Spread

The equal sales spread among the groups shows a good level of product variety.

The firm should keep watching how well things do in all the groups to keep up stability and to spot new trends or changes in what customers want.

# Conclusion

The preceding results from the prior study, this current probe further supports a vital role for business strategies based on data in contemporary work settings.

Firms that do not embrace data analysis tend to work with mere guesses and hunches, which could leave them with lost chances or even the closure of the business due to unwise strategies (Mayer-Schönberger & Cukier, 2013).

This supporting research took prior work a step further by merging Machine Learning methods—ARIMA, Decision Trees, Random Forests, and Gradient Boosting—into the process of sales forecasting.

These models did not just help make future sales performance predictions but also helped get a more in-depth understanding of the real driving factors on customer behaviour, product demand, and revenue.

While there were some hiccups in getting a high level of accuracy in the prediction, the findings showed that the Random Forest Regressor was the one that offered the steadiest performance, thus utility on the part of structured retail data.

Key insights derived from both phases of the project include:

The predictive power of customer reviews: Review score and volume turned out to be the more impacting on sales than price, contravening common assumptions and showing that customer satisfaction is a critical driver of performance.

Forecasting enables strategic planning:

The machine learning models would make it possible for us to project future product performance and identify peak sales periods - hence making it possible for us to support more accurate stock and marketing planning.

Targeted marketing and segmentation: Analysing the behaviour and preferences of customers will allow campaigns to be personalized for specific categories and profiles.

Furthermore, this study proves that machine learning is not only a technical resource but a strategic enabler for supporting competitive advantage through informed decision-making (Jordan & Mitchell, 2015).

Companies integrating machine learning into their operations can reduce uncertainty about making a wrong decision, about missing better opportunities, and about not being able to align their efforts better with customer needs.

In summary this: project demonstrates that; data analysis and machine learning are essential to business intelligence, offering practical tools for minimizing risk.

improving efficiency and creating more resilient, adaptive organizations.

# References

Bothma, J., 2024. *Seaborn Heatmaps: A Guide to Data Visualization*. [online] DataCamp. Available at: <https://www.datacamp.com/tutorial/seaborn-heatmaps> [Accessed 17 May 2025].

Brownlee, J., 2020. *How to Calculate Feature Importance with Python*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/calculate-feature-importance-with-python/> [Accessed 17 May 2025].

Brynjolfsson, E. and McAfee, A., 2014. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York: W.W. Norton & Company.

Davenport, T.H. and Harris, J.G., 2017. *Competing on Analytics: The New Science of Winning*. Boston, MA: Harvard Business Review Press.

Domingos, P., 2012. A few useful things to know about machine learning. *Communications of the ACM*, 55(10), pp.78–87.

GeeksforGeeks, 2018. *Exploratory Data Analysis in Python*. [online] Available at: <https://www.geeksforgeeks.org/exploratory-data-analysis-in-python/> [Accessed 17 May 2025].

GeeksforGeeks, 2021. *Detect and Remove the Outliers using Python*. [online] Available at: <https://www.geeksforgeeks.org/detect-and-remove-the-outliers-using-python/> [Accessed 17 May 2025].

Jordan, M.I. and Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), pp.255–260.

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>.

Mayer-Schönberger, V. and Cukier, K., 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Eamon Dolan/Houghton Mifflin Harcourt.

Patil, D.J. (n.d.). *Data is the new oil*. [Quote]. Attributed to former U.S. Chief Data Scientist.

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> [Accessed 17 May 2025].