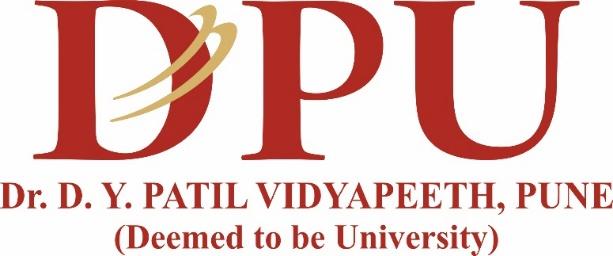
****

**DR. D. Y. PATIL SCHOOL OF SCIENCE AND TECHNOLOGY**

**TATHAWADE, PUNE**

**A Mini- Project Report on**

**DATA TO DECISION UNLOCKING BUISNESS GROWTH WITH DATA-DRIVEN INSIGHTS**

**SUBMITTED BY:**

**NAME OF STUDENT ROLL NUMBER**

**1. Suyash Raj BTAI-55**

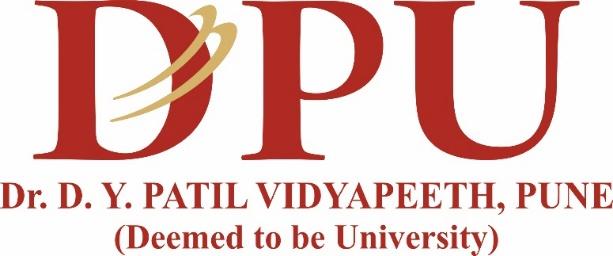
**2. Raunak Kumar BTAI-42**

**GUIDED BY:**

**Mrs . MILY LAL**

**ARTIFICIAL INTELLIGENCE & DATA SCIENCE**

**ACADEMIC YEAR 2023-2027**

****

**DR. D. Y. PATIL SCHOOL OF SCIENCE AND TECHNOLOGY**

**TATHAWADE, PUNE**

**CERTIFICATE**

**This is to certify that the Mini- Project Report entitled**

**DATA TO DECISION UNLOCKING BUISNESS GROWTH WITH DATA-DRIVEN INSIGHTS**

is a bonafide work carried out by Mr. Suyash Raj under the supervision of **Mrs. Mily Lal** and it is submitted towards the partial fulfillment of the requirement Project Based Learning-I.

Mrs. Mily Lal Prof. Manisha Bhende

**Project Guide HOD (DPSST)**

**ARTIFICIAL INTELLIGENCE & DATA SCIENCE**

**ACADEMIC YEAR 2023-2027**

**ABSTRACT**

E-commerce platforms generate vast amounts of data daily, yet many businesses struggle to extract meaningful insights from it. This inability to leverage data effectively hinders crucial aspects such as sales optimization, targeted marketing, and customer experience enhancement, ultimately affecting growth and operational efficiency. Without data-driven decision-making, businesses miss opportunities to maximize revenue, personalize customer interactions, and optimize inventory management. This project aims to bridge this gap by applying data science techniques to analyse sales trends, segment customers, and build predictive models for improved marketing strategies and customer retention. By leveraging machine learning and data analytics, businesses can gain actionable insights, enhance customer engagement, and improve overall efficiency. The project will also develop visualization tools to help stakeholders interpret complex data easily, facilitating better strategic decisions. The ultimate goal is to transform raw e-commerce data into valuable intelligence that drives growth, increases profitability, and ensures a competitive edge in the digital marketplace . .   
  
  
**Keywords:** E-commerce, Revenue Growth, Data-Driven Decision Making, Sales Optimization, Customer Experience

\

**INDEX**

|  |  |  |
| --- | --- | --- |
| **SR.NO** | **TOPIC** | **PAGE NO** |
| 1] | INTRODUCTION | 6-10 |
| 1.1 Problem Statements | 6 |
| 1.2 Objective | 7 |
| 1.3 Scope | 8 |
| 1.4 System Architecture | 9-10 |
| 2] | DATA COLLECTION & PREPROCESSING | 11 |
| 2.1 Dataset |  |
| 2.2 Data Preprocessing |  |
| 3] | MODEL SELECTION & TRAINING |  |
| 3.1 Feature Engineering |  |
| 3.2 Machine Learning Model |  |
| 4] | MODEL EVALUATION AND VALIDATION |  |
| 4.1 Performance Metrics |  |
| 5] | CONCLUSION & FUTURE SCOPE |  |
| 6] | REFERENCES |  |

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **SR.NO** | **FIGURE NO.** | **PAGE NO** |
| **1** | System Architecture | 9 |
|  |  |  |
|  |  |  |

**Chapter 1**

**INTRODUCTION**

In today’s digital world, e-commerce platforms generate a huge amount of data every day. This data comes from customer activities such as browsing, shopping, and making payments. While this data has the potential to offer valuable insights, many businesses are unable to use it effectively. As a result, they often miss out on chances to increase sales, improve marketing, manage inventory better, and provide a more personalized customer experience. This project aims to solve that problem by using data science techniques to make sense of e-commerce data. The goal is to help businesses analyze sales trends, group customers based on their behavior (customer segmentation), and create models that can predict future actions. These insights can then be used to develop better marketing strategies, improve customer satisfaction, and boost sales. One key part of the project is creating visual dashboards. These dashboards will make it easy for business owners and managers to understand complex data through clear graphs and charts. This helps them make smarter and faster decisions without needing deep technical knowledge. By using machine learning and data analytics, this project will turn raw data into useful information. It will allow businesses to understand their customers better, plan smarter marketing campaigns, and improve their overall performance. In the end, this project aims to give businesses a strong advantage in the competitive world of e-commerce by helping them grow, earn more, and make data-driven decisions confidently.

**1.1. Problem Statement**

E-commerce platforms generate vast amounts of data daily, yet many businesses fail to extract meaningful insights from it. This limits their ability to make informed decisions, optimize sales, personalize customer experiences, and manage inventory effectively. Without data-driven strategies, businesses miss out on growth opportunities and competitive advantage. This project aims to solve this problem by applying data science techniques to analyze e-commerce data and provide actionable insights for better decision-making.

**1.2.Objective**

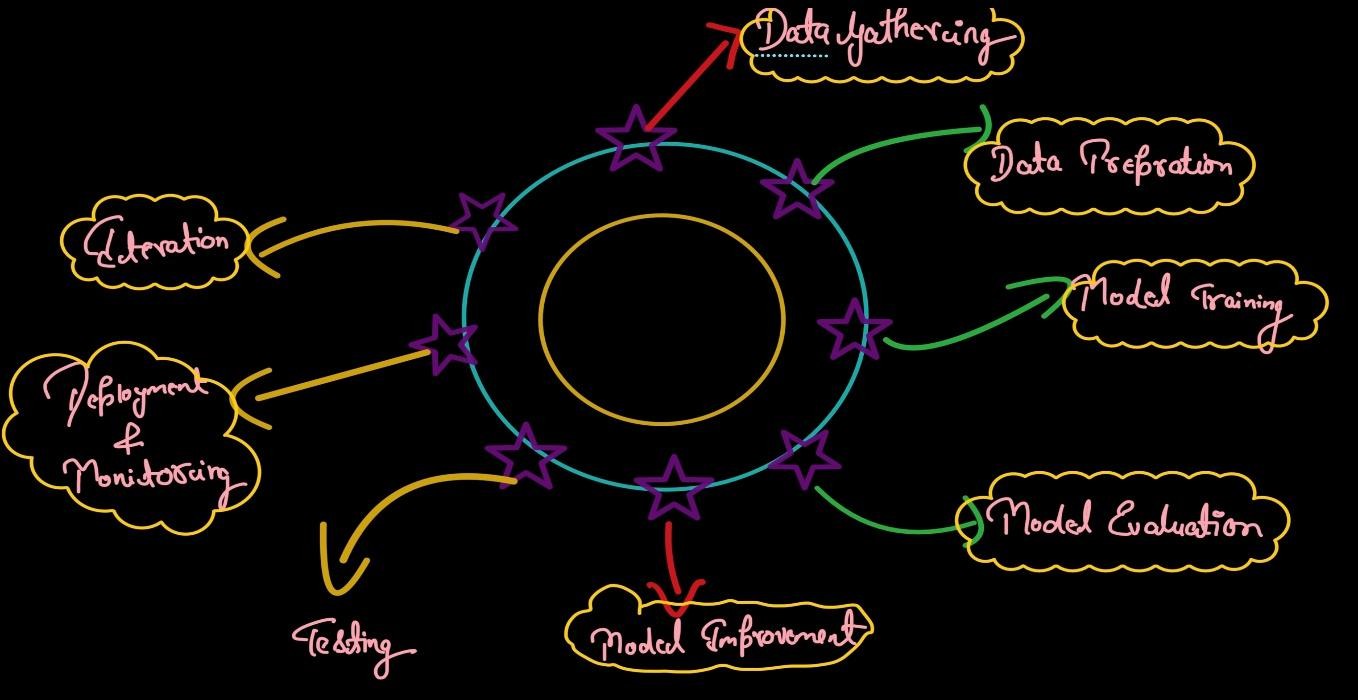
* Analyse e-commerce data to identify sales trends, improve pricing strategies, and enhance targeted marketing efforts for better customer engagement and revenue growth.
* Use data science techniques to segment customers based on preferences and purchasing patterns, enabling personalized recommendations and improved customer experience.
* Develop visualization tools and predictive models to help businesses interpret complex data, optimize inventory management, and make informed strategic decisions.

**1.3. Scope**

This project focuses on analyzing e-commerce data to help businesses make smarter decisions. It covers collecting and cleaning data, identifying sales trends, and understanding customer behavior. The project includes building customer segments, creating predictive models, and designing dashboards for easy data visualization.

The insights gained can help improve marketing, pricing, inventory management, and customer experience. The project is useful for e-commerce platforms of all sizes that want to grow using data-driven strategies.

**1.4 System Architecture**

****

**Fig.1**

The system architecture involves collecting e-commerce data, preprocessing and cleaning it, followed by feature engineering. Machine learning models are then applied for customer segmentation and sales prediction. Finally, results are visualized through dashboards, enabling businesses to make informed, data-driven decisions for growth and optimization.

**Chapter 2**

**DATA COLLECTION & PREPROCESSING**

**2.1. Dataset**

The dataset for this project consists of e-commerce transaction records, including customer details, product information, purchase history, browsing behaviour, and marketing engagement metrics. The data can be sourced from public e-commerce datasets such as Kaggle, UCI Machine Learning Repository, or real-time data collected from an e-commerce platform’s database via web scraping or APIs. This dataset is crucial as it provides insights into customer purchasing behaviour, sales trends, and marketing effectiveness. By analysing this data, businesses can improve sales forecasting, customer segmentation, personalized recommendations, and operational efficiency. The dataset's relevance lies in its ability to help businesses make data-driven decisions, optimize marketing strategies, and enhance customer experience, ultimately driving growth and profitability in the competitive e-commerce landscape.

It contains important information such as:

* **Customer Demographics**: Ship City, and Ship State, which help in understanding that in which state and city we are selling more or less.
* **Product Details**: Product\_Category, Size of clothes, and Quantity, which are crucial for understanding the sales trend of particular category and size.

This dataset is important because it helps businesses understand how customers shop, what products they like, and how they respond to marketing. By studying this data, companies can find patterns in sales, improve their marketing efforts, and make smarter business decisions. It also helps in giving customers a better experience through personalized recommendations. Overall, the dataset supports better planning, increases profits, and gives businesses an edge in the e-commerce market.

**2.2. Data Preprocessing**

### 1. Handling Missing Values

* Identification: Used df.isnull().sum() to check for missing values.
* Strategy:  
  + Numerical columns : Filled with the median to avoid skewness.
  + Categorical columns : Filled with the mode (most frequent value) to maintain consistency.
  + Columns with 100% missing values: Dropped using df.drop(columns=["Unnamed: 8"], errors="ignore").

### 2. Removing Duplicates

* Identification: Used df.duplicated().sum() to find duplicate entries.
* Strategy: Removed duplicate rows using df.drop\_duplicates(inplace=True) to ensure unique transactions.

### 3. Data Normalization

* Identification: Found that columns like Order\_Id and Amount had significantly different value ranges, which could impact model performance.

Strategy: Applied Min-Max Scaling using:  
  
 from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()

df[numeric\_columns] = scaler.fit\_transform(df[numeric\_columns])

* This ensured values were in the range of 0 to 1 for better model performance.

### 4. Handling Outliers

* Method Used: Interquartile Range (IQR)

Strategy: Outliers were removed using:  
  
 def remove\_outliers\_iqr(df, column):

Q1 = df[column].quantile(0.25)

Q3 = df[column].quantile(0.75)

IQR = Q3 - Q1

lower\_bound = Q1 - 1.5 \* IQR

upper\_bound = Q3 + 1.5 \* IQR

return df[(df[column] >= lower\_bound) & (df[column] <= upper\_bound)]

* This was applied to price-related columns such as:  
  + Amount

**Chapter 3**

**MODEL SELECTION & TRAINING**

**3.1. Feature Engineering**

Feature engineering plays a vital role in improving the performance and accuracy of data analysis and machine learning models. In this project, we transformed raw e-commerce data into meaningful and structured features to uncover valuable insights.

Category Normalization: To ensure consistent analysis, all clothing categories were converted to uppercase and stripped of extra spaces. This helped group similar categories accurately (e.g., "tshirt", "TShirt", and "T-SHIRT" were treated as the same).

Size Encoding: Sizes were standardized (e.g., "S", "M", "L", "XL") to ensure consistency in grouping and analysis. Any ambiguous or missing size values were cleaned or labeled as "Unknown".

Quantity Aggregation: The total quantity sold (Qty) was grouped by different features like Size, State, and City to understand demand patterns across different dimensions.

Geographic Mapping: By combining state and city information, we were able to drill down and analyze region-specific trends in sales, helping us identify top-performing and underperforming areas.

**3.2. Machine Learning Model**

In this project, several machine learning models were implemented to analyze and predict e-commerce trends. These models were used for tasks like sales prediction and customer behavior classification.

Random Forest Regressor

This model was used to predict future sales quantities based on features such as product category, size, and location. The Random Forest performed reasonably well with:

These results suggest that while the model captured some patterns, further tuning and more features could improve its predictive power.

Gradient Boosting Classifier

This model was applied for classification tasks, such as predicting whether a customer belongs to a high-value segment. It achieved:

Accuracy Score: 1.0

This perfect accuracy might be due to imbalanced data or overfitting, so further validation is recommended.

XGBoost Classifier

Like Gradient Boosting, XGBoost was used for classification and also achieved:

Accuracy Score: 1.0

XGBoost is known for its speed and performance on structured data. Similar to Gradient Boosting, its results should be validated using techniques like cross-validation.

By leveraging these models, the project demonstrated how machine learning can enhance understanding of customer behavior and forecast future trends in the e-commerce domain.

**Chapter 4**

**MODEL EVALUATION & VALIDATION**

**4.1. Performance Metrics**

To evaluate the effectiveness of the machine learning models used in this project, appropriate performance metrics were selected based on the type of task—regression or classification.

Regression Metrics (used with Random Forest Regressor):

Mean Absolute Error (MAE):

Measures the average magnitude of errors in predictions. Lower values indicate better accuracy.

MAE = 174.98

Mean Squared Error (MSE):

Measures the average squared difference between actual and predicted values. More sensitive to large errors.

MSE = 63,067.36

Root Mean Squared Error (RMSE):

The square root of MSE; provides error in the same units as the target variable.

RMSE = 251.13

R² Score (Coefficient of Determination):

Indicates how well the model explains the variance in the data. A score closer to 1 means better fit.

R² Score = 0.2636

Classification Metrics (used with Gradient Boosting and XGBoost Classifiers):

Accuracy Score:

The ratio of correctly predicted instances to the total instances. A perfect model scores 1.0.

Gradient Boosting Accuracy = 1.0

XGBoost Accuracy = 1.0

**Chapter 5**

**CONCLUSION & FUTURE SCOPE**

**Conclusion**

This project focused on analyzing e-commerce transaction data to identify key patterns in the sale of clothing items. By examining various factors such as product categories, sizes, and geographic locations, we were able to uncover valuable insights that can help businesses optimize their strategies. Our analysis revealed which clothing categories are the most popular, allowing businesses to focus on stocking and promoting high-demand products. We found that certain sizes of clothes are in higher demand than others, which helps in better inventory management, ensuring that the most popular sizes are always available. Geographically, the data highlighted the states and cities where sales are highest and lowest. By identifying top-performing regions, businesses can target marketing campaigns more effectively and allocate resources to areas with higher sales potential. Conversely, understanding where sales are low allows companies to investigate potential reasons and adjust strategies to increase sales in those areas. Overall, the insights derived from this analysis can guide decision-making in areas like inventory planning, marketing, and customer engagement. By knowing which products, sizes, and locations are performing well, businesses can create more effective, data-driven strategies to boost sales and improve customer satisfaction. This project demonstrates the power of data science in helping e-commerce platforms make smarter, informed decisions that can ultimately lead to increased profitability and growth in a competitive marketplace.

**Future Scope**

The future scope of this project includes several enhancements. Real-time data analysis could enable businesses to monitor sales trends, inventory, and customer preferences continuously, providing up-to-date insights. Predictive analytics can be integrated to forecast future sales and demand, helping businesses plan stock and marketing efforts more effectively. Machine learning algorithms can further segment customers based on behavior and purchasing patterns, enhancing personalized recommendations and marketing strategies. Additionally, incorporating customer sentiment analysis from reviews and feedback will allow businesses to gauge customer satisfaction and make necessary improvements. Expanding the analysis to cover a wider range of products, regions, and e-commerce platforms will enable businesses to optimize strategies for a global market. By refining the analysis continuously, businesses can remain competitive and thrive in the dynamic e-commerce landscape.

**REFERENCES**

1. **https://www.kaggle.com/datasets**
2. **https://colab.research.google.com/drive/1nXHg2OVO5tjoUhlV6thIUtU7exWw89bo?usp=drive\_link**
3. **https://www.researchgate.net/publication/380127515\_Data\_Science\_in\_E-commerce**