Project Report

On

"TRANSFORMING E-COMMERCE GROWTH: A DATA-DRIVEN MARKETING STRATEGY FOR ENHANCED CUSTOMER ACQUISITION AND RETENTION"

Submitted in the partial fulfillment for the award of

Post Graduate Diploma in Big Data Analytics (PG-DBDA)

from Know-IT ATC, CDAC ACTS, Pune

Guided by:

Mr. Milind Kapse

Submitted By:

Aman (240843025006)

Avirat (240843025010)

Rohit (240843025030)

Sahil (240843025034)

CERTIFICATE

TO WHOMSOEVER IT MAY CONCERN

This is to certify that

Aman (240843025006)

Avirat (240843025010)

Rohit (240843025030)

Sahil (240843025034)

have successfully completed their project on

"Transforming E-Commerce Growth: A Data-Driven Marketing Strategy for Enhanced Customer Acquisition and Retention"

ACKNOWLEDGEMENT

This project **"Transforming E-Commerce Growth: A Data-Driven Marketing Strategy for Enhanced Customer Acquisition and Retention"** was a great learning experience for us and we are submitting this work to Know-IT ATC, CDAC ACTS,Pune.

We are all very glad to mention the names of **Mr. Milind Kapse** for their valuable guidance on this project. Their continuous guidance and support helped us overcome various obstacles and intricacies during the course of the project work.

We are highly grateful to **Mr. Vaibhav Inamdar,** the Center Coordinator at Know-it in Pune, for his guidance and support whenever necessary while we were pursuing the Post Graduate Diploma in Big Data Analytics (PG-DBDA) through C-DAC ACTS in Pune.

Our most heartfelt thanks goes to **Mr. Shrinivas Jadhav** (Vice-President, Know-it, Pune) who provided all the required support and his kind co-ordination to provide all the necessities like required hardware, internet facility and extra lab hours to complete the project, throughout the course and till date, here in Know-IT ATC, CDAC ACTS, Pune.

TABLE OF CONTENTS

ABSTRACT

In the digital age, understanding customer behaviour is critical for businesses aiming to enhance customer retention and maximize profitability. This project utilizes advanced machine learning techniques to perform customer segmentation and predict future purchasing behaviour. By integrating datasets like online sales transactions, customer demographics, marketing spend, and discount coupons, we developed models to identify key purchasing patterns.

The primary objective is twofold: to segment customers based on their purchasing behaviour using RFM (Recency, Frequency, Monetary) analysis, and to predict the likelihood of future purchases along with potential product recommendations. A Random Forest Classifier was implemented for these predictions due to its robustness and high accuracy. Additionally, a user-friendly Streamlit application was developed to enable real-time interaction with the models, providing actionable insights for marketing teams.

1. INTRODUCTION

In today’s competitive market landscape, customer retention is as crucial as customer acquisition. Businesses that understand their customers' purchasing behaviour can tailor marketing strategies to enhance customer satisfaction and loyalty. This project aims to leverage machine learning techniques for effective customer segmentation and predictive analytics.

Customer Segmentation is the process of dividing customers into distinct groups based on common characteristics. This helps businesses personalize marketing efforts and improve service delivery.

Predictive Analytics involves using historical data to forecast future behaviour, which in this context refers to predicting the likelihood of a customer making a future purchase and identifying the products they are likely to buy.

The datasets used in this project include online sales records, customer demographic details, marketing expenditures, discount coupon usage, and tax information. By merging these datasets, we created a comprehensive view of customer interactions and purchasing patterns.

2. SYSTEM REQUIREMENTS

Hardware Requirements:

- **Processor:** Intel i5 or above / AMD Ryzen equivalent
- **RAM:** 8 GB (minimum), 16 GB (recommended)
- **Storage:** 500 MB for datasets and models, 5 GB for overall project files
- **Display:** 1080p resolution for optimal visualization
- **Internet Connection:** Minimum 2 Mbps for ngrok and data fetching

Software Requirements:

- **Programming Language:** Python 3.x
- **Libraries and Frameworks:**
  - pandas: For data manipulation and analysis
  - scikit-learn: For machine learning model development
  - streamlit: For creating interactive web applications
  - pyngrok: For creating secure tunnels to localhost
  - matplotlib & seaborn: For data visualization
- **Integrated Development Environment (IDE):** Jupyter Notebook, VSCode
- **Operating System:** Windows 10/Linux Ubuntu 18.04 or higher

3. FUNCTIONAL REQUIREMENTS

1. **Data Integration and Cleaning:**
    - Merge multiple datasets (Online Sales, Customer Data, Marketing Spend, Discount Coupons) based on CustomerID.
    - Handle missing values and ensure data consistency.

2. **Feature Engineering:**
    - Create new features such as Days_Since_Last_Transaction, Next_Purchase, and
    - Next_Product to enhance model performance.
    - Perform RFM analysis to categorize customers into value segments.

3. **Model Training and Evaluation:**
    - Implement machine learning models using Random Forest Classifier.
    - Evaluate models using accuracy, precision, recall, and F1-score.

4. **Interactive Web Application:**
    - Develop a Streamlit app to allow users to input CustomerID and get real-time predictions.
    - Use pyngrok to deploy the app for remote access.

5. **Data Visualization:**
    - Create visual representations of customer segments, purchasing behaviour, and
    - model predictions.

4. SYSTEM ARCHITECTURE

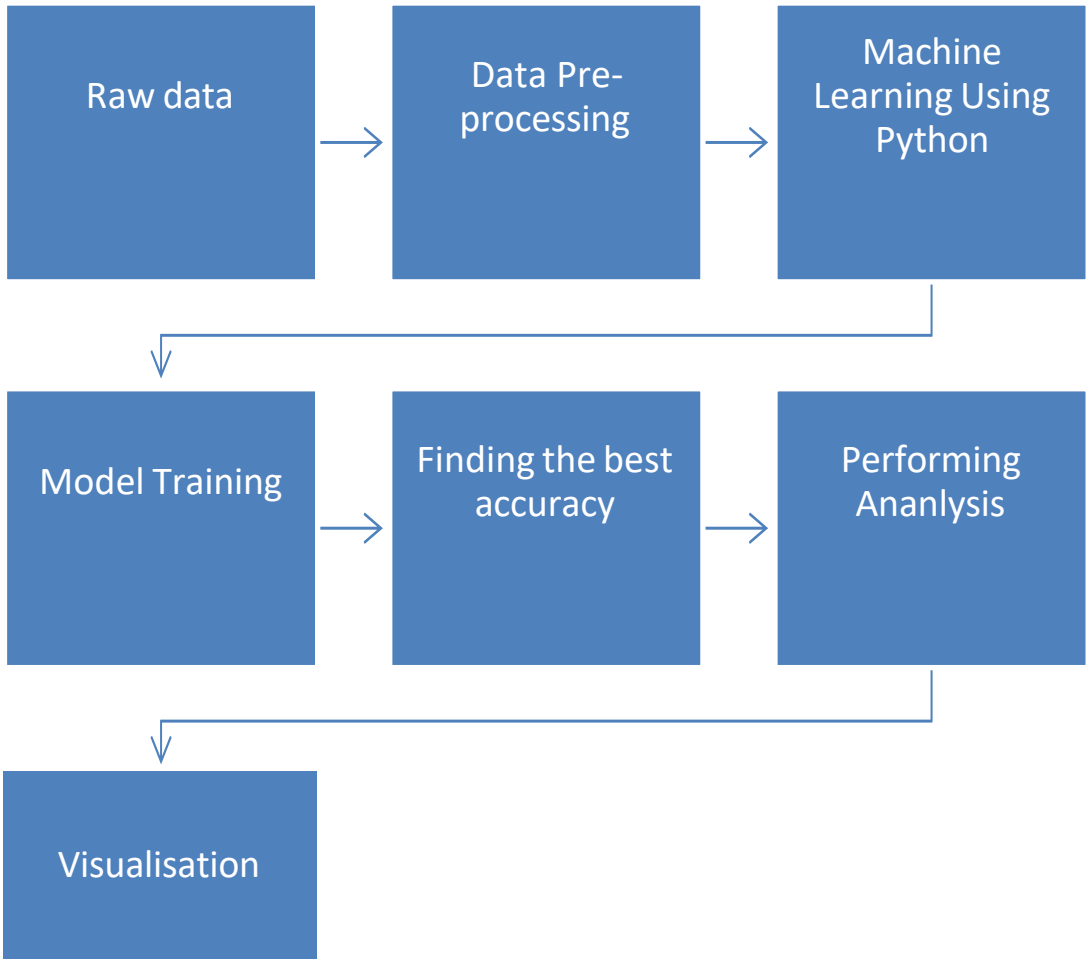


Fig: System Architecture of Crop Yield Prediction

The system architecture for this project involves the following components:

1. **Data Collection:**
   - Datasets collected from various sources including sales records, customer information, and marketing data.

2. **Data Preprocessing:**
   - Clean, merge, and prepare data for analysis.
   - Convert date fields to datetime format and sort transactions by CustomerID and Transaction_Date.

3. **Feature Engineering:**
   - Create features like Next_Purchase (binary indicator if a customer will make another purchase) and Next_Product (predicting the product category of the next purchase).

4. **Model Development:**
   - Train Random Forest models for purchase prediction and product recommendation.
   - Save models as model.pkl and product_model.pkl for deployment.

5. **Deployment:**
   - Develop a Streamlit app to visualize and interact with the model predictions.
   - Use pyngrok for creating a secure, publicly accessible URL.

6. **Visualization and Reporting:**
   - Generate visual insights into customer segmentation and purchasing patterns.

5. METHODOLOGY

1. **Data Loading:**
    - Load Online_Sales.csv, CustomersData.xlsx, Marketing_Spend.csv, and Discount_Coupon.csv using pandas.

2. **Data Merging and Cleaning:**
    - Merge datasets on CustomerID and handle missing values.
    - Convert Transaction_Date to datetime and sort transactions for each customer.

3. **Feature Engineering:**
    - Calculate Days_Since_Last_Transaction to understand purchasing intervals.
    - Create Next_Purchase as the target variable for predicting future purchases.
    - Create Next_Product by shifting the Product_Category to predict future product purchases.

4. **Model Training:**
    - Split data into training and testing sets using train_test_split.
    - Train Random Forest Classifiers for both purchase and product predictions.
    - Save models using the pickle library.

5. **Model Evaluation:**
    - Evaluate model performance using accuracy, precision, recall, and F1-score.
    - Compare model predictions with actual data to assess performance.

6. **App Development and Deployment:**
    - Develop a Streamlit app for interactive predictions.
    - Use pyngrok to create a public URL for the app, enabling remote access.

6. MACHINE LEARNING ALGORITHMS

- Machine learning is a subfield of artificial intelligence that involves developing algorithms and models that enable computers to learn from data and make predictions or decisions without being explicitly programmed. The goal of machine learning is to enable computers to improve their performance over time by learning from experience and feedback.
- In our project, we applied various Regression Algorithms such as Random Forest, Decision Tree, Linear Regression, Polynomial Regression, and Gradient-Boosting.. After the implementation, were able to analyze the accuracy of the algorithms on our data.
- Accuracy was one of the major factors that helped to decide which model has the accurate predictions.

K-Means Clustering:

K-Means is an unsupervised learning algorithm used for customer segmentation.

It partitions customers into clusters based on similarities in their purchasing behavior.

Pros:

- K-means is easy to understand and implement. The algorithm is straightforward, making it accessible for beginners.
- The algorithm can handle large datasets
- K-means can be applied to various types of data and is widely used in different fields, such as marketing, biology, and image processing.

Cons:

- The number of clusters (K) must be specified in advance, which can be challenging**.**
- The final clusters can depend on the initial placement of centroids. Poor initialization can lead to suboptimal clustering.
- K-means is sensitive to outliers, which can skew the results and affect the placement of centroids.

Random Forest:

Random forest is a machine learning algorithm that is used for classification, regression, and feature selection tasks. It is an ensemble method that combines multiple decision trees, where each tree is trained on a subset of the training data and a subset of the input features.

Pros:

- It is a highly accurate and powerful machine learning algorithm that can perform well on a wide range of classification and regression tasks.
- It can handle both categorical and continuous input variables, and it can detect and handle interactions between variables.

Cons:

- It may not perform well on small datasets or with rare or unseen classes, which may require more specialized techniques or models.
- It may not be suitable for online or real-time prediction tasks, which require faster and more lightweight models or techniques.

Support Vector Classifier (SVC)

Support Vector Classifier (SVC) is a supervised machine learning algorithm that is part of the Support Vector Machine (SVM) family. It is primarily used for classification tasks, although it can also be adapted for regression.

Pros:

- SVC is particularly effective in high-dimensional spaces, making it suitable for text classification
- SVC can be robust to overfitting, especially in high-dimensional datasets.
- SVC aims to maximize the margin between classes, which can lead to better generalization on unseen data.

Cons:

- SVC can be computationally expensive, especially with large datasets
- Selecting the appropriate kernel and tuning its parameters can be challenging and may require domain knowledge or experimentation.
- SVC can be sensitive to noisy data and outliers, which can affect the position of the hyperplane and the support vectors.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a widely used statistical technique for dimensionality reduction and data analysis. It transforms a dataset into a new coordinate system, where the greatest variance by any projection of the data lies on the first coordinate (the first principal component), the second greatest variance on the second coordinate, and so on. Here's an overview of PCA, including its working principle, advantages, disadvantages, and applications.

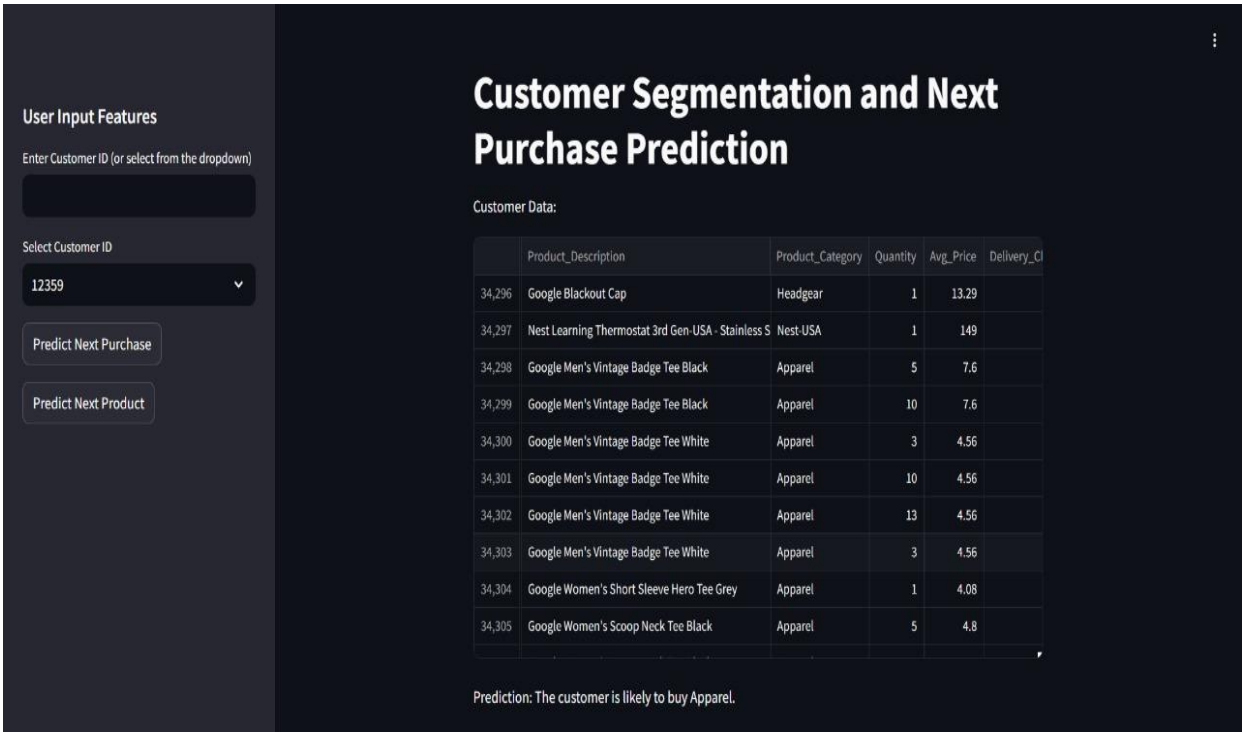
Pros:

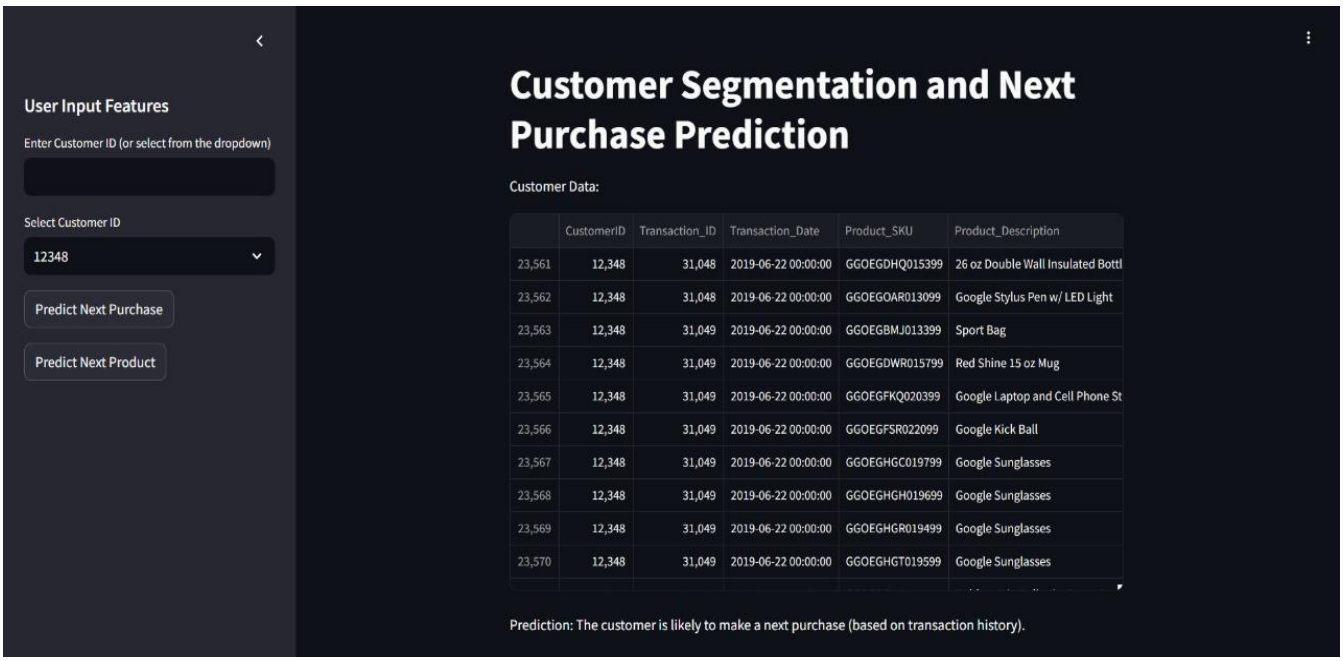
- PCA effectively reduces the number of features while preserving as much variance as possible.
- By focusing on the principal components that capture the most variance, PCA can help filter out noise and irrelevant features, leading to cleaner data.
- PCA can be used to create new features that are linear combinations of the original features.

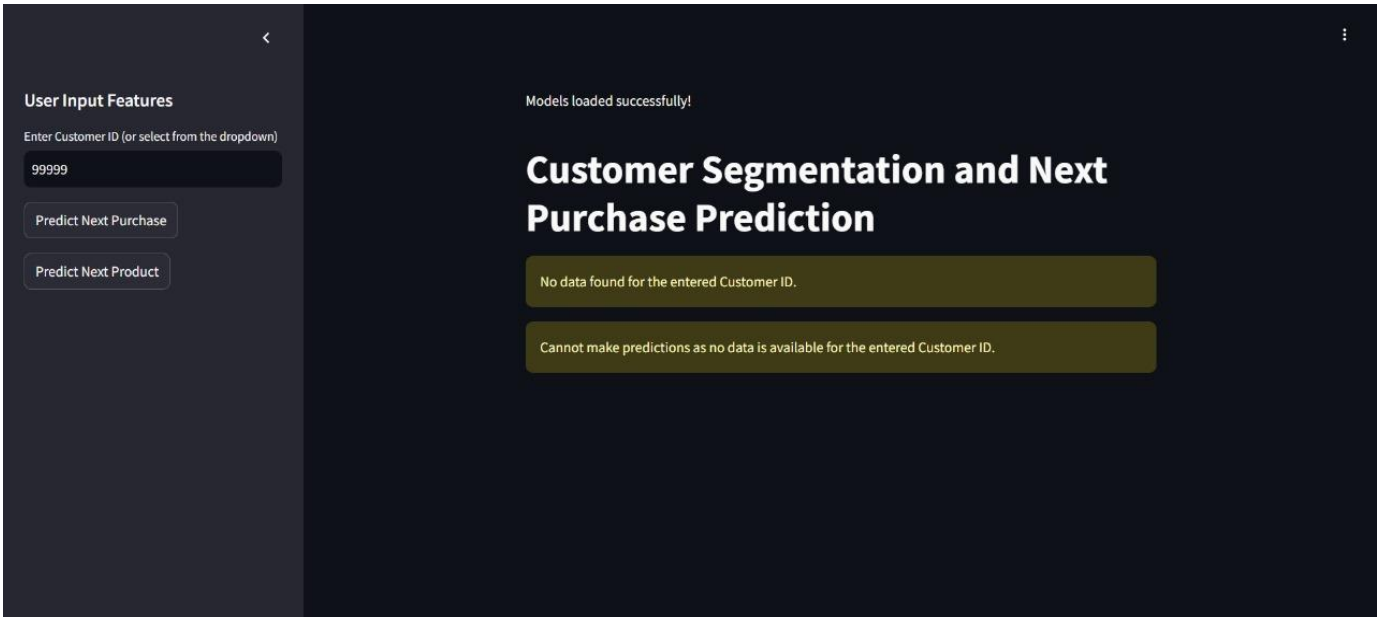
Cons:

- **Loss of Information:** Reducing dimensions may discard important data, affecting model performance.
- **Linear Assumption:** PCA assumes linear relationships, which limits its effectiveness for non-linear data.
- **Interpretability:** Principal components are hard to interpret, complicating the understanding of results.

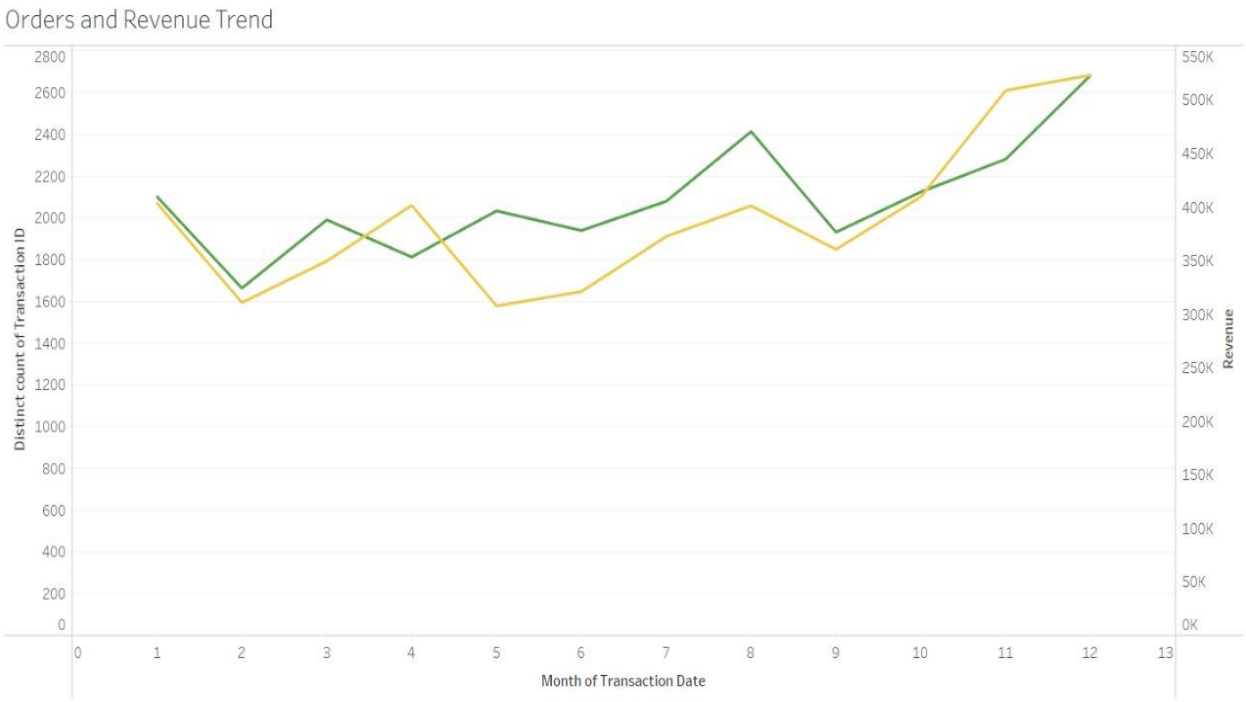
8.USER INTERFACE

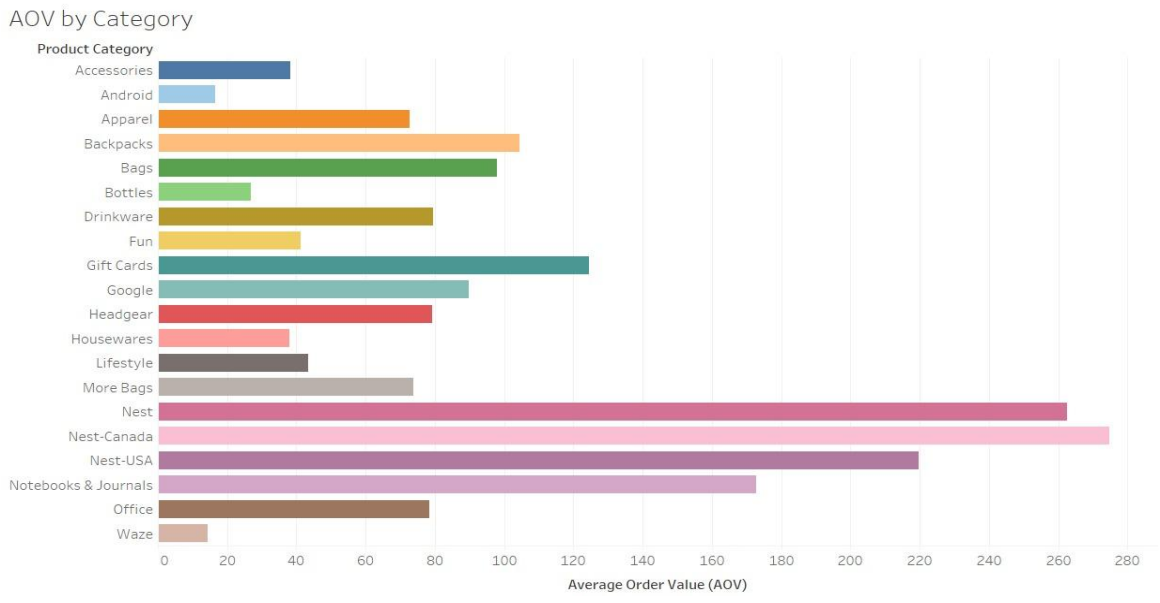


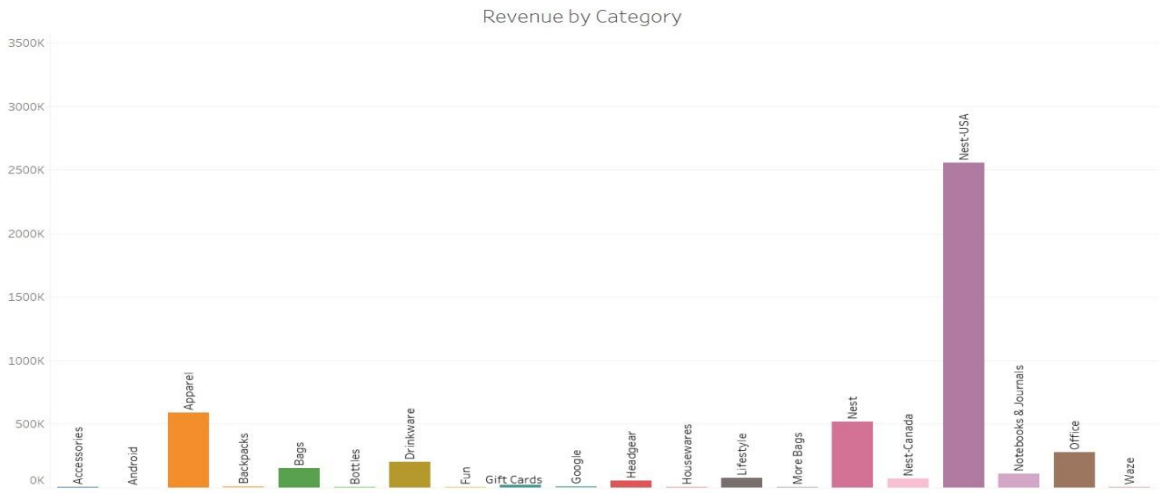


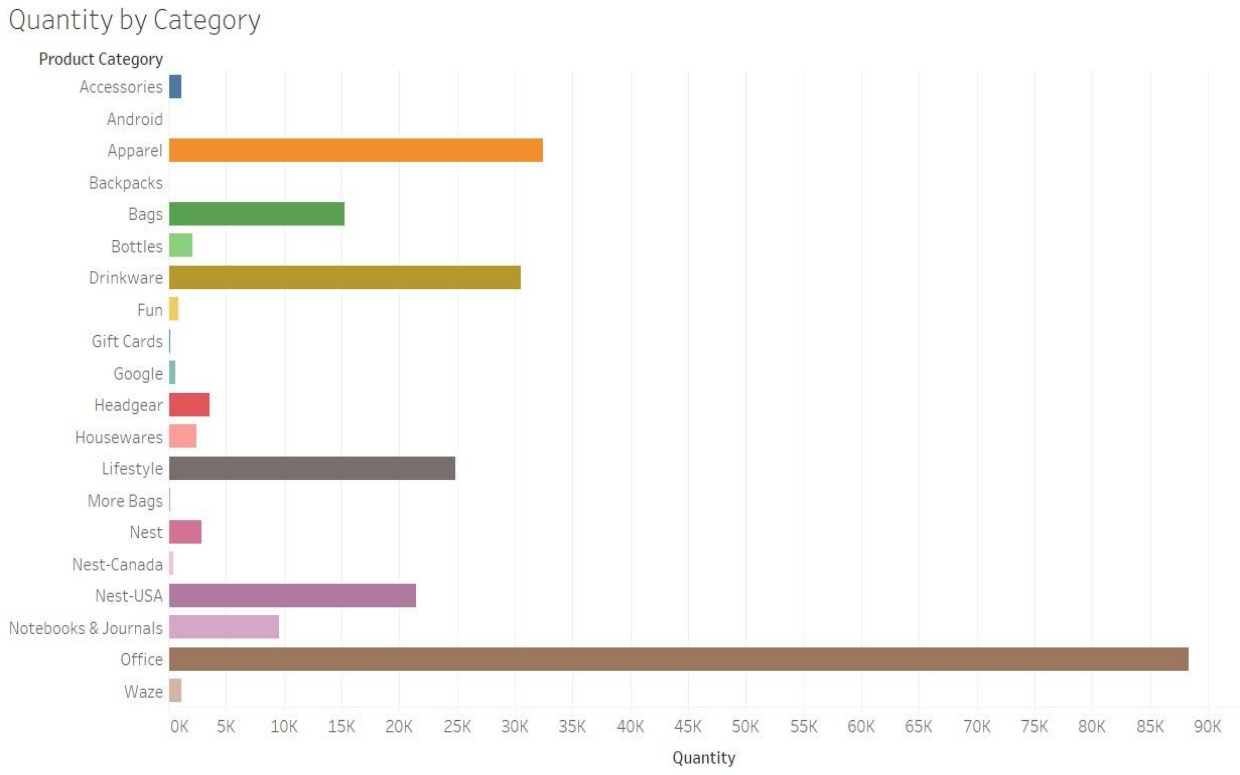

User Input Features
Enter Customer ID (or select from the dropdown)
99999
Predict Next Purchase
Predict Next Product
Models loaded successfully!
Customer Segmentation and Next Purchase Prediction
No data found for the entered Customer ID.
Cannot make predictions as no data is available for the entered Customer ID.

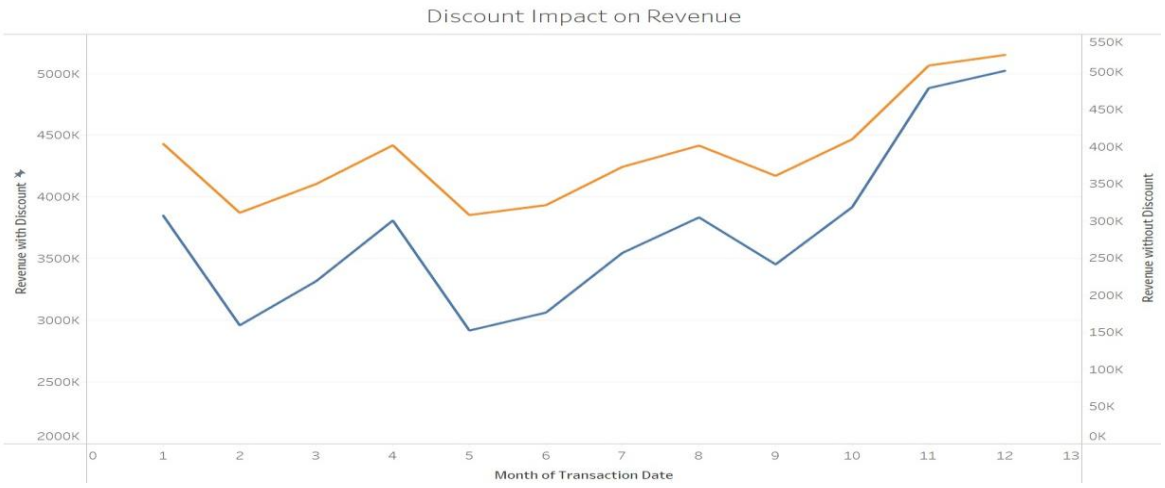
8.DATA VISUALIZATION AND REPRESENTATION






Revenue by Category
3500K
3000K
2500K
2000K
1500K
1000K
500K
0K
Accessories
Android
Apparel
Backpacks
Bags
Bottles
Drinkware
Fun
Gift Cards
Google
Headgear
Housewares
Lifestyle
More Bags
Nest
Nest-Canada
Nest-USA
Notebooks & Journals
Office
Waze


Quantity by Category
Product Category
Accessories
Android
Apparel
Backpacks
Bags
Bottles
Drinkware
Fun
Gift Cards
Google
Headgear
Housewares
Lifestyle
More Bags
Nest
Nest-Canada
Nest-USA
Notebooks & Journals
Office
Waze
0K
5K
10K
15K
20K
25K
30K
35K
40K
45K
50K
55K
60K
65K
70K
75K
80K
85K
90K
Quantity


Discount Impact on Revenue
Revenue with Discount
Revenue without Discount
Month of Transaction Date
5000K
4500K
4000K
3500K
3000K
2500K
2000K
550K
500K
450K
400K
350K
300K
250K
200K
150K
100K
50K
0K
0
1
2
3
4
5
6
7
8
9
10
11
12
13

Dashboard:-

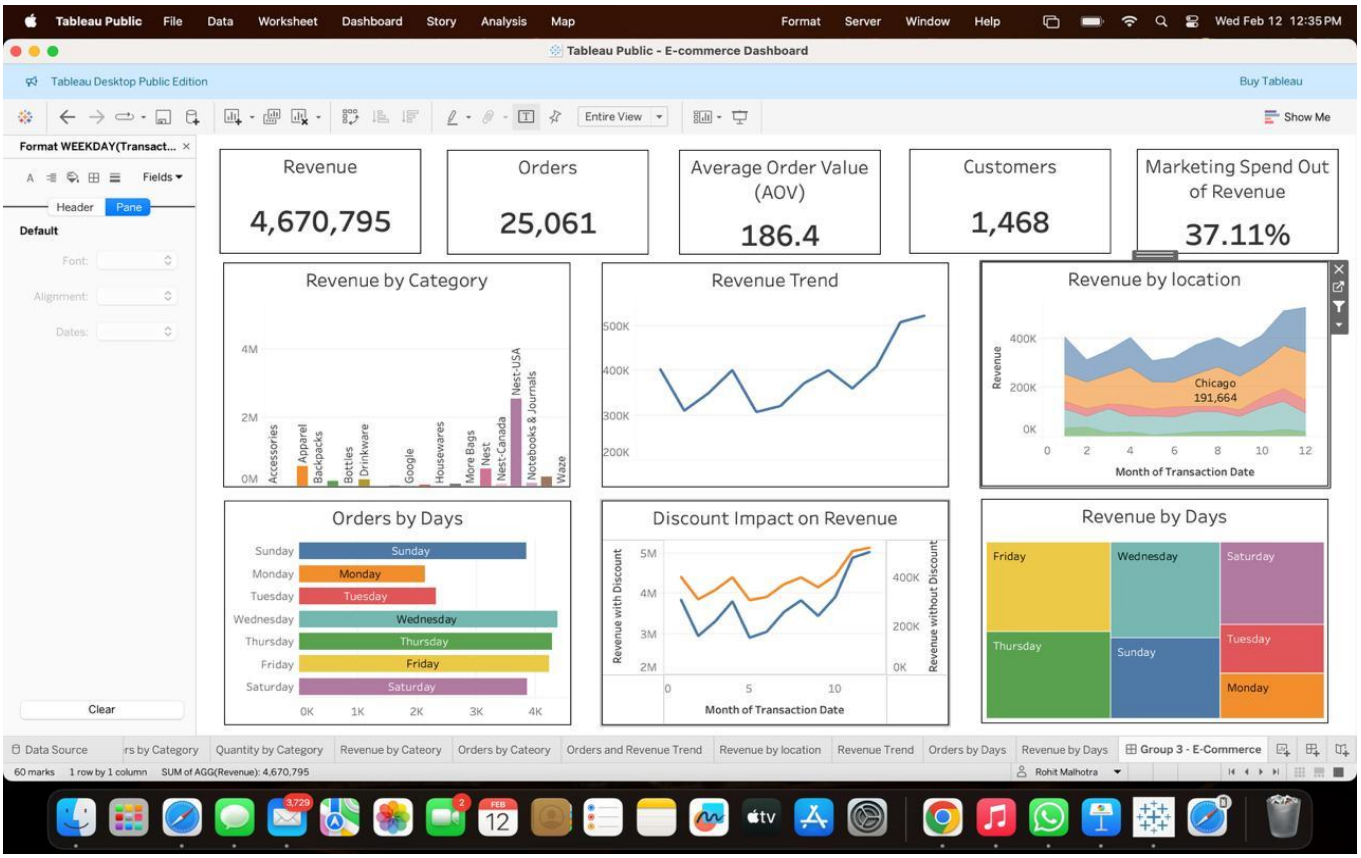


Tableau Public - E-commerce Dashboard
Revenue
4,670,795
Orders
25,061
Average Order Value (AOV)
186.4
Customers
1,468
Marketing Spend Out of Revenue
37.11%
Revenue by Category
Revenue Trend
Revenue by location
Chicago
191,664
Month of Transaction Date
Orders by Days
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Discount Impact on Revenue
Revenue with Discount
Revenue without Discount
Month of Transaction Date
Revenue by Days
Friday
Wednesday
Saturday
Thursday
Sunday
Tuesday
Monday

CONCLUSION AND FUTURE SCOPE

This project demonstrates the potential of machine learning in enhancing customer segmentation and predictive analytics. By leveraging Random Forest algorithms and integrating multiple datasets, we achieved reliable predictions for both future purchases and product recommendations.

Key Insights:

1. Customers with frequent transactions and higher monetary value are more likely to make future purchases.
2. Product recommendations are more accurate when recent purchasing behaviour is considered.
3. The Streamlit app provides an effective interface for real-time interaction with predictive models.

Future Enhancements:

1. Incorporate real-time data streaming for dynamic predictions.
2. Experiment with advanced algorithms like Gradient Boosting, XGBoost, and Neural Networks.
3. Perform hyperparameter tuning using Grid Search and Random Search to optimize model performance.
4. Deploy the application on cloud platforms like AWS or Azure for scalability.
5. Integrate additional data sources such as social media activity and customer feedback for holistic insights.