**Project Title:**

“Optimizing E-Commerce Strategy with Customer Segmentation and Predictive Analytics”

**Problem Statement**

E-commerce businesses need to improve customer retention, boost sales, and personalize marketing. However, without data-driven insights, decisions are often reactive and inefficient.

This project aims to:

1. **Segment customers** based on purchase behavior and demographics
2. **Predict future purchases or churn risk** using machine learning
3. **Provide actionable business insights** to improve marketing and product strategies

**Dataset Overview**

You’ll be using:

| **Table** | **Purpose** |
| --- | --- |
| users | Customer demographics |
| orders | Transaction history and status |
| products | Product details (category, brand, cost) |
| events (optional) | User interactions like views, adds to cart, etc. |
| inventory\_items | Pricing and inventory metadata |

**Key Questions to Answer**

* Who are the **most valuable customers**?
* Which customers are **likely to churn**?
* What **product categories** drive the most revenue?
* What is the **average customer lifecycle**?
* How do **demographics affect purchases**?

**Suggested Workflow**

**1. Data Exploration (EDA)**

* Join users and orders
* Analyze order volume over time
* Check repeat vs one-time customers
* Identify trends by age, location, product category

**2. Feature Engineering**

* Create RFM metrics (Recency, Frequency, Monetary)
* Calculate customer lifetime value (CLV)
* Extract product preferences by customer

**3. Segmentation (Unsupervised Learning)**

* Apply KMeans clustering on customer behavior features
* Profile each segment (e.g., bargain buyers, high spenders)

**4. Prediction (Supervised Learning)**

* Predict **repeat purchase likelihood** or **churn**
* Model type: Logistic Regression / Random Forest / XGBoost
* Target: Binary classification (will repurchase / will churn)

**5. Insights & Recommendations**

* Recommend personalized marketing strategies per segment
* Suggest products to promote or bundle based on trends
* Identify churn triggers

**Suggested Visualizations**

* RFM Heatmaps
* Customer segment distribution
* Purchase trend line charts
* Product-category performance bar charts

**Step-by-Step Project Plan (Multi-Table Dataset)**

**🔹 Step 1: Define the Business Question & Target**

**Objective:**  
Can we segment customers and predict who is likely to respond to targeted offers?

**Action Items:**

* Finalize your **target variable** (e.g., “repeat buyer” or “high spender” or proxy for "term deposit")
* Write your **problem statement** and **project scope**
* Identify who the stakeholders are (marketing, sales, etc.)

**🔹 Step 2: Understand the Data Landscape**

**Tables in your dataset:**

* users – customer demographics
* orders – order-level data
* order\_items – items within each order
* products – product-level info
* inventory\_items – metadata
* events – user interactions (clicks, views)
* distribution\_centers – logistics info

**Action Items:**

* Explore each table: column names, data types, nulls, value ranges
* Sketch a **data schema** (ERD) to understand how tables relate (via user\_id, order\_id, product\_id)

**🔹 Step 3: Design the Analytical Dataset**

You’ll need to **combine multiple tables** to get a customer-level dataset.

**Goal:** Build a clean, customer-centric table (1 row per user)

**Action Items:**

* Join users with orders (via user\_id)
* Join orders with order\_items and products to get category and item detail
* (Optional) Aggregate events by user to engineer engagement metrics

**🔹 Step 4: Feature Engineering**

**Goal:** Create meaningful customer-level features for segmentation and prediction

**Ideas:**

* total\_spent: Sum of sale\_price from orders
* avg\_order\_value
* num\_orders, last\_order\_date, days\_since\_last\_order
* most\_common\_category
* discount\_usage (based on price vs cost)
* return\_rate if returns exist
* active\_days: based on events

**🔹 Step 5: Target Label Creation**

**Goal:** Define who the “target customers” are for your model or segmentation

**Options:**

* Binary classification:
  + e.g., **“Repeat Buyer” = 1 if >1 order**
  + **“High Value Customer” = 1 if total\_spent > threshold**
* Or use clustering with no label (unsupervised)

**🔹 Step 6: Exploratory Data Analysis (EDA)**

**Goal:** Understand patterns and distributions

**Action Items:**

* Histograms: total spend, order frequency
* Bar plots: product categories, state distribution
* Correlation heatmaps for features
* Compare target groups vs non-target groups

**🔹 Step 7: Modeling**

**Classification (if predictive):**

* Split data (train/test)
* Test models: Logistic Regression, Random Forest, XGBoost
* Use metrics: Accuracy, F1-score, AUC-ROC
* Tune hyperparameters

**Segmentation (if clustering):**

* Scale features
* Use KMeans or DBSCAN
* Evaluate with Silhouette Score
* Label clusters and interpret them

**🔹 Step 8: Interpret & Recommend**

**Goal:** Translate model or segments into business value

**Action Items:**

* Identify:
  + High-value customers
  + Low-engagement customers
  + Product/category affinities
* Recommend marketing actions:
  + Who to target
  + What kind of offer works for each segment

**🔹 Step 9: Final Outputs**

**Deliverables:**

* Jupyter Notebook (clean, annotated)
* PowerPoint slides (for your capstone presentation)
* GitHub README with:
  + Problem statement
  + Dataset structure
  + Methodology
  + Findings & impact
* Optional: Power BI dashboard (even if basic)

**🔁 Recap of Core Steps (Simplified)**

1. **Frame the problem**
2. **Explore and join the data**
3. **Engineer features**
4. **Define your target**
5. **Perform EDA**
6. **Model or segment**
7. **Generate insights**
8. **Deliver your findings**

Table use casesTop of Form

**Order:**

* status is useful for filtering complete vs returned orders.
* created\_at vs delivered\_at can help calculate **delivery time**.
* num\_of\_item helps you derive **order size** and could feed into customer value features.
* If you joined orders + users + order\_items, then num\_of\_item is likely aggregated.

**Distribution center**

* You can use latitude and longitude for **mapping** fulfillment routes or measuring delivery distance.
* Join this table to inventory\_items or orders to analyze **shipping logistics by location**.
* distribution\_center\_geom can be used with geospatial functions in BigQuery or mapping tools.

**Events:**

* **Funnel analysis**: Track how users move from visit → cart → checkout → purchase.
* **Churn behavior**: Analyze repeated cancel events by user\_id or traffic source.
* **Marketing channel effectiveness**: Group by traffic\_source to assess conversion rates.
* **User behavior profiling**: Join with users table to identify segments with high intent or drop-off.

**Inventories item:**

* **Inventory Turnover**: Calculate time between created\_at and sold\_at to assess product velocity.
* **Profitability Analysis**: Compare cost vs. product\_retail\_price for margin estimation.
* **Unsold Inventory**: Filter rows with null sold\_at to monitor overstock or slow-moving items.
* **Distribution Efficiency**: Join with distribution\_centers to analyze regional inventory flow.
* **Category Sales**: Group by product\_category or product\_brand to see top performers.

**Order\_items:**

* **Order Fulfillment Analysis**: Time between created\_at, shipped\_at, and delivered\_at can help track fulfillment efficiency.
* **Return Rate Tracking**: Use returned\_at to compute product return rates.
* **Sales Revenue**: Aggregate sale\_price across status = 'Complete' to estimate revenue.
* **Cancellation Patterns**: Explore customer or product trends related to high cancellation.
* **Product Movement**: Join with inventories to connect products sold to specific inventory batches.

**Product\_df:**

* **Profit Margin Calculation**: Use retail\_price - cost to compute gross margin.
* **Category & Department Analysis**: Group products by category or department for inventory and marketing insights.
* **Inventory Management**: Join with inventories or order\_items to see sales and stock trends by product and distribution center.
* **Brand Performance**: Track top-selling brands and optimize sourcing strategies.
* **SKU-level Insights**: Use sku to uniquely identify products across sales, returns, and fulfillment.

**users:**

* **Customer Segmentation**: Use age, gender, traffic\_source, and location features for demographic clustering.
* **Marketing Attribution**: Analyze acquisition channels (traffic\_source) to understand effective marketing strategies.
* **Geo-Targeting**: Use latitude, longitude, and user\_geom for regional analysis and location-based personalization.
* **Lifetime Value**: Join with orders or order\_items to track customer value over time.
* **Retention Analysis**: Use created\_at to analyze cohort behavior and churn risk.

Bottom of Form