Major Project: Predictive Analysis and Customer Segmentation for an E-Commerce Platform.

Dataset Link: <https://www.kaggle.com/datasets/ineubytes/online-retail-ecommerce-dataset>

An online retail e-commerce dataset is a collection of structured data that records various aspects of online retail transactions and customer interactions with an e-commerce platform. Such datasets are commonly used for analyzing sales patterns, customer behavior, inventory management, and other aspects of e-commerce business operations.

These are the columns of dataset:

1. InvoiceNo: An identifier for each retail transaction.
2. StockCode: A code representing the product in the transaction.
3. Description: A description of the product.
4. Quantity: The quantity of the product purchased in the transaction.
5. InvoiceDate: The date and time when the transaction occurred.
6. UnitPrice: The price of a single unit of the product.
7. CustomerID: A unique identifier for each customer.
8. Country: The country where the customer is located or where the transaction took place.

Q/A for Major Project

1. What is the business model of the e-commerce platform you're working with?

Ans: The business model of the e-commerce platform I'm working with is:

B2C (Business-to-Consumer) - Online Retail:

The traditional model where businesses sell products directly to individual consumers through their websites.

1. What kind of data preprocessing and cleaning was required for the Online Retail I Dataset?

Ans: For the Online Retail I Dataset, the data preprocessing and cleaning involved:

Removing Null Values: Eliminating null or missing values from the dataset to ensure data completeness and accuracy.

Duplicate Data: Identifying and removing duplicate records within the dataset to prevent redundancy and maintain data quality.

1. How did you visualize and interpret the data distributions and relationships using Power BI/Tableau?

Ans: In Power BI/Tableau, I visualized sales trends over time using line charts or area charts, with the X-axis representing date or month and the Y-axis showing total sales amount or quantity sold. This helped in identifying sales patterns and seasonal variations. For customer segmentation analysis, I used bar charts to display the distribution of customers among segments created based on behavior or attributes, with the X-axis indicating segments and the Y-axis showing either customer counts or total sales amounts. These visualizations facilitated better decision-making and targeted marketing strategies.

1. What new features did you engineer from the existing dataset and why?

Ans: Several new features were engineered from the existing dataset to enhance its analytical potential. First, the "Total Number of Unique Products Purchased by Each Customer" was calculated, offering insights into customer behavior and preferences. Second, the "Total Price" for each item in a transaction was derived from the quantity and unit price, aiding in financial analysis and revenue assessment. Additionally, the "Invoice Date" column was split into day, month, year, and quarter components, enabling granular time-based analysis, seasonal trend identification, and informed decision-making regarding marketing, sales, and inventory management.

1. Which regression models did you test for predicting the annual spending of a customer?

Ans: I have used Linear Regression and Random Forest Regression models to predict the annual spending of a customer.

1. What metrics did you use to evaluate the performance of the predictive models?

Ans: I’ve used metrics like Mean squared error, Root mean squared error and mean absolute error to evaluate the performance of the predictive models.

1. What did you learn about the data science project lifecycle throughout this project?

Ans: Throughout this data science project, I've gained valuable insights into the entire project lifecycle. It all began with a clear understanding of the problem we aimed to solve – deciphering customer behavior and sales drivers in an e-commerce context. We sourced our data from Kaggle, a treasure trove of datasets, setting the stage for our analysis.

With data in hand, I delved into data cleaning and preprocessing, ensuring our dataset was spotless and ready for exploration. The next phase involved donning my detective hat, conducting exploratory data analysis (EDA) to unveil hidden patterns and glean meaningful insights. To supercharge our models, I even engineered new features from existing data.

This model efforts included deploying techniques like linear regression for sales predictions and k-means clustering to group customers with similar behaviors. To communicate our discoveries effectively, I turned to visualization tools like Power BI. Rigorous model evaluation, including cross-validation, was crucial to ensure the reliability of our findings.