Customer Lifetime Value (CLV) Prediction Project

# Abstract

This project focuses on predicting the Customer Lifetime Value (CLV) of customers based on their historical purchase behavior. Customer Lifetime Value is a key metric for businesses as it helps identify the total expected revenue a company can generate from a customer over the entire relationship. By leveraging advanced machine learning techniques, this project empowers businesses to identify and prioritize high-value customers, design effective targeted marketing campaigns, and optimize resource allocation.  
  
The solution integrates data from multiple sources, including customer, order, and payment data. Data preprocessing was performed using Excel Power Query to clean, merge, and prepare the data. This was followed by building and training a predictive model in Python using libraries like Scikit-learn and XGBoost. The model outputs predicted LTV values and segments customers into meaningful categories to help the business take action.

# Tools Used

1. Microsoft Excel (Power Query) - For initial data cleaning, merging, and creating customer-level features like Recency, Frequency, and Monetary (RFM).  
2. Python (Pandas, NumPy, Scikit-learn, XGBoost, Matplotlib, Seaborn) - For data modeling, visualization, and building the predictive machine learning model.  
3. Jupyter Notebook - For step-by-step analysis, model training, and documentation.  
4. CSV/Excel Output Files - For delivering final customer segmentation and predictions.

# Steps Involved in Building the Project

1. \*\*Data Collection\*\*: Gathered nine datasets containing customer information, orders, payments, products, reviews, and geographic data.  
2. \*\*Data Preprocessing\*\*: Cleaned and merged datasets using Excel Power Query to form a unified dataset at the customer level.  
 - Removed invalid or canceled orders.  
 - Created transaction tables and calculated total revenue per order.  
3. \*\*Feature Engineering\*\*: Built Recency, Frequency, and Monetary (RFM) features, along with derived metrics such as Average Order Value (AOV), tenure, and purchase rate to better understand customer behavior.  
4. \*\*Model Training\*\*: Implemented a regression model using XGBoost to predict future customer spend.  
5. \*\*Model Evaluation\*\*: Validated the model using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to assess accuracy.  
6. \*\*Customer Segmentation\*\*: Grouped customers into Low, Medium, High, and Very High value segments based on predicted LTV.  
7. \*\*Final Deliverables\*\*: Exported the results as a CSV file and created visualizations for feature importance and customer distribution.

# Conclusion

The Customer Lifetime Value Prediction project successfully demonstrated how businesses can utilize historical purchase data to make informed decisions. By integrating multiple datasets, cleaning and preparing them in Excel, and applying machine learning models in Python, we were able to predict the expected revenue from each customer with a high degree of accuracy.  
  
The segmentation of customers based on their predicted LTV allows businesses to:  
- Identify and nurture high-value customers through loyalty programs.  
- Allocate marketing budgets efficiently by targeting customers most likely to generate long-term revenue.  
- Design retention strategies for at-risk or low-value customers.  
- Improve overall customer experience by personalizing marketing efforts.  
  
This project not only produced a working model but also established a repeatable, scalable pipeline for future analytics initiatives. As businesses continue to collect more data, this framework can be expanded to include additional features such as product preferences, geographical analysis, and review sentiment. Ultimately, this solution empowers organizations to transform raw data into actionable insights that directly impact profitability and customer satisfaction.