iykra Data Fellowship

Building Customer Segmentation Pipeline

for an Effective Personalized Marketing

prepared by:
Group 2

Data Fellowship Program
Batch 12

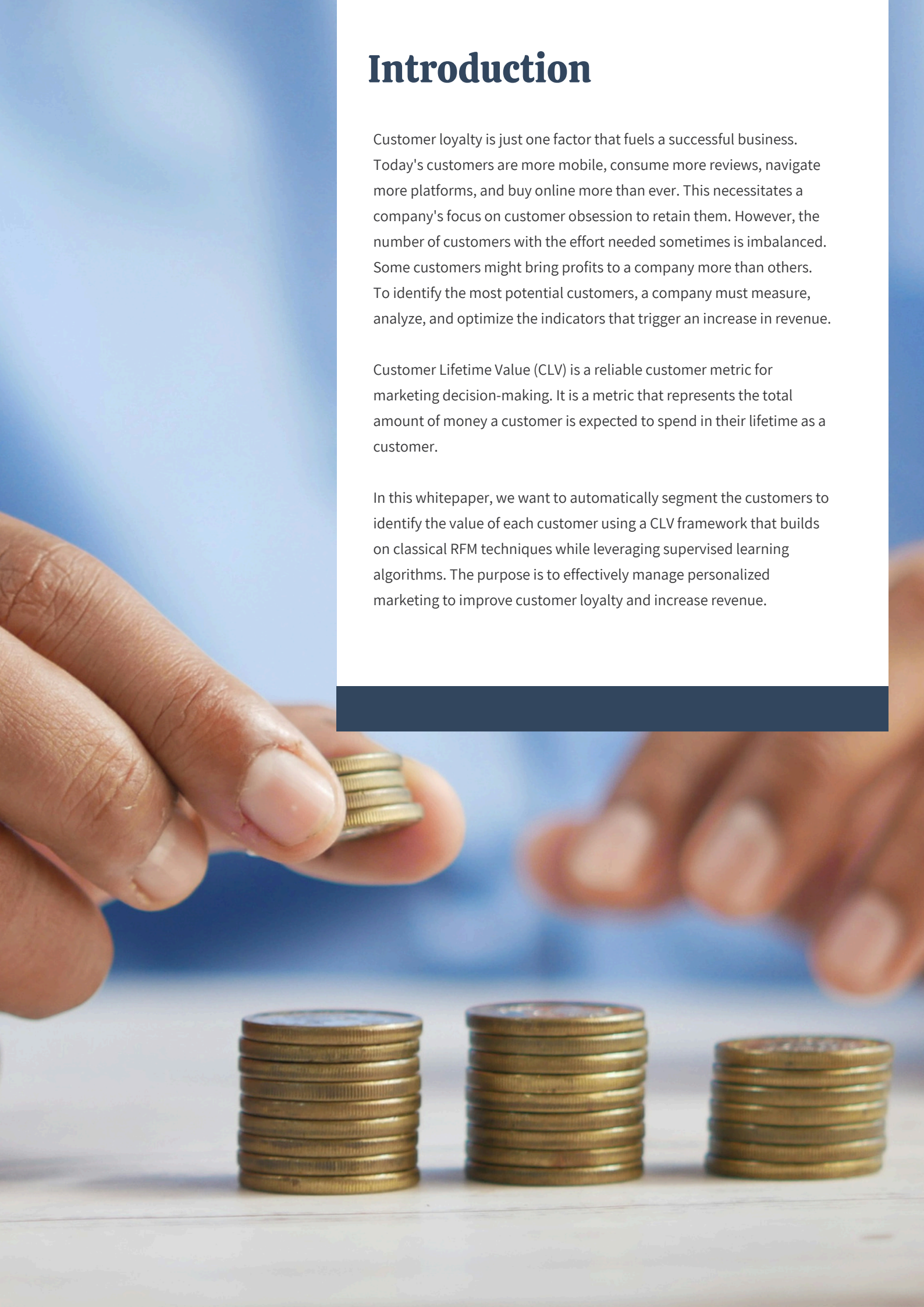
Table of Contents

Introduction

Customer loyalty is just one factor that fuels a successful business. Today's customers are more mobile, consume more reviews, navigate more platforms, and buy online more than ever. This necessitates a company's focus on customer obsession to retain them. However, the number of customers with the effort needed sometimes is imbalanced. Some customers might bring profits to a company more than others. To identify the most potential customers, a company must measure, analyze, and optimize the indicators that trigger an increase in revenue.

Customer Lifetime Value (CLV) is a reliable customer metric for marketing decision-making. It is a metric that represents the total amount of money a customer is expected to spend in their lifetime as a customer.

In this whitepaper, we want to automatically segment the customers to identify the value of each customer using a CLV framework that builds on classical RFM techniques while leveraging supervised learning algorithms. The purpose is to effectively manage personalized marketing to improve customer loyalty and increase revenue.



Problem and Goal

This research design investigates a dataset obtained from the retail sales channel of a consumer. Customer analytics techniques including Recency-Frequency-Monetary (RFM). An e-commerce company has a large number of customers to serve and wants to effectively manage personalized marketing to improve customer loyalty and decrease churning rate.

The introduction of e-commerce has benefitted both consumers and firms alike, with consumers benefitting proportionately more, largely due to the increased competition and product differentiation faced by suppliers in online distribution channels.

To compete for consumers' share of wallet, it is therefore essential to pay attention to customer databases and analytics important components of customer relationship management (CRM). In the domain of CRM, targeted marketing, built upon the foundations of data analytics, is recognized as a valuable tool, as marketing paradigm is becoming increasingly customer-focused and unsolicited marketing is often costly and ineffective due to low response rates. The use of statistical techniques to analyze customer data to support marketing decisions is an important element of database marketing (DBM)

New customers can be 5x to 25x more expensive than retaining customers. The probability of converting an existing customer is between 60-70%.

+67%

A 5% increase in retention produces a 25% increase in profit

Existing customers spend 67% more on average than new customers. It's common business knowledge that it's cheaper to retain a current customer than it is to attract a new one. Acquiring new customers costs five to 10 times more than selling to a current customer, and current customers spend 67 percent more on average than those new to your business, according to BIA Advisory.

Customer Lifetime Value

There are two ways to look for the customer lifetime value: the **historical CLV analysis technique** (how much each existing customer has spent in the business) and by **predictive CLV technique** (how much customers could spend in the business).

The **descriptive model** calculates CLV from the historical data of existing customers and identifies behavioral patterns of each consumer's group through simple manual analysis. It's helpful to understand what the customer brought as an initial indicator of CLV for potential decisions.

The **predictive model** uses historical data to forecast customers' buying behavior using regression or machine learning models. It can take into account customer acquisition costs, average purchase frequency rate, and more to give realistic predictions. Hence, this can act as supportive data to make a more effective decision. However, this model is more complex and requires comprehensive advanced analytics capabilities.

One popular technique to segment customers is by looking at the purchasing behavior, namely **recency**, **frequency**, and **monetary** (RFM). It calculates recent customers' last purchases, how frequently they make purchases and the profit they bring to business.

Customer Lifetime Value

$$CLV = [\text{Customer Value} \times \text{Average Customer Lifespan}]$$

Customer Value

$$CLV = [\text{Avg. Purchase Value} \times \text{Avg. Purchase Frequency Rate}]$$

25-95%

profit increase from 5% customer retention increase

The concept of the Pareto principle, also known as the 80/20 rule, perfectly captures this idea. It states that **80% of the profits come from 20% of your customers**. These high-value customers are cost-effective and generate significant revenue.

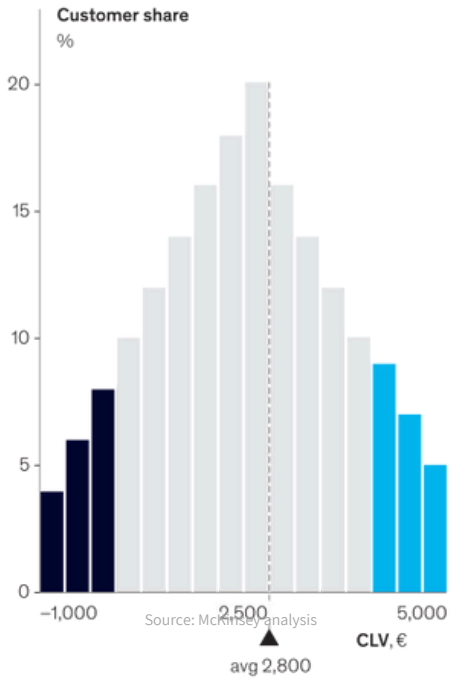
CLV Illustration

Customer lifetime value (CLV) is fundamental for the elaboration of customer profiles and marketing measures.

Low CLV
Illustrative customers
- Aged 40 to 50
- $-500 to $150

Customer types
- One-off customer: Follow-up purchase unlikely
- Bargain hunter: Buys variety of brands when on offer

Actions
- Reduce short-term marketing investment
- Perform long-term monitoring and potentially reactivate with targeted campaigns



High CLV
Illustrative consumers
- Aged 25 to 35
- CLV > $3,500

Customer types
- Loyal customer: regular shopper in same channel, often in same brand
- Life event customer: large purchase volume within brief period (eg, home decor)

Actions
- Make personalized offers for follow-up purchases
- Reward loyalty through additional service offers
- Identify life events in advance by drawing on external data

RFM Segmentation

Each value of RFM corresponds to some key customer trait. The frequency and monetary value affect a customer's lifetime value, while recency affects retention, a measure of engagement.

RECENCY

The freshness of the customer activity, be it purchases or visits

E.g. Time since last order or last engaged with the product

FREQUENCY

The frequency of the customer transactions or visits

E.g. Total number of transactions or average time between transactions/ engaged visits

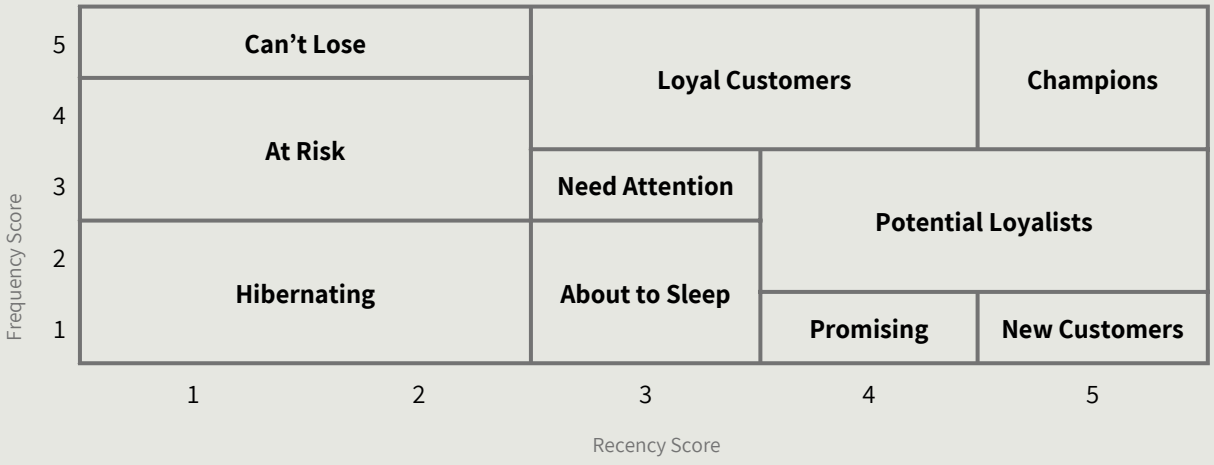
MONETARY

The intention of customer to spend or purchasing power of customer

E.g. Total or average transactions value

RFM factors illustrate these facts:

- the more recent the purchase, the more responsive the customer is to promotions
- the more frequently the customer buys, the more engaged and satisfied they are
- monetary value differentiates heavy spenders from low-value purchasers



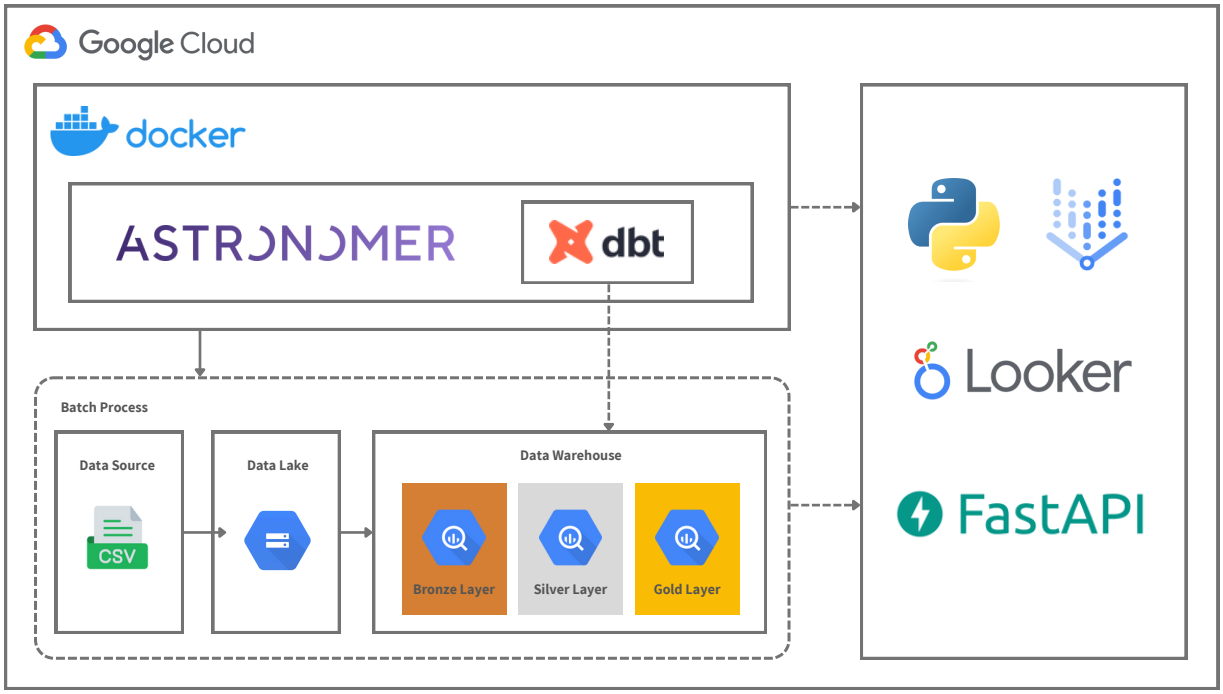
Example

In the dataset, we calculated the RFM Score and labeled it as follows.

CustomerID	Recency	Frequency	Monetary	R_Score	F_Score	M_Score
12348	-157	28	1487.24	2	2	4
18282	-35	7	100.21	4	2	3
18283	-4	447	1120.67	1	5	4

Data Pipeline

Data pipeline is an essential systematic process that orchestrates the movement, transformation, processing, and loading of data from one or multiple sources into an organized repository that enables analysis and visualization. The following diagram illustrates the architecture of the automatic customer segmentation prediction process. This section provides a conceptual overview of the data pipeline designed to predict CLV, leveraging modern data processing and analytics tools on the Google Cloud Platform (GCP).



The data pipeline for Customer Lifetime Value (CLV) prediction is designed to provide a seamless flow from data ingestion to model deployment using Google Cloud Platform (GCP) services. Initially, raw data is collected from CSV files, which are stored in Google Cloud Storage (GCS). This raw data, including customer interactions and transaction histories, is then ingested into a data lake, serving as a centralized repository for unprocessed data.

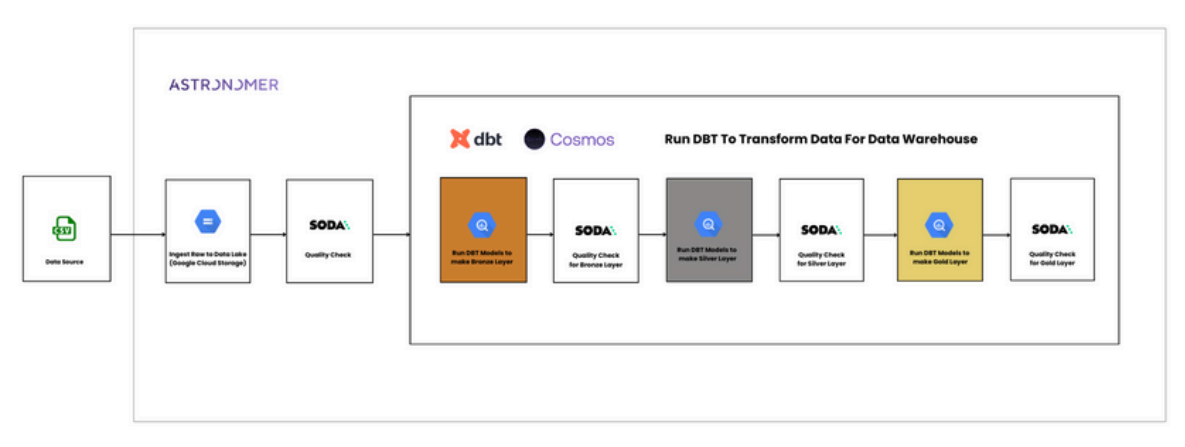
Following ingestion, data transformation and processing are orchestrated using Astronomer (Airflow) and dbt (Data Build Tool). These tools facilitate the cleaning, enrichment, and structuring of data within BigQuery, creating three distinct layers: Bronze (raw data), Silver (cleaned and enriched data), and Gold (highly processed data).

This structured data is then utilized for training machine learning models in Python using Google Colab, ensuring accurate and reliable CLV predictions.

Finally, the processed data and CLV predictions are visualized through Looker, providing intuitive dashboards and reports for stakeholders. To make the predictions accessible across various applications, FastAPI is employed to create APIs that serve the predicted CLV data. This comprehensive pipeline leverages the scalability and efficiency of GCP services, enabling businesses to derive actionable insights and make informed decisions to enhance customer engagement and maximize revenue.

Architecture Objective

Data Integration and Collection	Seamlessly gather data from various sources into Google Cloud Storage (GCS) using Docker for consistent deployment.
Data Cleaning and Transformation	Utilize Astronomer (Airflow) and dbt to clean, transform, and organize data in BigQuery across Bronze, Silver, and Gold layers for analysis.
Scalability	Automation and Repeatability: Use Airflow to automate workflows, ensuring consistent and repeatable data processing.
Deployment and Monitoring	Deploy applications with Docker and Astronomer, and continuously monitor for issues to maintain pipeline health.
Business Insights and Reporting	Employ Looker to visualize data and generate insightful reports, aiding in data-driven decision-making.



The data engineering project involves creating a robust and scalable architecture to support data collection, storage, processing, and analytics. This architecture will ensure efficient data management, real-time analytics, and support for machine learning models like CLV and RFM analysis.

Implementing a well-designed architecture for a data engineering project offers numerous benefits, particularly in terms of cost management. These benefits ensure that the project not only meets its technical and business objectives but also does so in a cost-efficient manner.

Cost

Google Cloud Storage

$0.023

PER GB PER MONTH

GCP VM Instances

$135.45

PER MONTH
for 4 vCPU + 16 GB memory + 30G persistent disk

Total Gross

$135.472

PER MONTH

Data Transformation

The data consists of 541909 records in 8 columns with description as follows.

Column	Description
InvoiceNo	a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation
StockCode	a 5-digit integral number uniquely assigned to each distinct product
Description	product name
Quantity	the quantities of each product (item) per transaction
InvoiceDate	the day and time when each transaction was generated
UnitPrice	product price per unit
CustomerID	a 5-digit integral number uniquely assigned to each customer
Country	the name of the country where each customer resides

Data Transformation and Processing

Raw Data:

- Drop rows with missing values in CustomerID.
- Correct data types (e.g., CustomerID from string to int).
- Extract components from InvoiceDate to create new columns: year, month, quarter, day of the week, and week of the year.
- Replace UnitPrice for each StockCode with its mode.

DBT:

- Processing phase where transformations are applied.

Preprocessed Data:

- Calculate TotalPrice by multiplying Quantity and UnitPrice.
- Remove rows where UnitPrice is zero or Quantity is negative.

Data Preprocessing

Data Cleaning and Feature Selection

Handling Missing Values:
Missing values in the dataset were identified and handled appropriately. Numerical missing values were imputed using the median, ensuring that extreme values did not overly influence the imputation. Categorical missing values were imputed using the mode, maintaining consistency in categorical distributions.

Feature Selection:
Features relevant to predicting Future CLV were selected based on domain knowledge and statistical methods. Key numerical features included Recency, Frequency, Density, Monetary, R_Score, F_Score, D_Score, M_Score, RFM_Score, AOV, profit_margin, CLV, FutureRevenue, Recency_Frequency, Recency_Monetary, Frequency_Monetary, and AvgTransactionValue. Categorical features included Segment and RF_Segment.

Feature Engineering

Creating New Features:
Composite scores such as Recency_Frequency, Recency_Monetary, and Frequency_Monetary were created to capture interactions between individual RFM components. These composite scores provided additional predictive power by highlighting customer behaviors.

Transforming Target Variable:
The continuous target variable, Future CLV, was transformed into categorical segments (Low, Mid, High) using quantiles. This transformation facilitated classification modeling and enabled a better understanding of customer value segmentation.

Encoding and Scaling

Encoding Categorical Features:
Categorical features were encoded using One-Hot Encoding and Label Encoding. One-Hot Encoding was used for features with nominal categories, while Label Encoding was applied to ordinal categories, ensuring that categorical data was appropriately transformed for modeling.

Scaling Numerical Features:
Numerical features were scaled using StandardScaler without centering (with_mean=False), ensuring that all features had equal variance and contributed equally to the model training process.

Machine Learning Model

In predicting Customer Lifetime Value (CLV), a range of regression models was explored to capture the complex relationships between various predictors and future CLV estimates. The models selected included Linear Regression, Random Forest Regressor, Decision Tree Regressor, and Support Vector Regressor (SVR). Each model underwent rigorous hyperparameter tuning using GridSearchCV to optimize performance metrics such as Mean Squared Error (MSE), R-squared (R2), Mean Absolute Percentage Error (MAPE), and cross-validation scores. This approach ensured that the models were fine-tuned to effectively predict and quantify the expected CLV over time. The following are the results of the machine learning model that has been created

Model	Accuracy	Evaluation			
		Precision	Recall	f1-score	support
XGBoost	Training set: 0.945 Test set: 0.947	0: 0.93 1: 1.00 2: 0.88	0: 0.92 1: 1.00 2: 0.89	0: 0.92 1: 1.00 2: 0.88	0: 345 1: 429 2: 234
Support Vector Classifier	Train set: 0.651 Test set: 0.646	0: 0.77 1: 0.61 2: 0.50	0: 0.70 1: 0.87 2: 0.18	0: 0.73 1: 0.72 2: 0.26	0: 345 1: 429 2: 234
Logistic Regression	Train set: 0.842 Test set: 0.821	0: 0.93 1: 0.81 2: 0.77	0: 0.88 1: 0.98 2: 0.54	0: 0.90 1: 0.89 2: 0.63	0: 345 1: 429 2: 234
Random Forest	Train set: 0.932 Test set: 0.930	0: 0.92 1: 0.99 2: 0.85	0: 0.90 1: 1.00 2: 0.85	0: 0.91 1: 0.99 2: 0.85	0: 345 1: 429 2: 234
Gradient Boosting	Train set: 0.945 Test set: 0.948	0: 0.94 1: 1.00 2: 0.87	0: 0.90 1: 1.00 2: 0.91	0: 0.92 1: 1.00 2: 0.89	0: 345 1: 429 2: 234

Ensemble modeling was also employed to further enhance prediction accuracy and robustness. Specifically, a Voting Regressor combined the strengths of Random Forest and SVR models. This ensemble approach leveraged the diversity of predictions from each base model to provide more reliable estimates of CLV. Evaluation of the ensemble model included comprehensive metrics like MSE, R2, MAPE, and cross-validation scores, demonstrating its superior performance compared to individual models alone.

For CLV segmentation and classification tasks, a suite of classification models was evaluated, including Logistic Regression, Random Forest Classifier, Decision Tree Classifier, Support Vector Classifier, and Gradient Boosting Classifier. These models were trained and evaluated on accuracy, cross-validation accuracy, classification reports, and confusion matrices to effectively segment customers based on their CLV profiles. The insights derived from these models not only enhanced segmentation accuracy but also provided actionable intelligence for targeted marketing and retention strategies tailored to different customer segments based on their CLV characteristics.

This modeling framework culminated in the implementation of advanced techniques such as XGBoost, a powerful gradient boosting algorithm known for its ability to handle large datasets and capture intricate patterns in data. Hyperparameter tuning using GridSearchCV further optimized XGBoost's performance, making it an effective tool for refining CLV predictions and supporting strategic decision-making within business operations.

Analysis Model

Regression Model Performance

Comparison of Regression Models: The performance of each regression model was compared based on MSE, R2, MAPE, and cross-validation scores. Random Forest exhibited strong performance with high accuracy and low error rates. The ensemble model outperformed individual models, demonstrating the benefits of combining Random Forest and SVR.

Key Findings:

- Random Forest achieved high predictive accuracy.
- The ensemble model further improved performance, indicating the effectiveness of model combination.

Classification Model Performance

Comparison of Classification Models: Classification models were compared based on accuracy, cross-validation accuracy, classification reports, and confusion matrices. Gradient Boosting Classifier and XGBoost showed the highest accuracy and balanced performance across different CLV segments.

Key Findings:

- Gradient Boosting Classifier and XGBoost were most effective in classifying Future CLV segments.
- Confusion matrix analysis revealed balanced predictions across all segments.

Business Implications

Customer Segmentation: Classifying customers into different CLV segments provides valuable insights for targeted marketing strategies. High-value customers can be prioritized, while strategies can be developed to nurture mid and low-value segments.

- Predictive Insights: The predictive models offer a robust framework for forecasting future customer value. Businesses can leverage these insights to make informed decisions on customer retention and resource allocation, ultimately driving growth and profitability.

Conclusion

The process of predicting Future CLV using various regression and classification models. Ensemble models and XGBoost showed strong predictive capabilities, providing actionable insights for customer segmentation.

Future Work:
Future research could explore additional features and alternative modeling techniques to enhance predictive accuracy. Implementing these models in real-time systems could provide ongoing insights and further optimize customer value strategies.

Our Team

Alivia Rahma Sakina
Data Analyst

Alya Mutiara Firdausyi
Data Scientist

Fathurrahman Hernanda Khasan
Data Engineer

M Zakie Arfiansyah
Data Scientist

Ridho Alfayet Umar
Analytics Engineer

References

Mukherjee, S., & Mukherjee, S. (2024, May 2). Customer Lifetime Value: What is it and how to calculate. CleverTap - All-in One Customer Engagement Platform. https://clevertap.com/blog/customer-lifetime-value/

Vanhaesebroeck, J. (2024, February 28). Learn your customer lifetime value first, if you want to improve loyalty. StriveCloud. https://strivecloud.io/blog/cltv-improves-loyalty/

Customer lifetime value: The customer compass. (2021, October 27). McKinsey & Company. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/customer-lifetime-value-the-customer-compass

Makhija, P., & Makhija, P. (2024, May 2). RFM analysis for Customer Segmentation - CleverTap. CleverTap - All-in One Customer Engagement Platform. https://clevertap.com/blog/rfm-analysis/