IYKRA
Data Fellowship Program

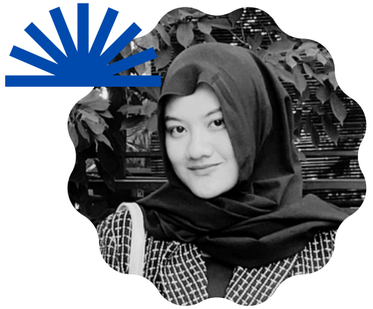
Building Customer Segmentation Pipeline for an Effective Personalized Marketing

By Group 2

14 June 2024

Our Team

Data Analyst

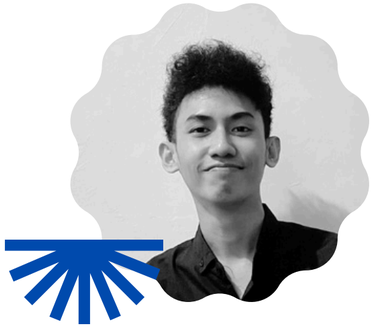


ALIVIA RAHMA SAKINA (ALIVIA)

Data Scientist

ALYA MUTIARA FIRDAUSYI (ALYA)

Data Engineer



FATHURRAHMAN HERNANDA KHASAN (FAFA)

Analyticst Engineer

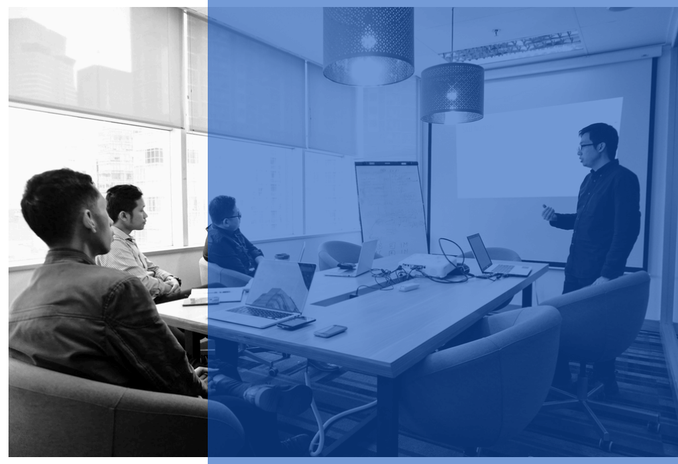
RIDHO ALFAYET UMAR (RIDHO)

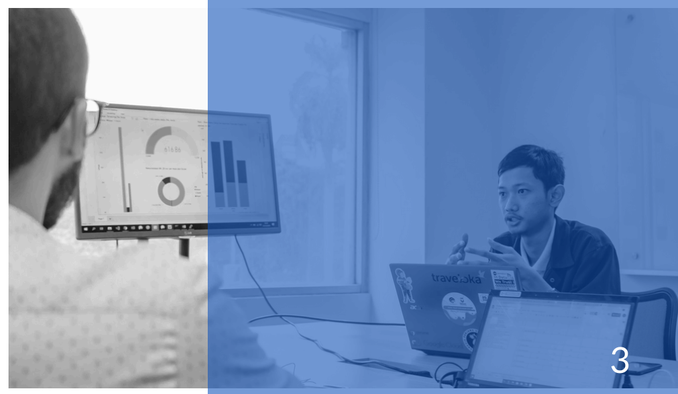
Data Scientist

M ZAKIE ARFIANSYAH (ARFI)

Let's Connect On LinkedIn!

Table of Contents





Data Fellowship

Project Objective

Problem & Goal

Customers are the puzzle pieces of any successful business.

While all contribute to bringing profits, not every customer fits the picture perfectly. Some might bring more value to a company than others.

Is there any way to find the best pieces--the ones that complete the picture perfectly?

Problem & Goal

Problem Statement

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**.

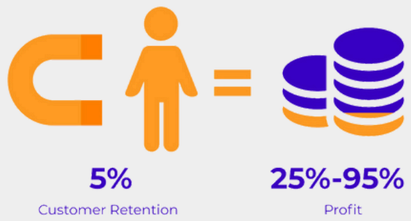
Goal

Create a data pipeline from existing data that can be utilized to **segment the current customers' value** and predict the segment for the next period.

Why we need to do this?

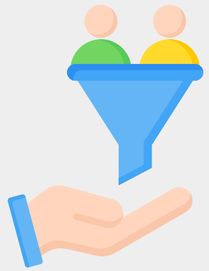
Pareto Principle

80% of the profits come from 20% of your customers.



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

New customers can be 5x to 25x **more expensive** than retaining customers.



The probability of converting an **existing customer** is between 60-70%.

Existing customers spend **67% more** on average than new customers.

Objective

Predict Customer Lifetime Value

Develop a machine learning model that can predict the future value a customer will bring to the company.

Automate the Prediction Process

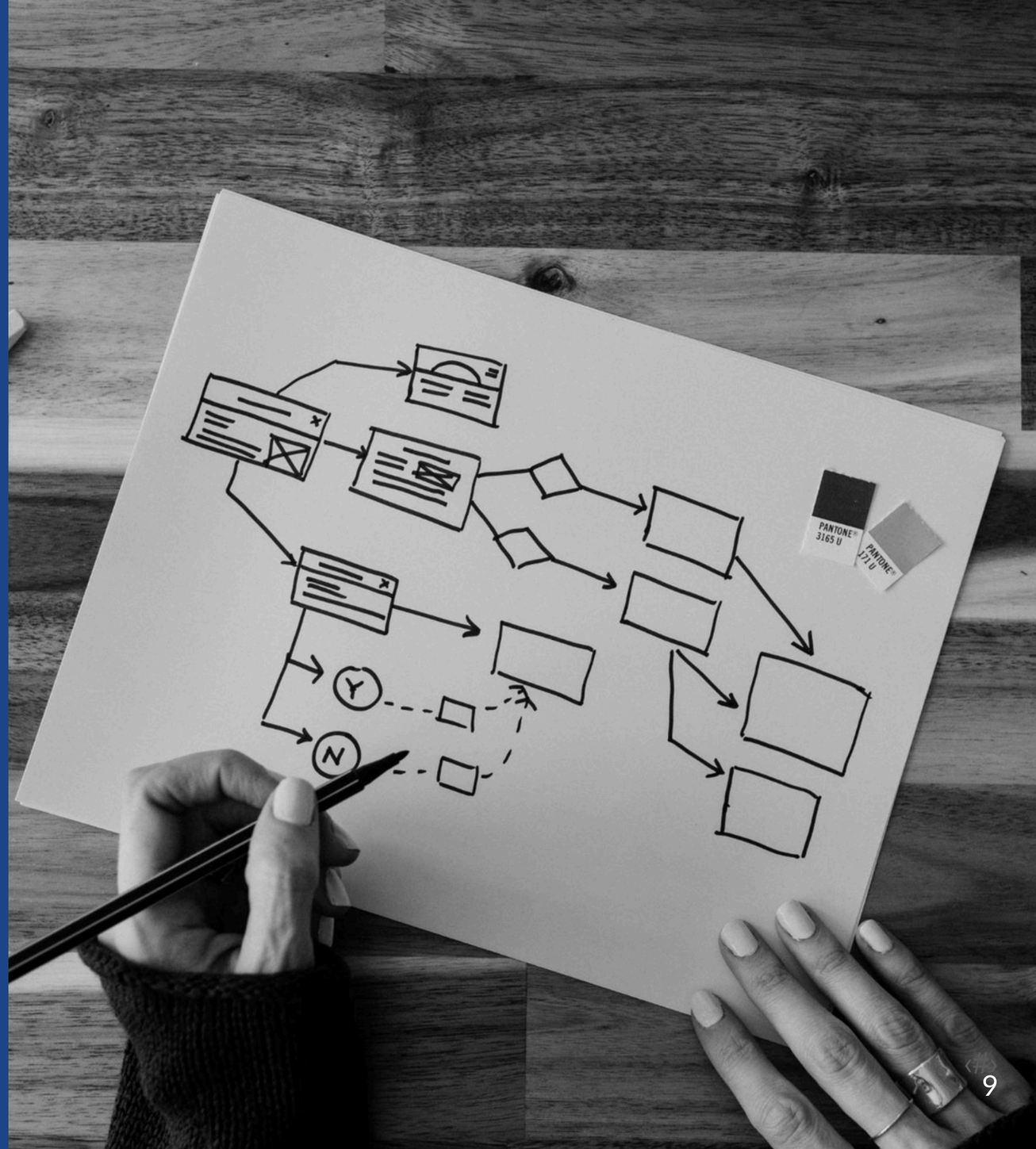
Create an end-to-end automated pipeline that handles data collection, preprocessing, model training, and prediction.

Dashboard for Data Visualization

Build a dashboard to improve data visualization and analysis for better marketing decisions.

Data Fellowship

Proposed Architecture



Architecture Objective

01 Data Integration and Collection

02 Data Cleaning and Transformation

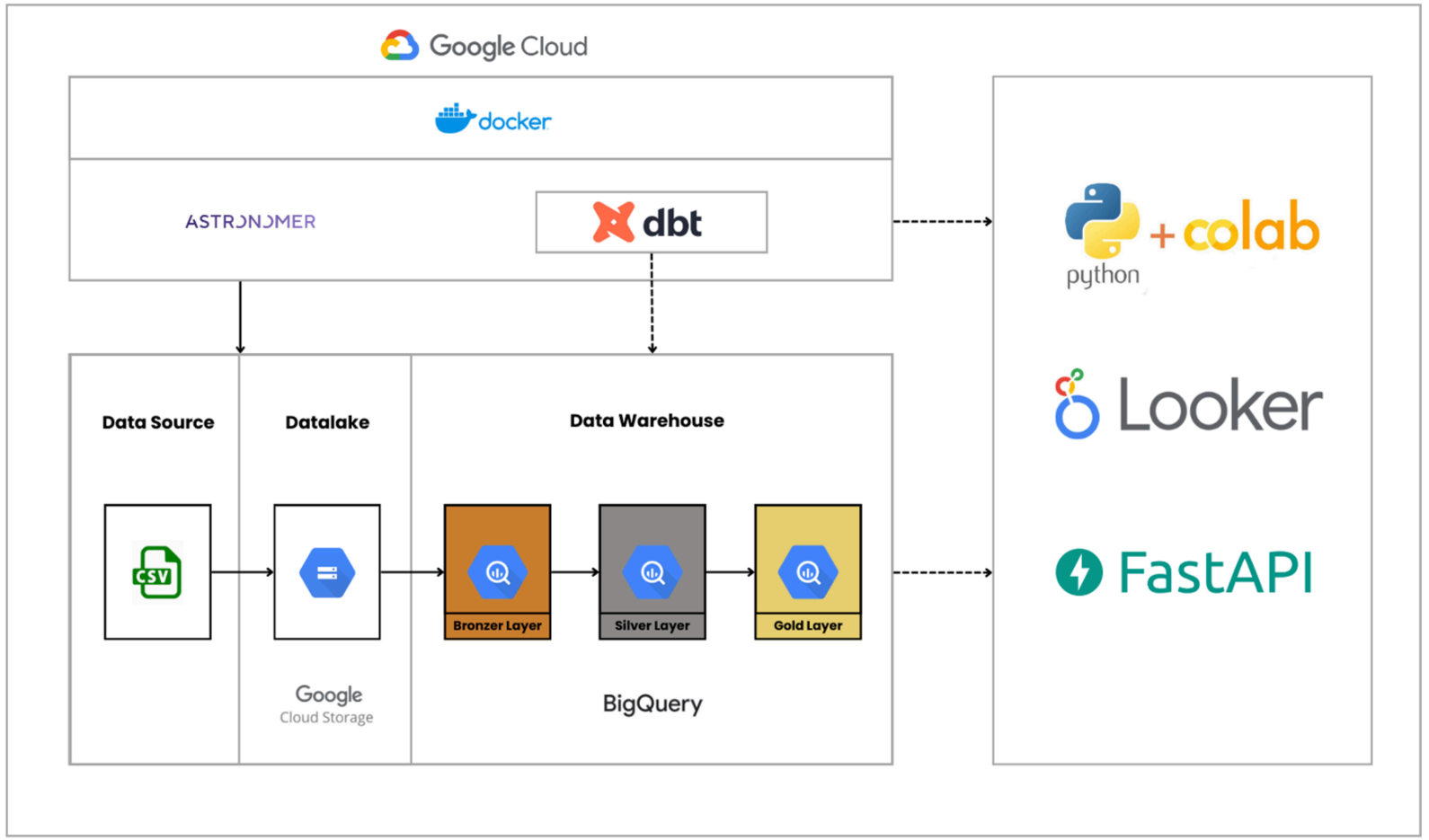
03 Scalability

04 Automation and Repeatability

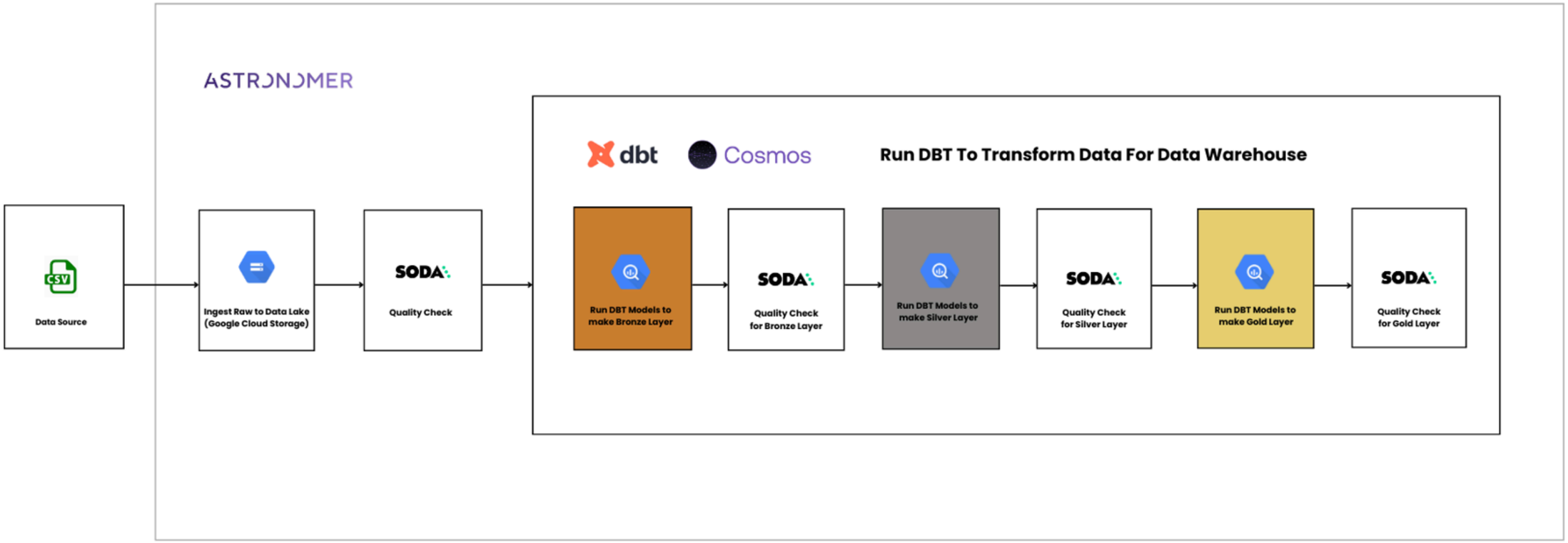
05 Deployment and Monitoring

05 Business Insights and Reporting

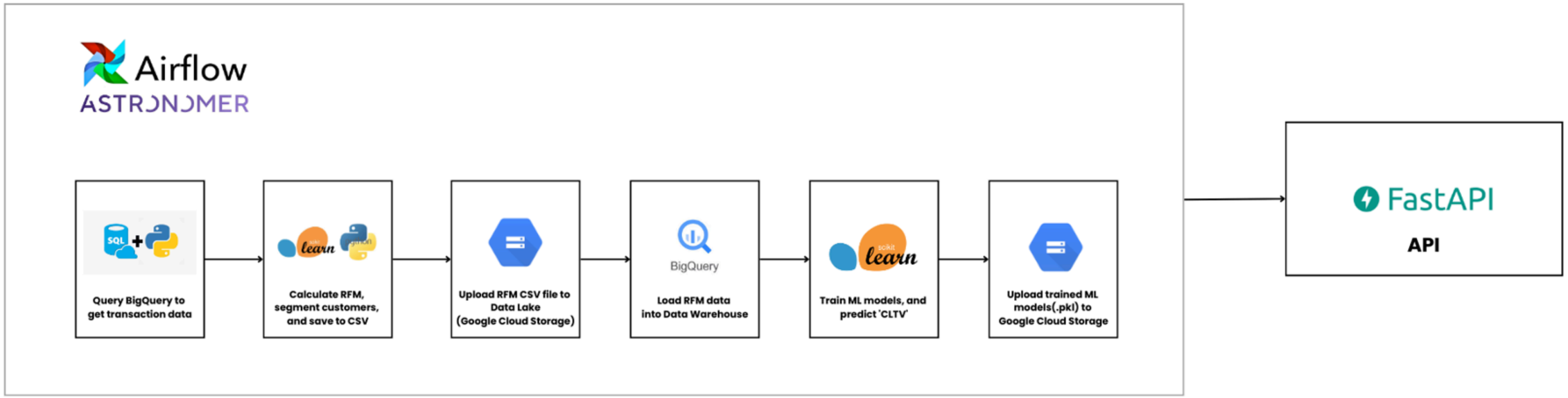
Architecture Design



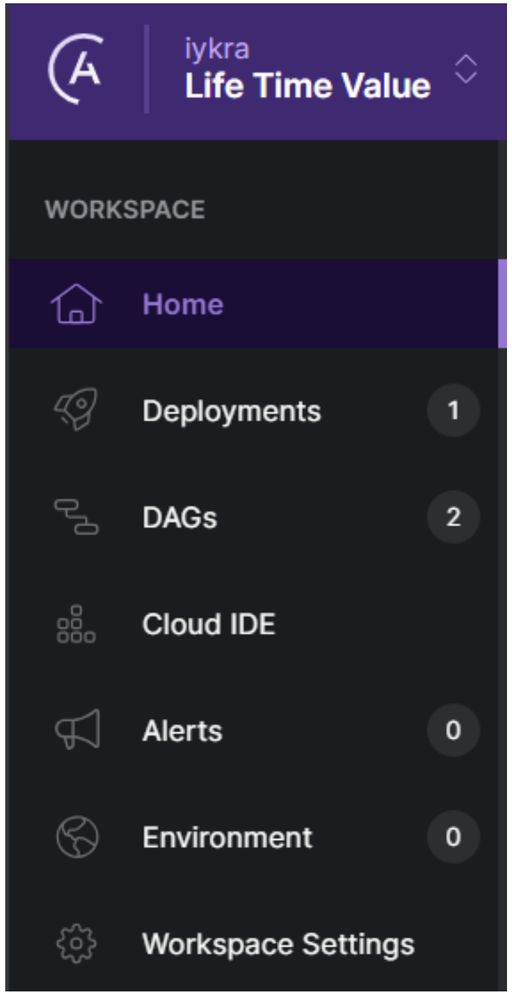
Data Pipeline

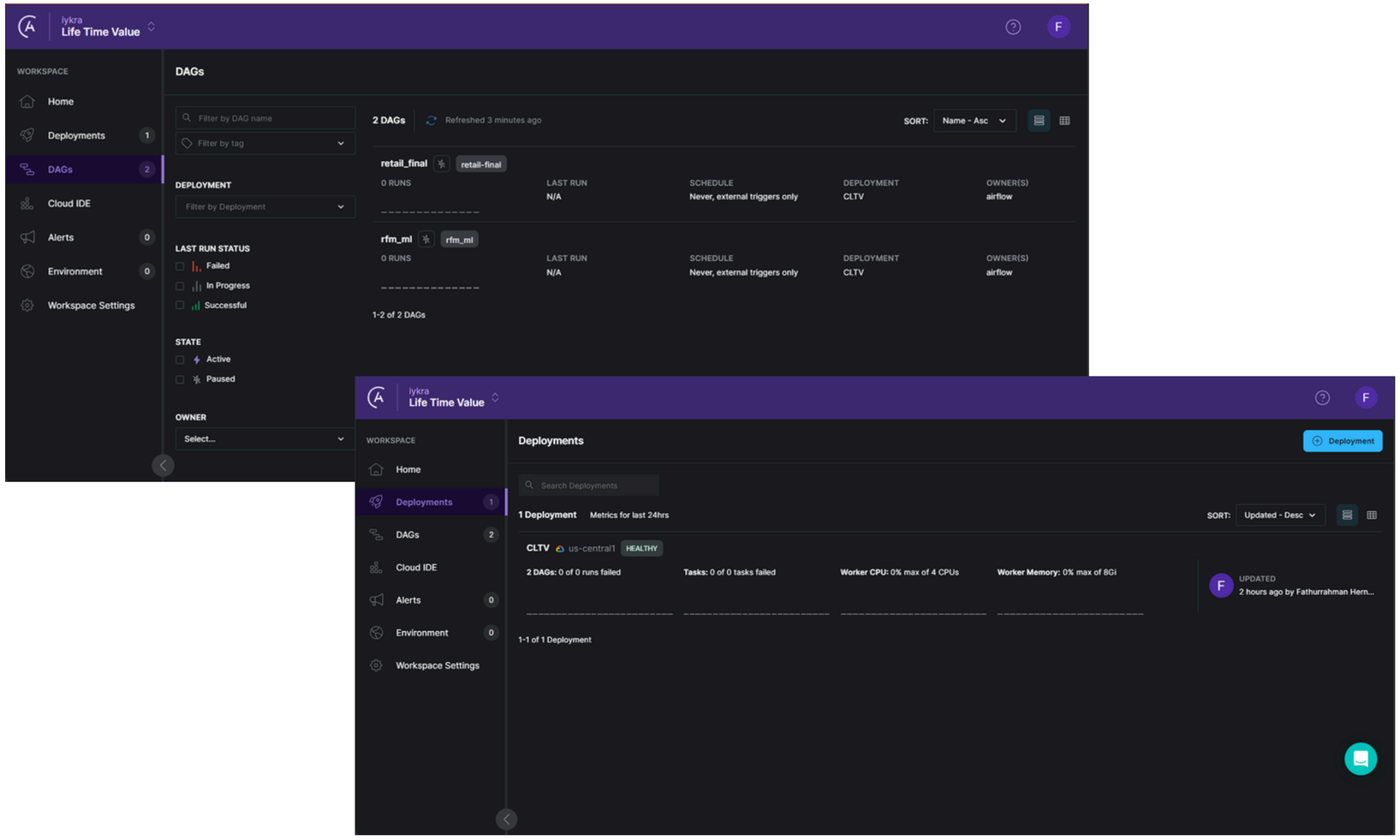


Machine Learning Pipeline



Deployment


iykra
Life Time Value
WORKSPACE
Home
Deployments 1
DAGs 2
Cloud IDE
Alerts 0
Environment 0
Workspace Settings


iykra
Life Time Value
WORKSPACE
Home
Deployments 1
DAGs 2
Cloud IDE
Alerts 0
Environment 0
Workspace Settings
DAGs
Filter by DAG name
Filter by tag
DEPLOYMENT
Filter by Deployment
LAST RUN STATUS
Failed
In Progress
Successful
STATE
Active
Paused
OWNER
Select...
2 DAGs
Refreshed 3 minutes ago
SORT: Name - Asc
retail_final
retail-final
0 RUNS
LAST RUN N/A
SCHEDULE Never, external triggers only
DEPLOYMENT CLTV
OWNER(S) airflow
rfm_ml
rfm_ml
1-2 of 2 DAGs
Deployments
Deployment
Search Deployments
1 Deployment
Metrics for last 24hrs
SORT: Updated - Desc
CLTV us-central1 HEALTHY
2 DAGs: 0 of 0 runs failed
Tasks: 0 of 0 tasks failed
Worker CPU: 0% max of 4 CPUs
Worker Memory: 0% max of 8Gi
UPDATED
2 hours ago by Fathurrahman Hern...
1-1 of 1 Deployment

Cost

Free

Data Warehouse	BigQuery
Orchestrator	Astronomer
Transformation Workflow	dbt
Data Governance	Soda
Data Visualization	Looker Studio

Paid

A. Data Lake: GCP Cloud Storage

Location	Standard storage (per GB per Month)
Jakarta (`asia-southeast2`)	$0.023

B. Compute Engine: GCP VM Instance

Monthly estimate

$135.45

That's about $0.19 hourly

Pay for what you use: no upfront costs and per second billing

Item	Monthly estimate
4 vCPU + 16 GB memory	$131.55
30 GB balanced persistent disk	$3.90
Total	$135.45

Data Fellowship

Data Ingestion

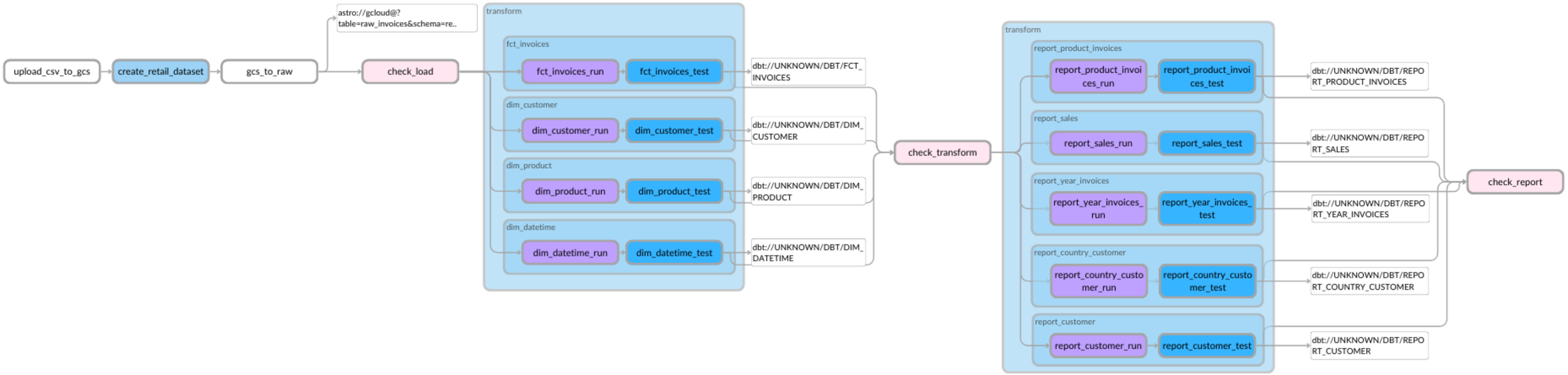
Data Source

Online_Retail.xlsx
Source: UCI Machine Learning Repository
This is a transactional data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.

- 45.78 MB
- 541909 records
- 8 columns:
  - 5 data categorical
  - 1 data continuous
  - 1 data integer
  - 1 data date

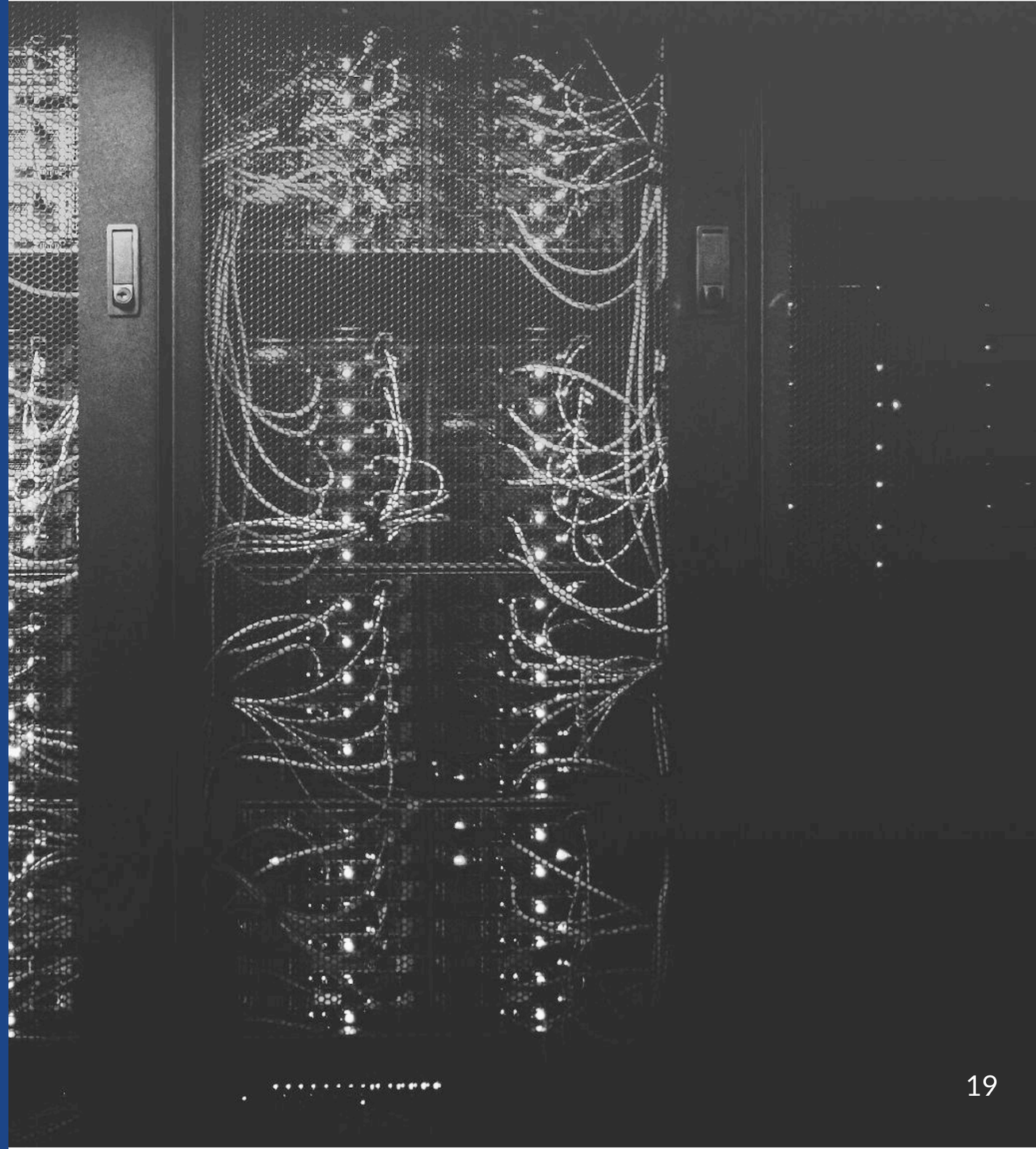
Column	Description
InvoiceNo	a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation
StockCode	a 5-digit integral number uniquely assigned to each distinct product
Description	product name
Quantity	the quantities of each product (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

Orchestrating Flow (Batch Pipeline)



Data Fellowship

Data Warehouse



Data Transformation

Raw Data → **dbt** → **Preprocessed Data**

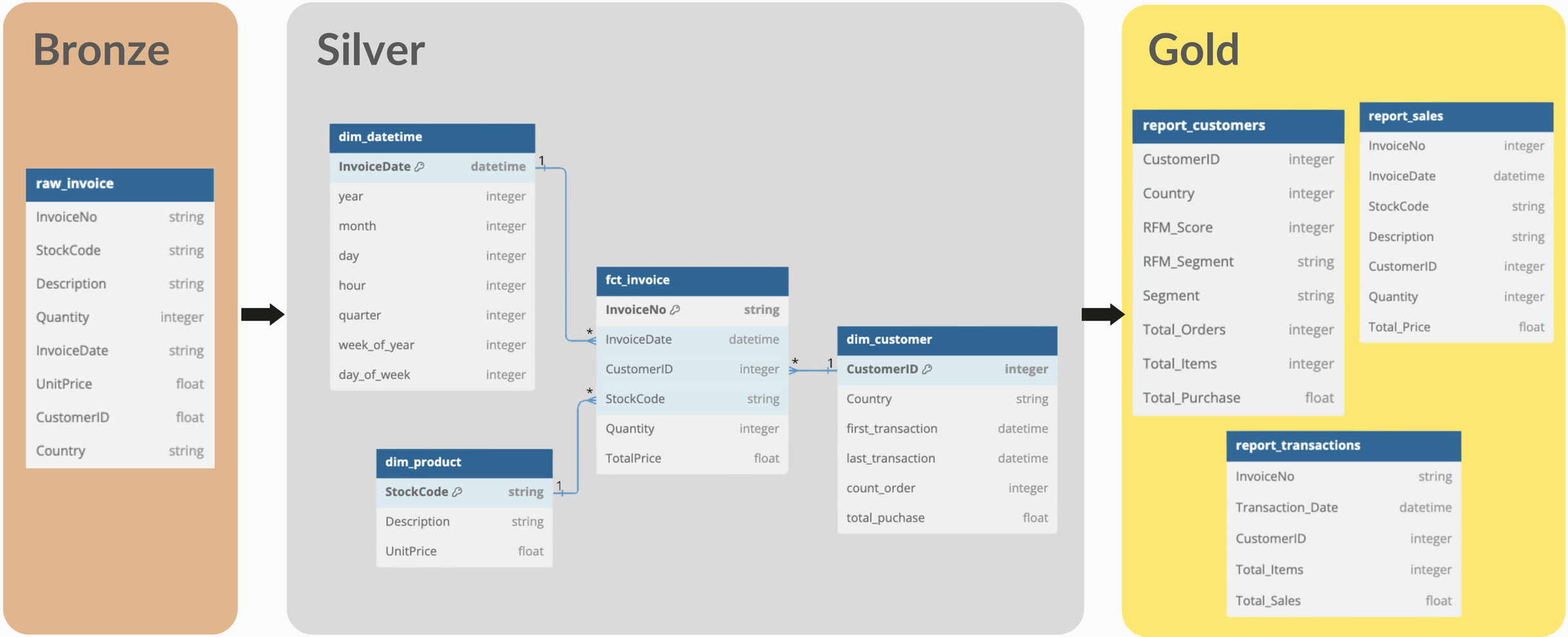
- Drop missing values in CustomerID.
- Change incorrect data types.
  `CustomerID: string → int`
- Extract InvoiceDate to add columns year, month, quarter, day of week, and week of year.
- Change UnitPrice for each StockCode with mode.

- Calculate TotalPrice using Quantity * UnitPrice

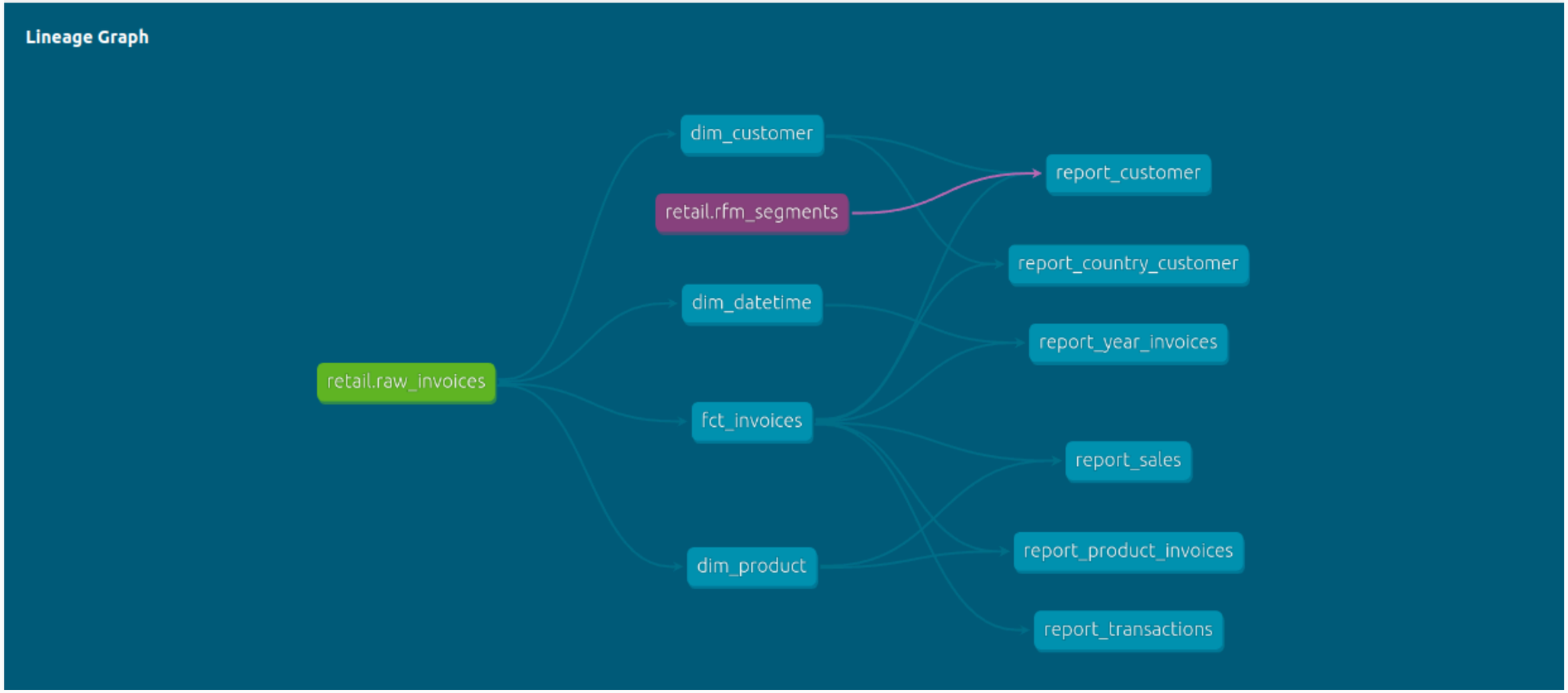
Quantity	UnitPrice	TotalPrice
2	2.5	5.0

- Drop rows with zero UnitPrice and negative Quantity.

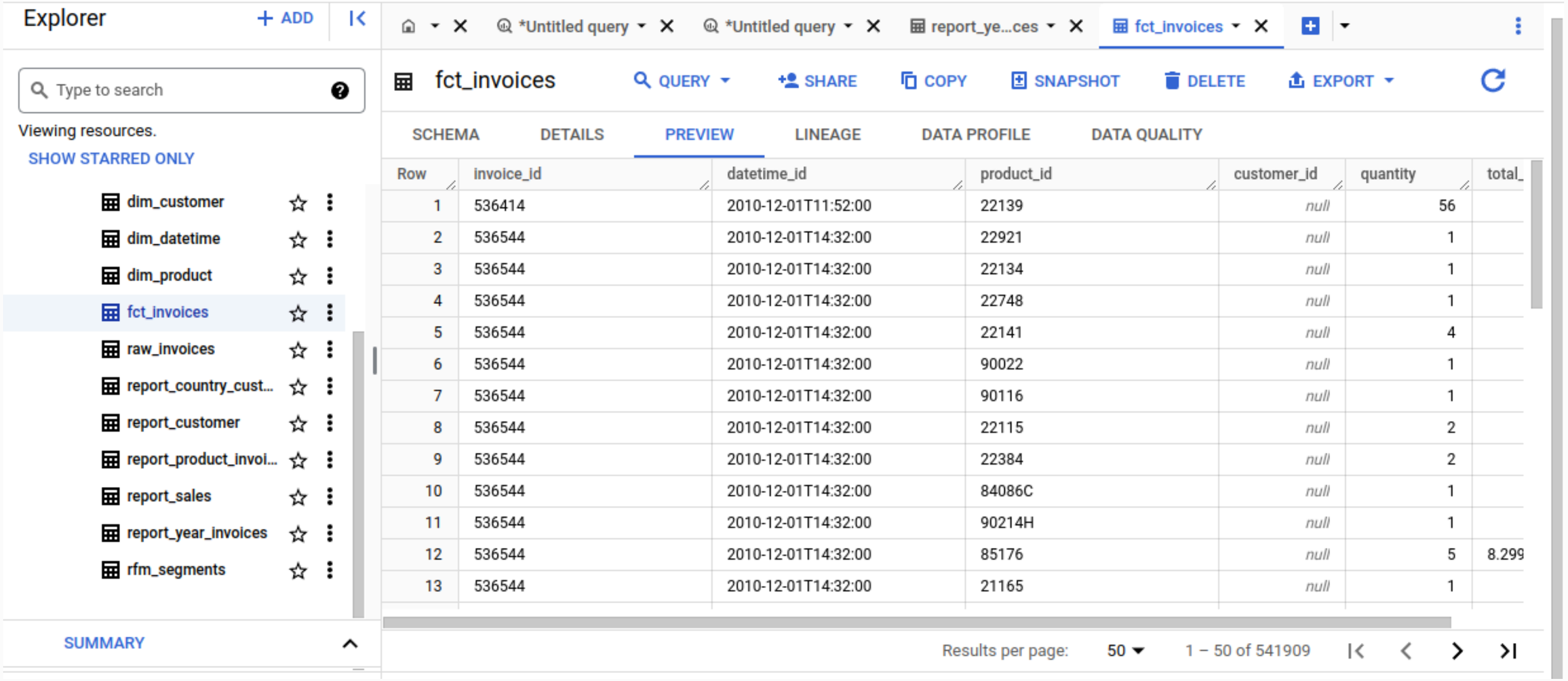
Database Schema



Data Lineage



Data Warehouse



Data Fellowship

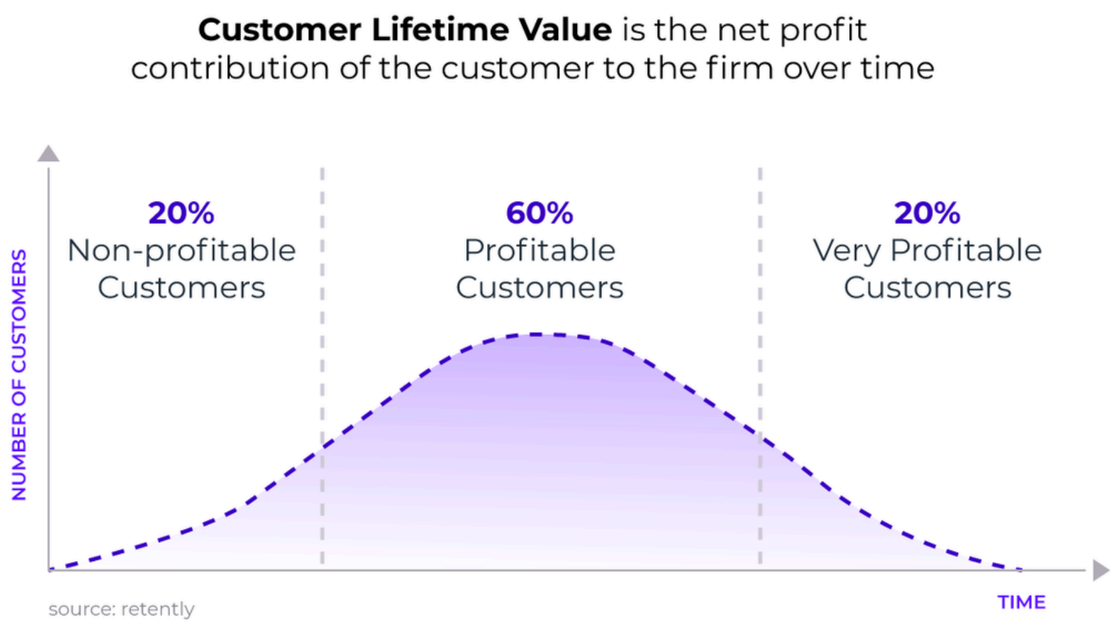
Machine Learning Model

Customer Lifetime Value

Customer Lifetime Value (CLV) is the monetary value that represents the amount of income or profit a customer will provide to the company during the relationship period.

The questions CLV tries to answer are:

- How to Identify **the most profitable customers**?
- How to **segment** profitable customers?
- How can a company **spend cost as little** and **make the most** money?



https://strivecloud.io/blog/cltv-improves-loyalty/

Customer Lifetime Value

Customer Lifetime Value (CLV) relies on informed assumptions, such as estimating average sale value, transaction frequency, and customer relationship duration.

Customer Lifetime Value = **Average Value of Sale** x **Average Number of Transactions** x **Average Customer Lifespan**

Average Value of Sales = Total Sales / Total number of orders

Average Number of Transactions = Total number of orders / Total unique customers

Steps of CLV Modeling

Take a dataset of customer purchases and **divide it into pre- and post-threshold periods**; the first 9 months and the 3 last months.

Calculate the R, F, M features in the pre-threshold period and the Monetary value in the post-threshold period.

Train a machine learning algorithm on the pre-threshold features, and use it to predict the post-threshold Monetary Value.

https://towardsdatascience.com/from-probabilistic-to-predictive-methods-for-mastering-customer-lifetime-value-72f090ebcde2

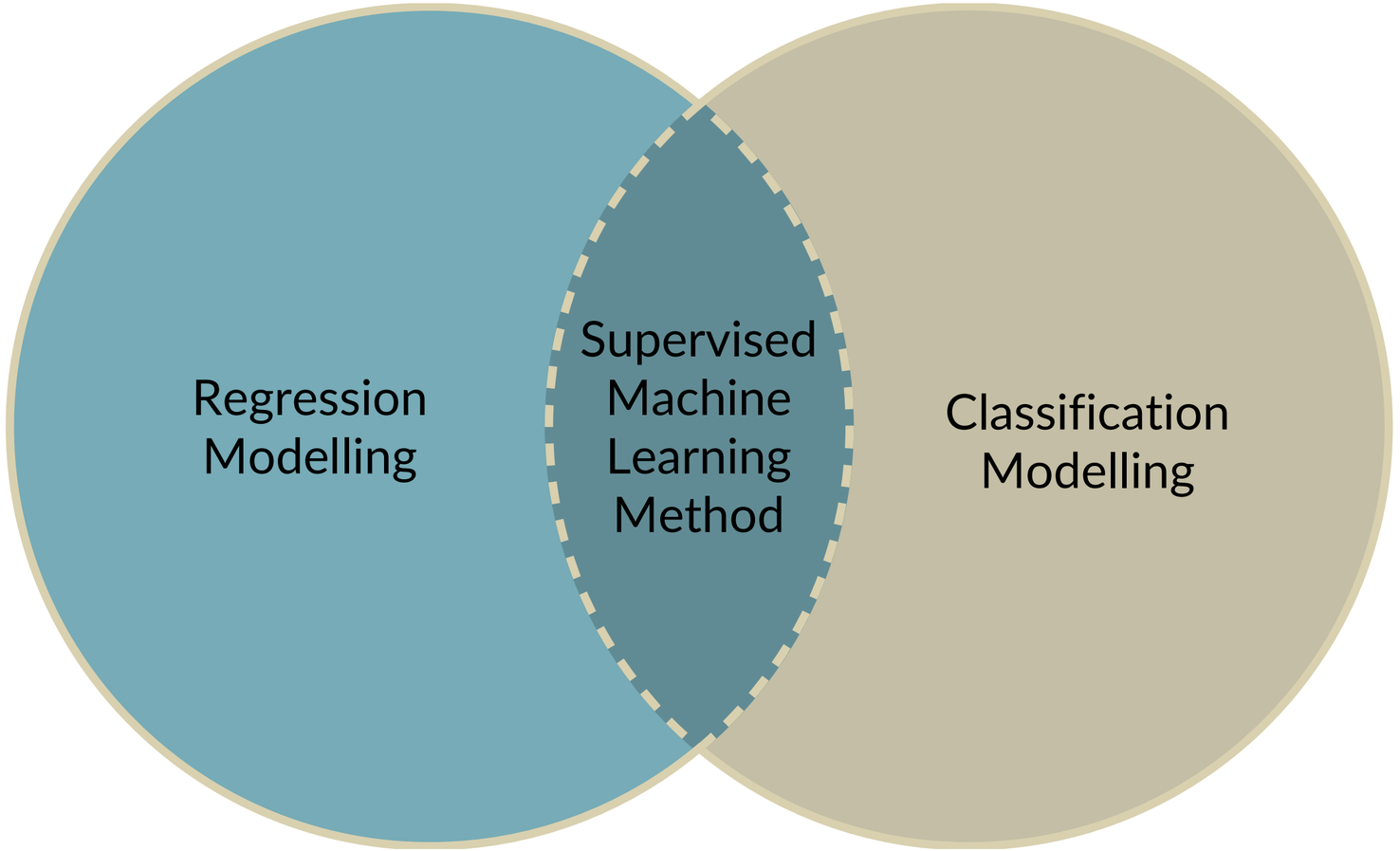
CLV and RFM

RFM analysis is a customer segmentation technique that evaluates which customers are most and least valuable to a company, based on recency, frequency and monetary value.

The next step is to link a description and marketing actions to each segment to improve your relationship with all the customers within the clusters.

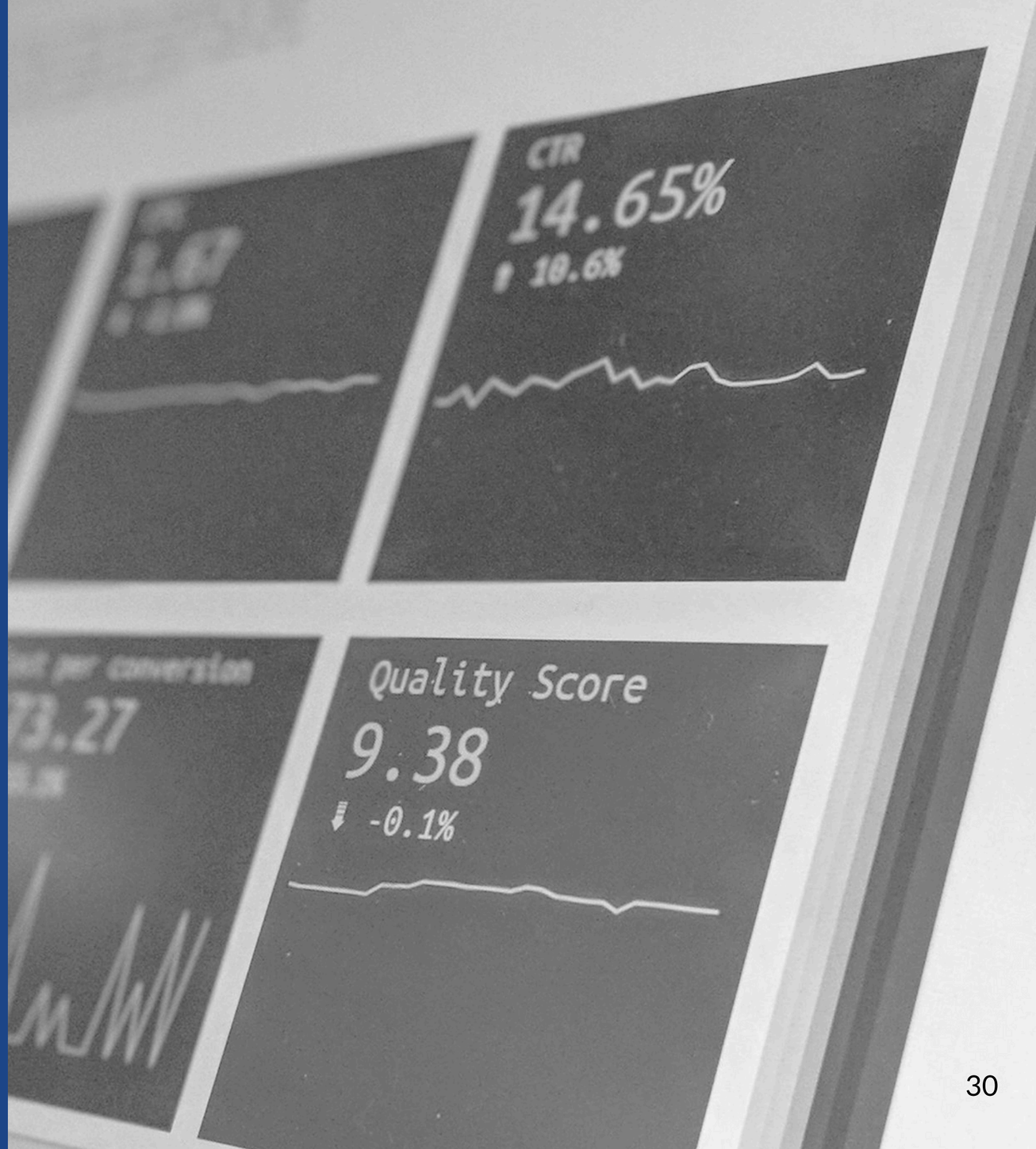
Through the segmentation of customers using RFM analysis, we can evaluate each group individually to ascertain which subset of customers exhibits the highest Customer Lifetime Value (CLV).

Modelling


Regression Modelling
Supervised Machine Learning Method
Classification Modelling

Data Fellowship

Reporting & Analytics



RFM Segmentation

1. Customers are segmented into 9 categories based on the combination of Recency, Frequency, and Monetary scores with quartile calculation.
2. Customers are also segmented based on their Monetary value into Low, Mid, and High value.
3. The most potential customers are Champions and Loyal Customers.
4. Customers that we need to attract and put effort into are Cannot Lose and At Risk.

rfm_segmentation

QUERY | SHARE | COPY | SNAPSHOT | DELETE | EXPORT

SCHEMA | DETAILS | PREVIEW | LINEAGE | DATA PROFILE | DATA QUALITY

Row	CustomerID	Recency	RecencyCluster	Frequency	FrequencyClust	Revenue	RevenueCluster	OverallScore	RFM_Score	RFM_Segment	RF_Segment	Segment
1	17949	23	5	70	1	58510.48	2	8	512	New Customers	New Customers	High
2	16013	25	5	139	1	37130.6	2	8	512	New Customers	New Customers	High
3	17857	26	5	54	1	26879.04	2	8	512	New Customers	New Customers	High
4	15769	29	5	130	1	56252.72	2	8	512	New Customers	New Customers	High
5	16333	29	5	45	1	26626.8	2	8	512	New Customers	New Customers	High
6	15838	33	5	167	1	33643.08	2	8	512	New Customers	New Customers	High
7	12931	43	5	82	1	42055.96	2	8	512	New Customers	New Customers	High
8	16446	22	5	3	1	168472.5	3	9	513	Promising	New Customers	High
9	12433	22	5	420	2	13375.87	1	8	521	Potential Loyalist	Potential Loyalists	High
10	12662	22	5	230	2	3849.78	1	8	521	Potential Loyalist	Potential Loyalists	High
11	17581	22	5	440	2	11045.04	1	8	521	Potential Loyalist	Potential Loyalists	High
12	13069	22	5	425	2	4436.12	1	8	521	Potential Loyalist	Potential Loyalists	High
13	17428	22	5	328	2	17256.85	1	8	521	Potential Loyalist	Potential Loyalists	High
14	16558	22	5	460	2	8338.49	1	8	521	Potential Loyalist	Potential Loyalists	High

Prediction of Customer Lifetime Value

Result of Regression Model

	Model	Best Params	MSE	R2	MAPE	CV_MSE
0	Linear Regression	{}	365361829984127808.0	0.070737	127250453195177334931456.0	157590469238921536.0
1	Random Forest	{'regressor__max_depth': 20, 'regressor__n_est...	155631170024300192.0	0.604167	53755494681312472596480.0	200416472617369984.0
2	Decision Tree	{'regressor__max_depth': None, 'regressor__min...	484908667511618432.0	-0.233319	103485350516024093442048.0	292986448746209408.0
3	Support Vector Regressor	{'regressor__C': 10, 'regressor__epsilon': 0.01}	394385270530325952.0	-0.003082	1340129814396766257152.0	234006138001537472.0

Example of Prediction

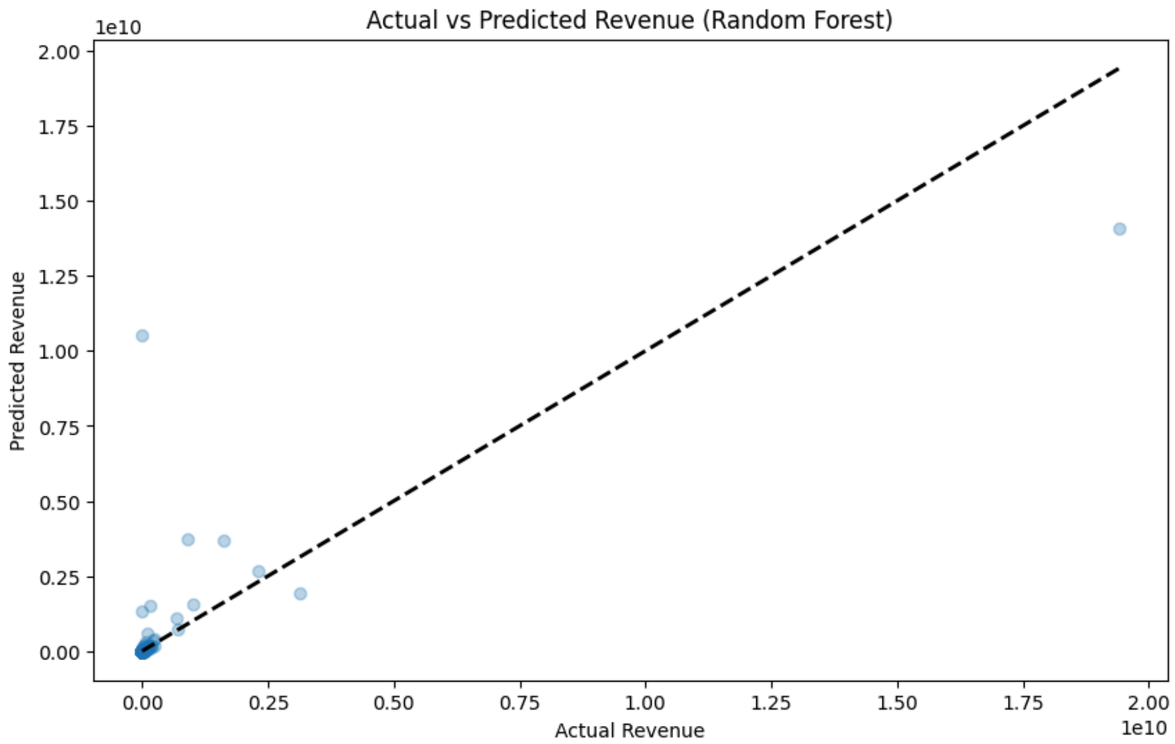
	Recency	Frequency	Density	Monetary	R_Score	F_Score	D_Score	M_Score
0	10	50	0.1	5000	4	3	3	4
1	5	30	0.2	3000	3	3	2	4

Result of Prediction

	Recency	Frequency	Density	Monetary	R_Score	F_Score	D_Score	M_Score	Regression_Predictions
0	10	50	0.1	5000	4	3	3	4	699089.184160
1	5	30	0.2	3000	3	3	2	4	699248.467888

Prediction of Customer Lifetime Value

	CustomerID	Actual CLV	Predicted CLV
0	14276	1864841.172626	698460.009900
1	15197	1454783.119899	696704.695647
2	14667	24193565.559516	700944.411365
3	16109	0.0	697989.387784
4	12664	134214636.722269	699372.483297
5	15397	0.0	696224.190256
6	17870	2445029.76453	699693.080162
7	14314	0.0	698040.729215
8	14280	0.0	696725.197006
9	13246	42382.409397	699632.879585

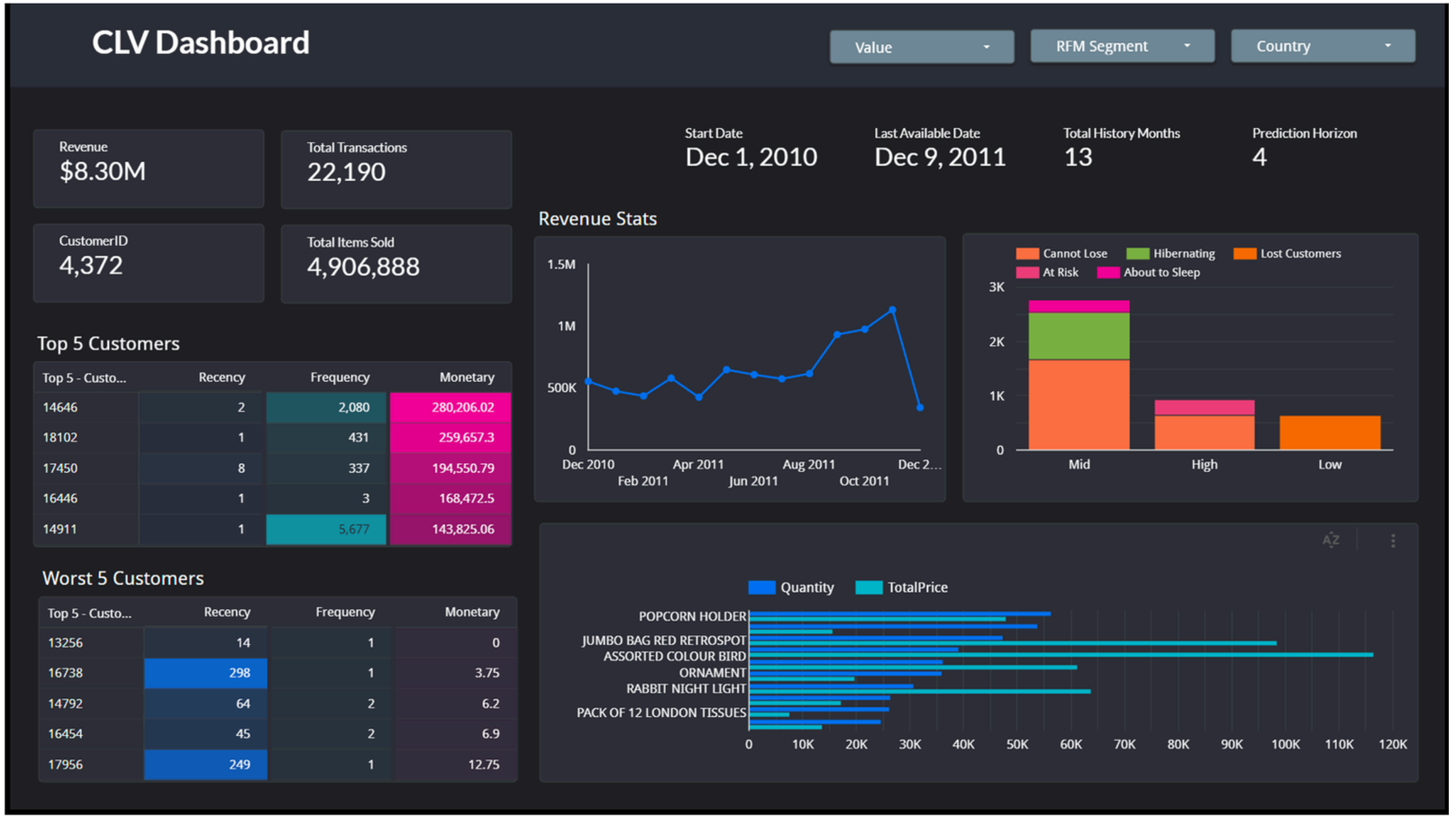


Prediction of Customer Lifetime Value

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

Based on the results of the model that we processed, we determined that the XGBoost model was chosen for prediction because the results were high and had a small difference between the train set and test set values so that the prediction results would later reach a high level of accuracy.

Dashboard


CLV Dashboard
Value
RFM Segment
Country
Revenue
$8.30M
Total Transactions
22,190
CustomerID
4,372
Total Items Sold
4,906,888
Start Date
Dec 1, 2010
Last Available Date
Dec 9, 2011
Total History Months
13
Prediction Horizon
4
Top 5 Customers
Top 5 - Custo... Recency Frequency Monetary
14646 2 2,080 280,206.02
18102 1 431 259,657.3
17450 8 337 194,550.79
16446 1 3 168,472.5
14911 1 5,677 143,825.06
Worst 5 Customers
Top 5 - Custo... Recency Frequency Monetary
13256 14 1 0
16738 298 1 3.75
14792 64 2 6.2
16454 45 2 6.9
17956 249 1 12.75
Revenue Stats
1.5M
1M
500K
0
Dec 2010
Feb 2011
Apr 2011
Jun 2011
Aug 2011
Oct 2011
Dec 2...
Cannot Lose
Hibernating
Lost Customers
At Risk
About to Sleep
3K
2K
1K
0
Mid
High
Low
Quantity
TotalPrice
POPCORN HOLDER
JUMBO BAG RED RETROSPOT
ASSORTED COLOUR BIRD
ORNAMENT
RABBIT NIGHT LIGHT
PACK OF 12 LONDON TISSUES
0
10K
20K
30K
40K
50K
60K
70K
80K
90K
100K
110K
120K

How to segment profitable customers?

For our top customers in these segments we would recommend the following:

1. Organize loyalty programs especially targeting at **Champions, Loyal Customers**
2. **Advertise Limited Edition** products mainly for **value monetary High**

How to increase and get returning customers?

For our critical customers in these segments we would recommend the following:

1. **Pay attention** to feedbacks of **Cannot Lose** and **At Risk** segment
2. Provide special discounts and free shipments

How can a company spend cost as little and make the most money?

Top sales in November - December

1. **Efficiency** of the **amount of goods** supplied in that month
2. **Start planning** your **marketing campaign** before the increase in sales occurs
3. **Low-Cost Offers**: Provide low-cost incentives such as free shipping or small discounts for **Hibernating** segment

Data Fellowship

Conclusion & Improvement

Conclusion

1. The system can be used to segment the customers based on past transaction behavior using the Recency, Frequency, and Monetary method into 9 segments.
2. The system can automatically run each period of time whether it is monthly, quarterly, or every 6 months to generate a report that can be accessed through the dashboard.
3. Based on the segmentation from historical data and prediction, decision-making can be made to create personalized segment marketing.

Improvement

1. Improve the machine learning modeling to create a prediction with better accuracy, because the segmentation highly depends on the predicted CLV.
2. Need to elaborate more demographic data regarding customers to make a better feature engineering in the machine learning model.

Thank you!