Project | CLV |

536543

PAGE NO.:

(i) Invoice (Bill ID) → 6 digit "xxxxxx"
→ each bill has multiple items.
Single customer → Unique invoices
↳ Frequency Of Purchase.

Item per invoice
↳ Average basket size (Per transaction Avg ~~that~~ ₹)

Some invoice starts with "Cxxxxxx" are returned orders with Qty → negative.
- Work around return.
  - Flag return (to find frequency if it matter)
  - Net spend (Remove return if it creates difference)
  - Value of item returned.

(ii) Stock Code (Unique product ID)
variety a customer bought.

Will think more!.
Later...

(iii) Description (Product specification)
Can't see any use...
Have 1 idea (Advance ₹ !. ><)
NLP use ~~such~~ Categorise product as "Fashion", "Decor", "Hygiene" then encode and use... Let see Later!!
Currently No Use!...

(iv) Quantity: (units of product)
Total Qty by a customer (bought)
Avg qty per order. (invoice)
Detect returns
buyer category → Bulk, avg, Light.
Trend of buying overtime, Decrease, Const, Increase.

(v) InvoiceDate

DD-MM-YYYY, HH:MM

↓ No use as such !!.

Recency very useful ☺

Extract Months → Group it into Quarters over the Year.

! Check the trend ——

→ Based on orders (Qty, invoices)

→ Amt. spend.

Category: Increasing, constant, Decreasing.

Few terms:

Recency: Pivot Dat (Let say end of year) − Last Purchase.

⇒ Lower the number recent the customer.

First Purchase − Last Purchase = Tenure

Total Purtase (Qty) / Tenure = Purchase Frequency.

says * Month matter.

Avg (Gap between each purchase)

⇒ "Frequent" or "infrequent" buyer.

(vi) Price (Per unit product)

total spend → Qty × Rice.

avg spend per invoice

Spend pattern over time

Observation: for −ve Qty somewhere the price → 0

somewhere same as at what they bought.

Mostly no errors, but Yes 'bulk pricing' is there. → no effect

Small discount on huge Qty.

(vii) Customer_ID (unique identifier for customer).

Basically we have 2 Years of Data

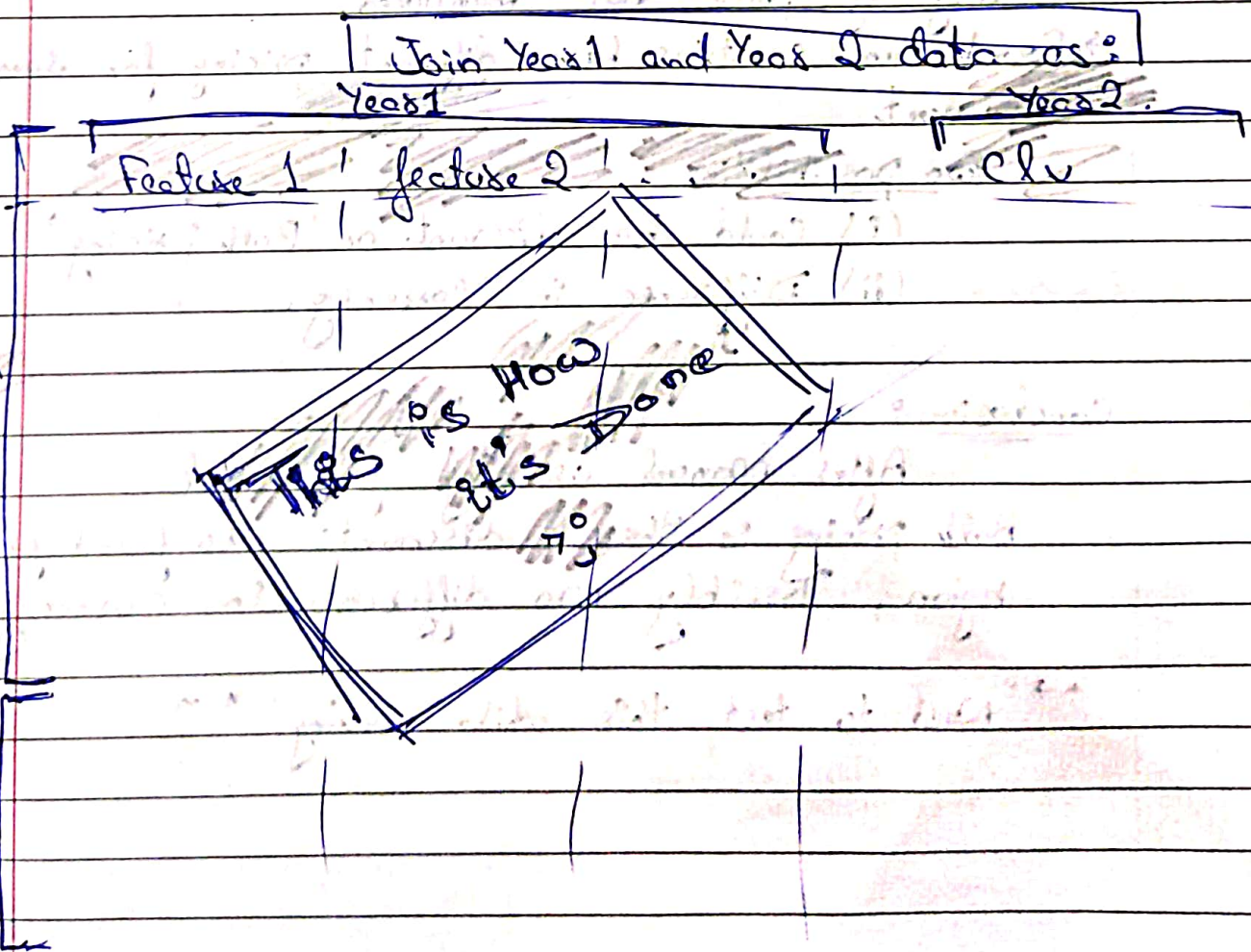
2009-2010 ; 2010-2011

#Note: Observation Says:

few 2010 data in Year 1, copy also present in Year 2

∴ Redundancy (Handle it)

Year 1	Year 2.
→ Group by customer ID	→ Same customer Group from Year 1.
→ Find features	→ ~~Map~~ Calculate clv of Year 2.
→ Train model to learn it.	→ Map each customer to its corresponding clv of Year 2



(viii) Country :- (country of the customer)

Comm...

Can categorize based on some 'continet' or 'Region' wise,
Need to find correlation with CLV to say wheather it impose any impact or not ?.

OBSERVATIONS :

#Manual Data Assessing

Let's go through the CSV and try observing things and note it down, other that what we have discovered Yet ?.

Observation 1 :

Currency of 'price' not mentioned.

→ Found some countries have different pricing for same product.

Two possibilities

(i) Could be discount or 'Bulk pricing'

(ii) Difference in currency.

Conclusion :

After Manual assesing found that Bulk pricing is there, discounts also found for region, Possibly no difference in currency...

** "Need to test this while working" **

PAGE NO.:

Observation 2:

Distribution:

Year 1: (2009-2010)

5,24,461 → Rows

8 → Columns (Features)

But; Out of 5,24,461 : _____ Customer ID are blank 'NULL'.

→ Most Blank Customer ID is associated with

(i) Invalid product description - '?' '???' or even null description

(ii) Null Stock Code.

(iii) Priced → 0 (Returns)

(iv) Qty → -ve (Returns)

Year 2: (2010-2011)

5,41,900 → Rows

8 → Colums.

_____ Customer ID's are Blank.

Rest observation Same as Year 1...

Can we do something about Customer ID...

→ If consider 'Country' ⇒ Multiple Invoice (can be of diff persons)

→ Considering grouping Invoice as a single,

⇒ Might miss other orders of same person

⇒ No exact CLV...

⇒ Won't be able to map 2 customer over Year1 and Year2....

Therefore: Lets' think more,

else last option 'dropNA' !!

Observation 3:

Unusual StockID

includes : Gifts, PADS, POST, M, S, SPxxx, TESTxxx

⇒ Gifts are ; guess free items (complimentary) things.
- Gift with → Price Zero(0) are free.
- Gift with → Some price are vaoucheers.

at end doesn't Matter as none of the column with Gift Category is associated with CustomerID have to drop it anyway.

⇒ M → Manual entries,
⇒ Mostly [redacted] looks artificial.
⇒ Looked deeply, best to drop.

⇒ PADS → It is an item {associated with cusions}
thus, no need to drop, treat usual.

⇒ POST → Postage charges (Service charges)
and contributes to a customers expense.
thus, will keep it...

⇒ Samples (S) → Drop (Null CustomerID)
⇒ SPxxx → 2 valid entry
⇒ TESTxxx → Not real Sales, thus drop it.
↓
Description tell its a test entry possibly created by the one who made this Dataset.
very few entries, no effect if dropped.

PAGE NO.:

Finalised features:

- → Customer ID
- → TotalSpend (Imp)
- → PurchaseFrequency
- → OrderHabit
- → Recency
- → AvgPurchaseGap
- → Trend
- → Churn
- → ReturnRate

= Features.

→ CLV (Target)

Customer ID: Do use for these:
∴ Drop !.

TotalSpend: ⟶ "Log-TotalSpend"
- Skewed; Normalised (using log Transform)

PurchaseFrequency
- Skewed;
- Normalised (Log transformation, Box(cox))

Reduced: Skewness 0.9 Skewness 0.08

But: Distribution is not Normal
Due to missing and more zeroes

- Binning: Converted to labels (categories)
  Shows better trend, will go for it!
  → "PurchaseFreq-Label"
- IsFrequentBuyer: binary feature if Label is 'High' as 'SuperBuyer'

OrderHabit : (Avg. Qty / invoice)
- Highly Skewed.
- Log transformation {Bell curve achived}
- Label_OrderHabit
  = Light Buyer, Regular, Moderate, Heavy, Bulk
  ; Gives more information.
  ∴ good...

Will try both features ;

Recency :
- Bit skewed (Right)
- long right tail
- Log transformation ⇒ Skewness reduced
  but the distribution is not bell curved
  lot of ~~missing~~ zero values.
- BoxCox Handles '0' well but yet not satisfied distribution.

- Best ⇒ Recency to Recency_QBin

V.Rec Rec Mid old Dormant

Average Purchase Gap :
- Skewed.
- Log transform. (Good) {less skewed, but - long tail}
- Gap_Label (Binning).
  ↳ shows better trend with
  "Churn, trend, CLV"...

Trend → Trend-Refined.

Classified 'Flag' → New, seasonal based on features like their clv and spend trend.

∴ Trend is diverse now with 5 value (categories

Churn:

Initial churn logic:

- last 2 Quarter no spend.
- Hard, Not trustworthy.

∘∘ Pseudo churn.

Churn (New)

- If inactive in last 2 Quarters (Pseudo churn == 1)
- PurchaseFrequency is ~~X~~ = 2. (Not single buyer)
- clv < median_clv.

Then 1 - churn

0 - Not churn.

Better and Strong churn

{ We already know they have produced value in clv, hence definitely not churn }

but featured it to incorporate few things ...

Returnrate:

Lots of zeros, Can't perform log transformation.

∘∘ Binning.

~~No~~ No cross Fee High

Trend is not much good!.

High returner also have High CLV.

will try and decide while No returner have competiti to keep or drop. low clv.

Final Set of features

- TotalSpend : Log-TotalSpend.
- PurchaseFrequency : PurchaseFreq_Label.
- OrderHabit : Log-OrderHabit , Label-OrderHabit.
- Recency : Recency-QBin.
- Avg-PurchaseGap : Log-PurchaseGap , Gap-Label.
- Trend-Refined.
- Churn
- Returnrate_Label.
- CLV. — — — Target.