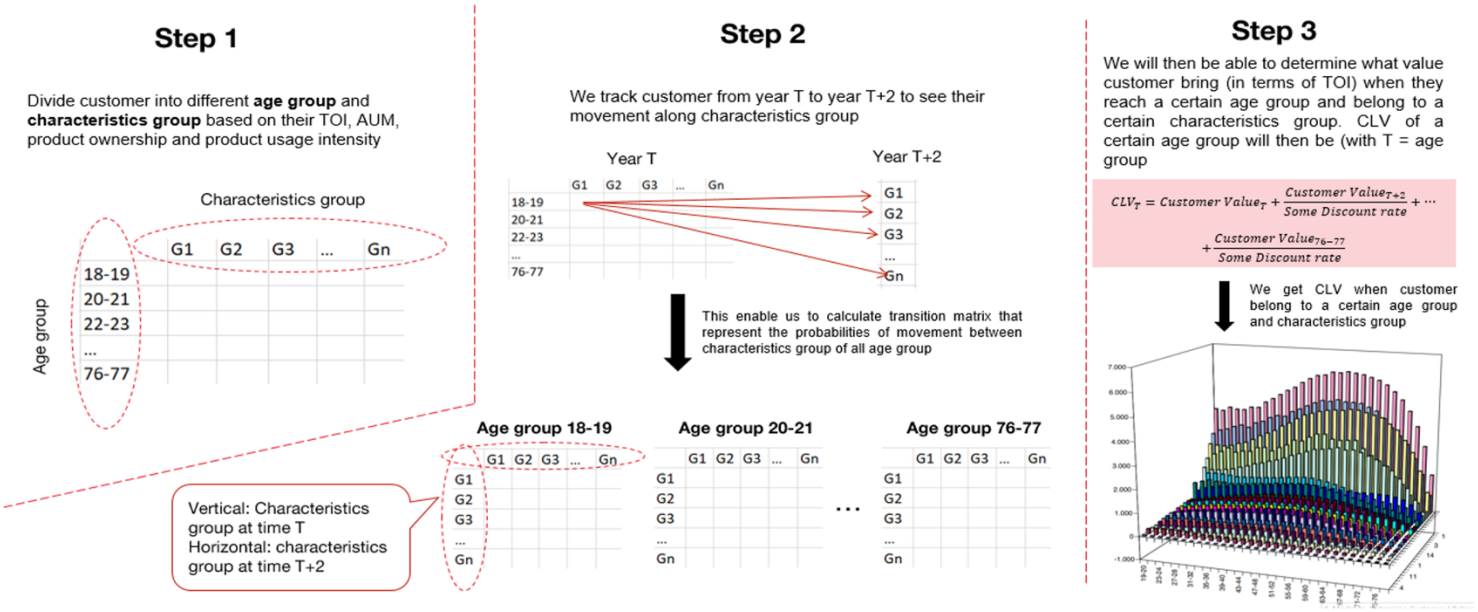
Customer Life Time Value (CLV) Calculation

This calculation process was based on the research paper A Model to Determine Customer Lifetime Value in a Retail Banking Context published in 2007. CLV is one of the most foundational and useful tool any enterprise can develop. Yet it is very challenging due to the need of a long-time range dataset and the unpredictability nature of forecasting the future, especially in this case, decades forward. This is my attempt to calculate the bank CLV. Many tests were carried out during the calculation process to make us aware of the limitation of the method.

1. Methodology overview



In here TOI means Total Operating Income that customer brings the bank

2. Data preparation

AFF: 285,786
MAF: 625,043
MAS: 1,608,300

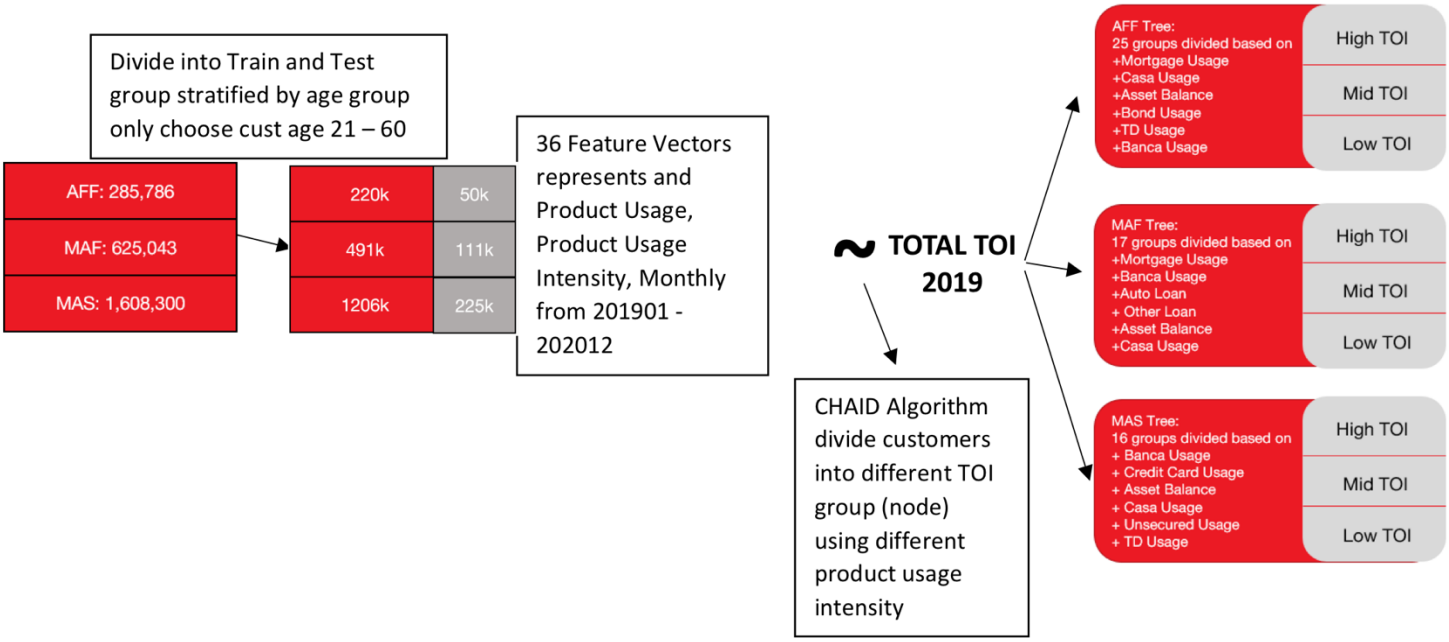
Customer in scope is all TCB customer onboard before 202001 and is active as of 201912 (have product balance, have active transaction)

Customer is divided into 3 different economic segments: Affluence, Mass Affluence and Mass. Each segment has very different characteristic

37 Feature Vectors represents TOI and Product Usage, Product Usage Intensity, Monthly from 201901 - 202112

toi	casa_flag
sum_tranfer_txn	td_flag
sum_tranfer_amt	bond_flag
sum_wd_txn	fund_flag
sum_wd_amt	auto_loan_avg_bal
avg_tranfer_txn	mortgage_avg_bal
avg_tranfer_amt	unsecured_avg_bal
avg_wd_txn	he_avg_bal
avg_wd_amt	hh_avg_bal
sum_db_amt	other_loan_avg_bal
sum_db_txn	auto_loan_flag
sum_cr_amt	mortgage_flag
sum_cr_txn	unsecured_flag
avg_bal_casa_12m	he_flag
avg_td_amt_12m	hh_flag
bond_avg_bal_12m	other_loan_flag
fund_avg_bal_12m	ape
avg_asset_12m	no_contract_ape
	banca_flag

3. Divide customer into node(groups) using CHAID algorithm (Step 1 in the Methodology Overview)



CHAID (Chi-squared Automatic Interaction Detection) was used to create decision tree for each economic segment. The algorithm helps divide customers into different groups using their product usage intensity. The algorithm optimises the fact that each node will have clearest separation from each other in terms of TOI. The advantage of this algorithm is that branches from node is not binary. Many branches can be separated from one single node.

3.1 Characteristics groups (nodes) of AFF customers resulted from the tree using train data

Group	TOI	No of Cust	% Population	MORTGAGE_AVG_BAL	AVG_TD_AMT_12M	AVG_BAL_CASA_12M	APE	Bond	AVG_ASSET_12M
node_13	653,495,388	2,085	0.77%	> 7,826,391,853					
node_21	223,373,951	2,570	0.94%	No		> 155,235,339		Yes	
node_12	221,382,860	4,012	1.47%	2,897,268,401 : 7,826,391,853					
node_62	144,344,470	5,234	1.92%	No	> 849,414,783	> 155,235,339		No	
node_10	115,434,477	4,035	1.48%	1,619,116,155 : 2,897,268,401					
node_20	81,933,679	3,777	1.39%	No		47,554,903 : 155,235,339		Yes	
node_42	77,719,515	2,672	0.98%	No	No		> 36,002,100	No	
node_8	70,094,755	1,948	0.72%	1,356,725,880 : 1,619,116,155					
node_39	57,850,416	7,024	2.58%	No	No		No	No	> 216,867,199
node_7	54,487,558	3,931	1.44%	937,563,247 : 1,356,725,880					
node_6	53,533,949	4,100	1.51%	<= 937,563,247					
node_61	53,316,447	17,048	6.26%	No	> 849,414,783	<= 155,235,339		No	
node_18	47,654,708	5,994	2.20%	No		<= 47,554,903		Yes	
node_55	37,908,329	11,457	4.21%	No	(299,999,999 : 849,414,783)	> 47,554,903		No	
node_50	33,389,312	5,506	2.02%	No	(84,545,454 : 299,999,999)	> 78,847,927		No	
node_44	32,867,731	2,556	0.94%	No	<= 84,545,454		Yes	No	
node_41	27,944,009	4,825	1.77%	No	No		<= 36002100	No	
node_54	18,738,080	12,072	4.43%	No	(299,999,999 : 849,414,783)	<= 47,554,903		No	
node_37	18,693,606	9,719	3.57%	No	No		No	No	101,698,370 : 216,867,199
node_48	11,016,040	18,974	6.97%	No	84,545,454 : 299,999,999	<= 78,847,927		No	
node_36	10,305,320	14,842	5.45%	No			No	No	49,538,597 : 101,698,370
node_43	6,919,046	22,935	8.42%	No	<= 84,545,454		No	No	
node_35	5,030,833	40,099	14.73%	No	No		No	No	12,732,726 : 49,538,597
node_31	3,727,450	19,494	7.16%	No	No		No	No	<= 1,920,785
node_33	2,596,990	45,677	16.78%	No	No		No	No	1,920,785 : 12,732,726

+ High TOI nodes mostly customer using mortgage, if not using mortgage, they will have high average casa balance compared to other nodes.
+ Mid TOI nodes tends have moderate TD (Term Deposit) balance and average casa balance
+ Low TOI nodes don't use mortgage, TD, banca, bond. They have low average asset balance

Model Accuracy
We test of CHAID tree model accuracy in test data to assess the validity of groups:
The test assesses the error terms, calculated by the absolute value of predicted TOI minus actual TOI of customer, all divide by the actual TOI. The smaller the number, the more accurate the model.

AFF NODE	Error
node_31	57.74%
node 33	41.23%
node_7	35.39%
node_8	34.92%
node_35	28.02%
node_10	24.21%
node_13	22.60%
node_42	20.78%
node 12	19.88%
node_37	19.04%
node_36	14.93%
node_61	14.88%
node_6	14.59%
node_43	14.20%
node_39	12.46%
node_48	11.93%
node_41	10.91%
node_55	10.01%
node_44	9.40%
node_54	9.10%
node_62	6.70%
node_50	2.96%
node_20	2.78%
node 18	1.67%
node_21	1.36%

Attention: The CHAID AFF Tree have low accuracy when predicting value of node 31 and 33 which are low TOI node with high number of customers

Overall, the model can predict TOI of test group with error of 17%

Model validity when further divide nodes into small age groups
It is also important that the predicted TOI are consistent among age groups because later in the CLV calculation process, customers will be divided by age groups. We have to check if the predicted TOI are equally applicable for all age groups. We divide nodes into smaller age group (21-30, 31-40, 41-50, 51-60) and use pairwise T-test test to see if those groups have the similar distribution.

AFF NODE	No of pairs have pvals >0.1
node_31	0
node_33	1
node_48	1
node_37	2
node_54	1
node_35	1
node_62	2
node_36	1
node_61	0
node_43	1
node_50	3
node_20	1
node_39	4
node_55	1
node_18	0
node_8	3
node_44	6
node_21	2
node_6	2
node_7	3
node 12	2

Attention: Pairwise t-test result show many of the groups don't have similar predicted TOI between groups

Model validity if data of different time range was used 2019 vs 2020

Model result should also be tested if it is stable through time. We fit the data of 2020 instead of 2019 to the model and use t-test to see if the TOI of nodes from the two time ranges are similar.

AFF NODE	pvals
node_62	0.127315
node_6	0.000558
node_8	0.011744
node_54	0.017917
node_55	1.32E-05
node_13	0.652307
node_35	8.31E-34
node_61	0.415888
node_48	6.55E-09
node_31	9.66E-05
node_20	5.90E-10
node_39	0.002398
node_18	1.71E-13
node_7	3.85E-17
node_42	0.293024
node_37	0.662524
node_43	3.37E-49
node_12	6.20E-06
node_36	6.70E-05
node_50	0.161667
node_33	5.59E-41
node_21	0.008391
node_10	0.091099
node_41	0.081112
node_44	0.772733

Attention: Only 9/25 nodes pass the t-test (at 10% significant)

3.2 Characteristics groups (nodes) of MAF customers resulted from the tree using train data

	TOI	No of Cust	% Population	Use Mortgage	APE	AVG_BAL_CASA_12M	AUTO_LOAN_AVG_BAL	OTHER_LOAN_AVG_BAL	AVG_ASSET_12M
Note_7	47,382,723	6677	1.36%	Yes		> 1,342,022			
Note_18	40,403,909	2311	0.47%	No	> 40,000,000				
Note_6	35,666,918	2215	0.45%	Yes		317,936 : 1,342,022			
Note_17	30,447,351	2364	0.48%	No	30,010,000 : 40,000,000				
Note_16	25,308,103	9180	1.87%	No	18,000,000 : 30,010,000				
Note_24	25,075,885	2356	0.48%	No	No		>299,499,102		
Note_5	24,620,392	3782	0.77%	Yes		<= 317,936			
Note_34	19,847,058	2816	0.57%	No	No	<= 4,348,707	No	> 24,464	
Note_37	16,197,203	3286	0.67%	No	No		<=299,499,102		> 663571
Note_35	15,701,555	3349	0.68%	No	No	> 4,348,707	No	> 24,464	
Note_36	14,749,299	2441	0.50%	No	No		<=299,499,102		<= 663,571
Note_13	14,489,133	9377	1.91%	No	<= 18,000,000				
Note_33	13,429,549	41862	8.52%	No	No		No	<= 24,464	>198,954,160
Note_32	7,054,112	44180	8.99%	No	No		No	<= 24,464	86,510,758 : 198,954,160
Note_31	4,712,329	45032	9.16%	No	No		No	<= 24,464	42,536,486: 86,510,758
Note_30	3,293,963	45613	9.28%	No	No		No	<= 24,464	22,692,835 : 42,536,486
Note_29	1,842,389	264739	53.85%	No	No		No	<= 24,464	<= 22,692,835

We can observe that
+ High TOI nodes: Use mortgage or high APE (Banca value)
+ Mid TOI nodes: Moderate level of usage in CASA or Loan
+ Low TOI nodes: No mortgage, APE, little or no loan usage

Model Accuracy

Similar method as above was applied

MAF Node	Error
Node 5	62.51%
Node 6	42.55%
Node 29	36.92%
Node 24	27.99%
Node 18	23.59%
Node 36	21.65%
Node 7	20.28%
Node 30	19.12%
Node 37	15.56%
Node 31	15.44%
Node 32	11.21%
Node 33	10.43%
Node 17	9.19%
Node 13	6.29%
Node 16	5.42%
Node 34	3.50%
Node 35	0.35%

Attention: The CHAID MAF Tree have low accuracy when predicting value of node 5 and 6 which are the node of customer using mortgage and node 29 which have the least product usage and also has the biggest number of customers

Overall the model can predict TOI of test group with error of 26%

Model validity when further divide nodes into small age group

Node	No of pairs have pvals >0.1
Node_30	1
Node_33	0
Node_29	1
Node 31	0
Node_32	0
Node_32	0
Node_7	0
Node_37	2
Node_36	3
Node_17	1
Node_13	2
Node_5	0
Node_16	1
Node_18	1
Node_24	3
Node_35	1

Attention: Most of nodes don't have similar distribution between age group

Model validity if data of different time range was used 2019 vs 2020

MAF_NODE	P vals
NODE_29	0
NODE_30	1.75E-99
NODE_32	2.16E-297
NODE_16	9.80E-57
NODE_33	1.13E-25
NODE_7	8.99E-09
NODE_31	0
NODE_37	0.252861199
NODE_18	1.37E-22
NODE_6	1.17E-34
NODE_35	0.138891436
NODE_34	1.99E-16
NODE_5	2.97E-104
NODE_13	2.35E-39
NODE_17	3.82E-17
NODE_24	8.19E-14
NODE_36	0.472705883

Attention: Only 3/17 node pass t-test at 10% significant level

3.3 Characteristics groups (nodes) of MAS customers resulted from the tree using train data

	TOI	No of Cust	% Population	APE	Credit Card	UNSECURED_AVG_BAL	AVG_TD_AMT_12M	AVG_ASSET_12M	AVG_BAL_CASA_12M
Note_7	16,186,848	4,855	0.40%	>19,999,500					
Note_6	10,634,314	4,452	0.37%	13,788,000 : 19,999,500					
Note_9	7,834,351	3,330	0.28%	No	No	>11,986,295			
Note_10	6,425,217	6,593	0.55%	No	Yes				<= 803,799
Note_5	6,099,452	6,634	0.55%	<=13,788,000					
Note_13	5,546,032	3,059	0.25%	No	Yes				>9,285,972
Note_11	5,335,901	3,440	0.29%	No	Yes				803,799 : 2,940,251
Note_12	3,972,885	4,537	0.38%	No	Yes				2,940,251 : 9,285,972
Note_23	3,401,359	44,851	3.72%	No	No	<=11,986,298	No	>= 16,519,694	
Note_27	2,587,207	69,716	5.78%	No	No	<=11,986,302	Yes	>16,519,694	
Note_22	1,316,830	97,689	8.10%	No	No	<=11,986,297	No	6,720,225 : 16,519,694	
Note_21	705,292	108,786	9.02%	No	No	<=11,986,296	No	3,706,851 : 6,720,225	
Note_26	677,113	18,438	1.53%	No	No	<=11,986,301	Yes	6,720,225 : 16,519,694	
Note_20	458,717	813,104	67.41%	No	No	<=11,986,295	No	<= 3,706,851	
Note_25	439,503	8,137	0.67%	No	No	<=11,986,300	Yes	3,706,851 : 6,720,225	
Note_24	379,359	8,581	0.71%	No	No	<=11,986,299	Yes	<= 3,706,851	

+ High TOI Node: High APE, high unsecured average balance
+ Mid TOI Node: Moderate level of Average Casa Bal
+ Low TOI Node: Little or no usage of unsecured average balance

Model Accuracy

MAS Node	Error
Note_20	38.90%
Note 25	8.96%
Note_6	23.99%
Note_26	23.96%
Note_21	20.41%
Note_7	19.63%
Note 27	19.16%
Note 22	17.31%
Note 23	16.73%
Note_24	15.14%
Note_10	12.80%
Note_11	9.32%
Note_13	8.33%
Note_12	6.99%
Note_5	5.12%
Note_9	2.71%

Attention: The CHAID MAS Tree have low accuracy when predicting value of node 20 which are the node of customers using mortgage which have the least product usage and also has the biggest number of customers (over 60% of pop)

Overall the model can predict TOI of test group with error of 14%

Model validity when further divide nodes into small age group

MAS Node	No of pairs have pvals >0.1
Note_20	0
Note 22	0
Note_21	0
Note_25	1
Note_27	1
Note_26	1
Note_9	1
Note_24	1
Note_23	0
Note_12	3
Note_6	2
Note_5	3
Note_10	3
Note_7	1
Note 11	1
Note 13	2

Attention: Most of nodes don't have similar distribution between age group

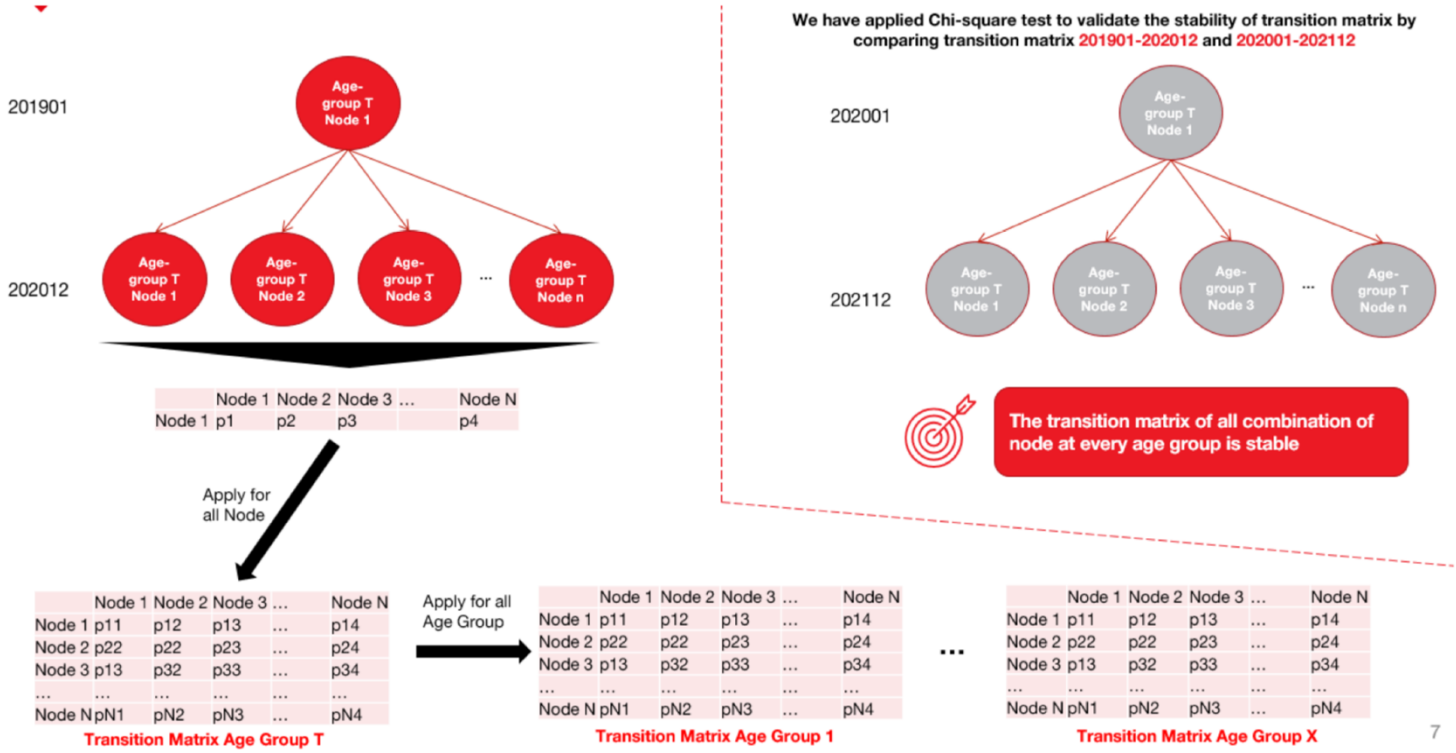
Model validity if data of different time range was used 2019 vs 2020

MAS Node	P vals
9	0.112401278
7	4.47E-10
6	8.71E-50
5	2.44E-43
27	0.821720658
26	1.62E-14
25	6.68E-11
24	1.06E-05
23	1.68E-21
22	1.24E-21
21	3.65E-36
20	6.80E-158
13	0.088235766
12	8.90E-29
11	2.54E-55
10	3.92E-65

Attention: Only 3/15 node pass t-test at 10% significant level

4. Find transition matrix for each age groups (Step 2 in the Methodology Overview)

After finding the rule to create characteristic groups, we can now see movement of customers of a certain age group after 2 years when they are 2 years older, jumping from one characteristic group to another. Age of customer range from 21-60 which then form 20 age groups (21-22,23-24…59-60). Each economic segment (AFF, MAF, MAS) will have their own set of transition matrixes.



Attention: It is assumed that customer movement within nodes have a Markov Property which means all customers at a particular age group and current year T node will move to future nodes at year T+2 based on a same transition matrix regardless of their past journey to reach the current node

5. Calculate CLV based on Transition Matrix using backward induction (Step 3 in the Methodology Overview)

Some more data preparation: Median TOI contribution 2 years from 202001-202112, group by all different age groups and nodes. When combined with transition matrixes of all age groups, we will be able to calculate Expected Customer Value at every age group by multiplying transition matrix of age group T-2 with median TOI of customer of age group T as illustrated in the examples below. In this case we will use a discount rate of 7% which is appropriate for Vietnam. Then we have the CLV of every age group and node using the formular:

$$CLV_T = Custom\ Expected\ Value_T + \frac{Custom\ Expected\ Value_{T+2}}{(1 + Discount\ Rate)^2} + \cdots + \frac{Custom\ Expected\ Value_{60}}{(1 + Discount\ Rate)^{60-T}}$$

T: The age group of customers

Detail examples:

Median total TOI contribution 2y 201901-202012 divided age group and Node

NEW_AGE_GROUP	NODE_6	NODE_7	NODE_8	NODE_10	NODE_12	NODE_13	NODE_18	NODE_20	NODE_21	NODE_31	NODE_33	NODE_35	NO
AGE_21_24	33,001,523	45,362,087	56,161,383	71,031,471	186,908,137	452,832,655	24,210,209	52,834,663	83,557,396	389,252	1,184,976	4,059,133	
AGE_25_26	37,375,134	44,645,882	59,362,096	89,724,479	227,877,458	523,029,282	29,372,960	37,069,435	75,888,387	626,754	1,612,863	4,395,203	
AGE_27_28	39,297,592	46,473,161	55,407,159	95,517,273	194,411,524	465,769,598	23,087,240	40,050,583	92,140,697	737,961	1,849,002	4,740,756	
AGE_29_30	42,622,304	49,666,889	61,056,306	111,694,488	228,857,818	497,142,105	25,017,426	38,832,665	104,181,464	850,382	2,149,398	5,122,388	
AGE_31_32	43,111,370	53,459,230	63,181,306	100,489,323	201,020,740	456,026,794	22,663,351	37,484,662	85,084,074	1,124,704	2,264,641	5,167,727	
AGE_33_34	48,106,161	50,266,477	71,547,828	112,902,404	199,850,844	589,397,449	26,067,591	38,937,668	95,170,688	1,409,880	2,346,228	5,429,008	
AGE_35_36	42,349,346	55,266,947	66,567,261	106,231,283	199,939,947	450,216,726	28,260,100	42,466,097	100,504,493	1,542,248	2,411,126	5,672,776	
AGE_37_38	47,819,399	59,868,113	68,033,412	112,219,073	220,008,012	502,808,390	27,529,186	49,592,804	93,415,609	1,880,983	2,791,242	6,087,898	
AGE_39_40	50,515,528	58,202,597	80,467,085	116,444,747	206,190,499	533,735,005	27,355,029	44,599,934	110,301,112	2,133,168	2,865,773	6,252,510	
AGE_41_42	48,274,539	56,688,280	71,145,064	117,951,225	237,433,867	515,455,967	31,282,538	48,687,553	117,284,790	2,802,548	2,762,022	5,855,672	
AGE_43_44	50,050,653	57,101,428	61,863,829	120,492,687	230,971,378	535,201,505	29,803,640	54,103,445	134,795,511	3,140,603	2,902,957	6,053,581	
AGE_45_46	47,326,685	65,990,393	75,593,403	122,364,613	236,339,253	623,120,136	34,485,461	56,031,636	125,037,480	2,883,914	2,632,812	6,097,458	
AGE_47_48	49,982,140	54,257,087	63,867,941	111,835,700	232,125,477	564,628,654	28,311,799	62,018,875	119,074,922	2,602,881	2,495,184	5,741,951	
AGE_49_50	67,251,328	52,184,687	97,622,468	108,154,870	219,433,169	563,353,538	36,315,689	65,034,913	118,646,409	3,168,310	2,421,032	5,956,121	
AGE_51_52	47,101,936	51,826,454	59,886,597	124,477,192	235,670,092	502,415,308	34,510,491	75,194,692	158,346,176	2,957,710	2,117,585	5,575,674	
AGE_53_54	44,329,230	53,316,807	97,308,411	101,512,779	232,596,034	493,116,905	34,430,434	61,884,787	150,353,852	1,280,529	3,040,206	5,323,736	
AGE_55_60	56,270,413	54,515,762	65,712,411	109,251,583	207,039,758	504,532,688	31,125,043	59,293,004	137,068,998	2,070,825	2,129,557	5,400,449	

Value all node of age group 55_60

Dot Product

= Expected value of Node X at AGE_53_54

X

Apply for all Node within AGE_53_54

Apply similarly for all age group then then we have expected value for all age group and node

Summing up appropriate expected value and applied discount rate = 7% per year depended on the age group

Probability Node X move to other node after 2y

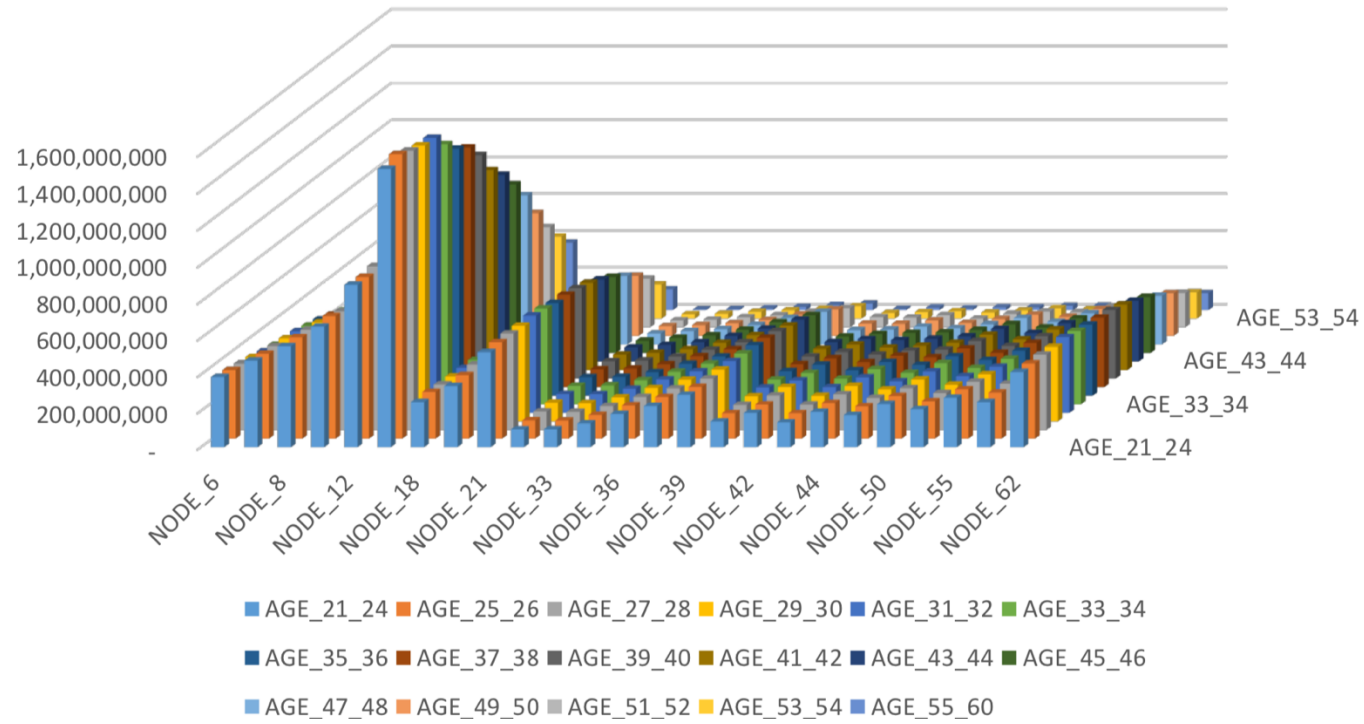
Future Node

Transition Matrix Age Group 53-54

Current Node

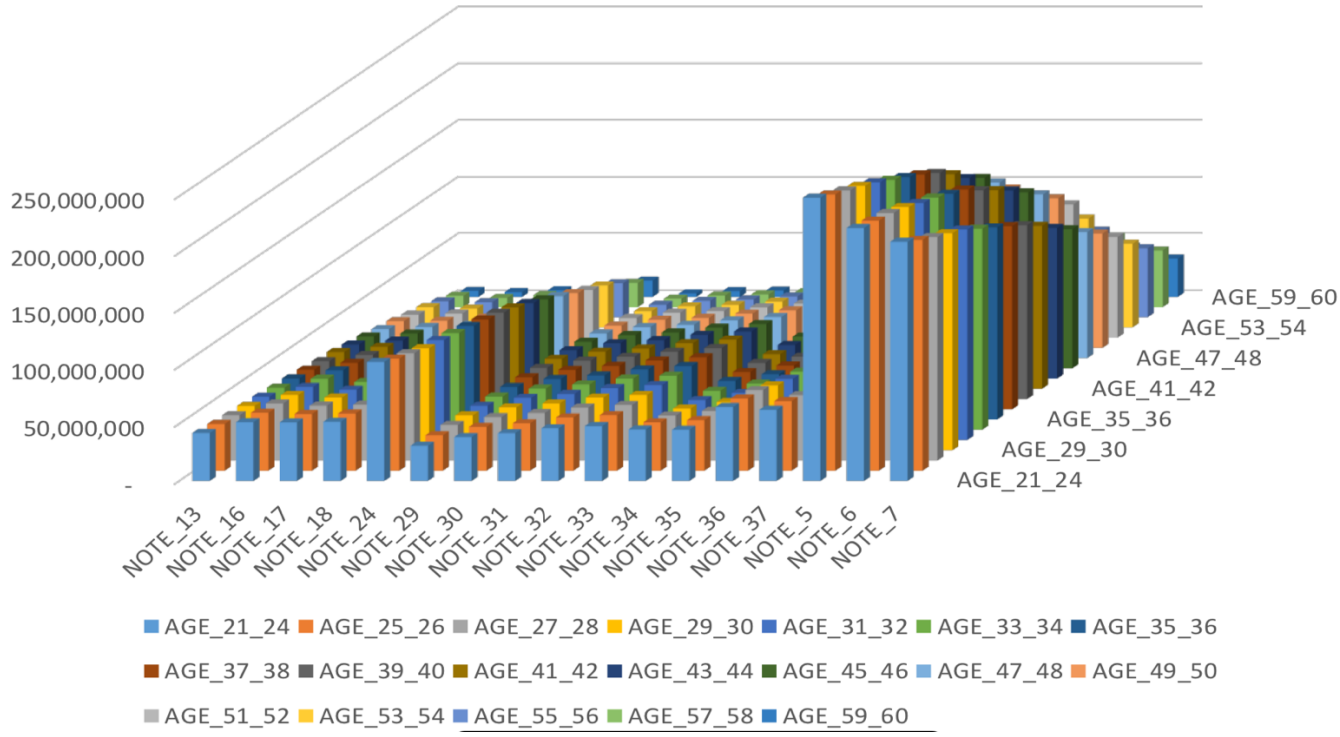
AGE_GROUP	NODE_2018	NODE_6	NODE_7	NODE_8	NODE_10	NODE_12	NODE_13	NODE_18	NODE_20	NODE_21	NODE_31	NODE_33	NODE_35	NODE_36	NODE_37	NODE_39	NODE_
AGE_53_54	node_20	0	0.02	0	0.01	0	0.02	0.13	0.41	0.21	0.02	0	0	0.02	0	0.03	
AGE_53_54	node_21	0	0	0	0	0.05	0.01	0.02	0.11	0.52	0.02	0	0	0.02	0.02	0.04	0.
AGE_53_54	node_31	0.01	0	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.51	0.14	0.06	0.03	0.04	0.01	
AGE_53_54	node_33	0.01	0.01	0	0.01	0	0.01	0.01	0.02	0.01	0.1	0.25	0.23	0.07	0.04	0.05	
AGE_53_54	node_35	0.01	0	0	0.01	0.01	0.01	0.01	0.02	0.02	0.06	0.09	0.19	0.16	0.1	0.08	
AGE_53_54	node_36	0	0	0	0	0.02	0.01	0.01	0.04	0.01	0.03	0.05	0.23	0.13	0.18	0.16	
AGE_53_54	node_37	0	0	0	0	0	0	0.02	0.02	0.01	0.03	0.03	0.08	0.14	0.18	0.31	
AGE_53_54	node_39	0.01	0.01	0	0	0	0	0.06	0.01	0.05	0.04	0.02	0.03	0.04	0.13	0.43	0.
AGE_53_54	node_41	0	0	0	0.02	0.07	0	0.02	0	0	0.2	0.02	0	0.02	0	0.07	0.
AGE_53_54	node_42	0	0	0	0.02	0	0.02	0.02	0.02	0	0.27	0.04	0.04	0.02	0.02	0	
AGE_53_54	node_43	0.01	0	0.01	0	0.01	0.02	0.05	0.04	0.02	0.04	0.05	0.12	0.05	0.03	0.03	
AGE_53_54	node_44	0	0	0	0	0	0	0	0.15	0	0.08	0	0.08	0	0	0	0.
AGE_53_54	node_48	0.01	0	0	0.01	0.02	0	0.07	0.06	0.03	0.04	0.04	0.04	0.05	0.01	0.04	0.
AGE_53_54	node_50	0	0.02	0	0	0	0	0.02	0.02	0.06	0	0.02	0.06	0.05	0.15	0.1	0.
AGE_53_54	node_54	0	0	0	0.01	0	0	0.13	0.06	0.02	0.03	0.03	0.02	0.02	0.01	0.04	0.
AGE_53_54	node_55	0	0.01	0	0.01	0	0	0.02	0.07	0.07	0.01	0.01	0.04	0.07	0.04	0.07	
AGE_53_54	node_61	0	0	0	0	0.01	0	0.1	0.05	0.03	0.12	0.02	0.02	0.01	0.02	0.03	0.
AGE_53_54	node_62	0	0	0.01	0.01	0.02	0.01	0.01	0.02	0.07	0.02	0.02	0.02	0.01	0.07	0.1	0.

AFF segment CLV results

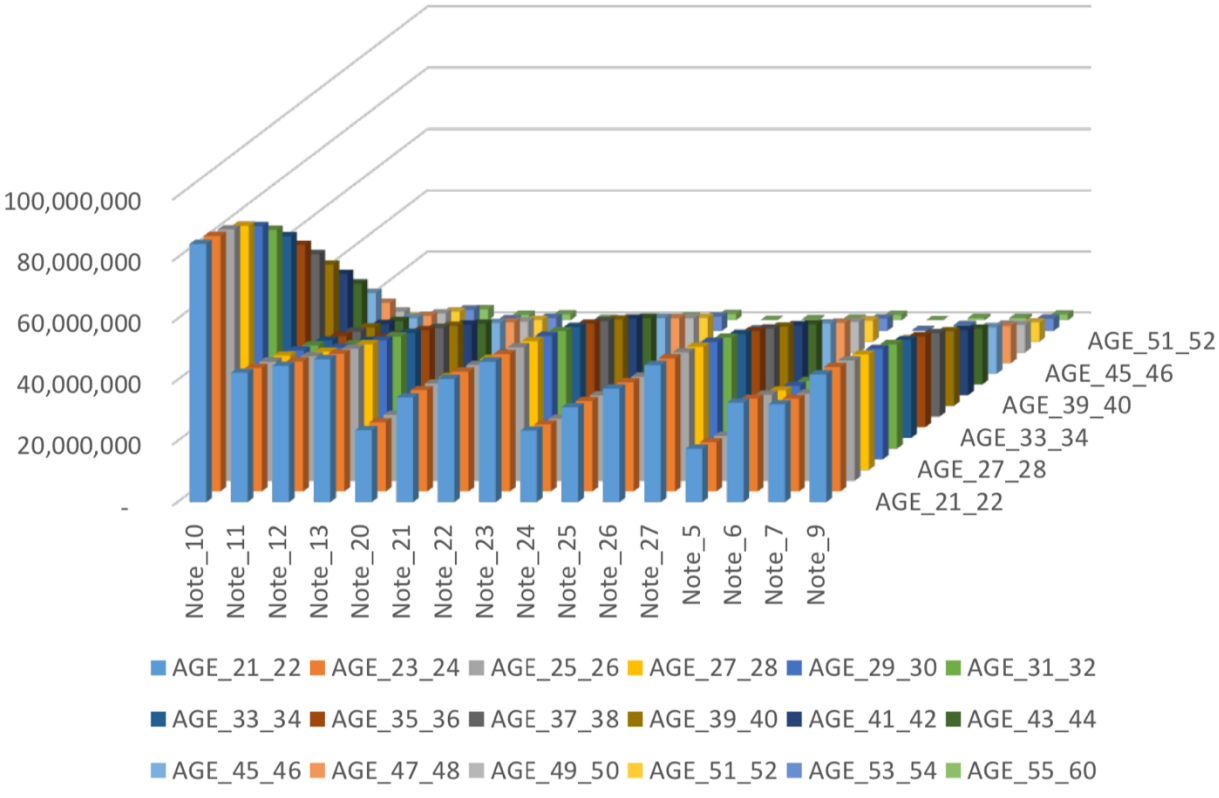


We observe that for most node, the group that have highest CLV seems to be from 29-32 years old. This is the age range when customers have relatively high current value and still have many years in the futures to contribute value to the banks.

MAF segment CLV results



MAS segment CLV results



6. CLV Application

After the CLV is calculated for all customer, there are two main usages we can follow:
+ Focus on high CLV customer, either by looking at the current customer base or looking for them outside of the bank.
+ Finding pathway to increase CLV of customers.
+ It should be aware that the calculation method does not pass many of the test and Markov property of customers is a very big assumption.