INFO 251

UC BERKELEY, USA – DECEMBER 2017

"Customer attrition prediction model: A bank in Chile, South America"

RIVERS JENKINS, CRISTOBAL PAIS, YIYU SHI

Berkeley UNIVERSITY OF CALIFORNIA

UC Berkeley School of Information

Agenda

- Problem Description
- Main Objectives
- Methodology
- Results and Discussion
- Conclusions

The **Problem**: customer attrition

- **ABC, a real bank located in Chile, has a problem related to customers retention.**
- **10% of the total portfolio finish its contract with the bank annually.**
- **Estimations: around 30% of their total portfolio may potentially finish their contract.**
- **Resources: to contact and persuade about 5% of its total portfolio.**

Objectives: **keep my clients!**

- **"Develop an efficient and effective customer attrition prediction model for a real bank in South America, helping the bank to focus its resources for its retention actions/policies".**

The Instance: **ABC bank** dataset

Two Data sets
- DS1: 1248 observations
- DS2: 2807 Observations
- 17 Features (1 ID)
- 1 Binary Label

Table 1: Attributes in Data Base

Cod_cliente	Client Identification
COD_OFI	Office code
COM	Province
ED	Age in years
SX	Sex/Genre (M: men; F: women)
NIV_EDUC	Educational level
RENTA	Rent in thousand of pesos by month
E_CIVIL	Civil state
VIG	Validity (in month)
TRX_T	Number of transactions in month T
TRX_T-1	Number of transactions in month T-1
TRX_T-2	Number of transactions in month T-2
SALDO_T	Average balance in CL pesos in month T
SALDO_T-1	Average balance in CL pesos in month T-1
SALDO_T-2	Average balance in CL pesos month T-2
SALDO_T-3	Average balance in CL pesos month T-3
SALDO_T-4	Average balance in CL pesos month T-4
CERRO	Indicator if the person closed (1) or not (0) his account

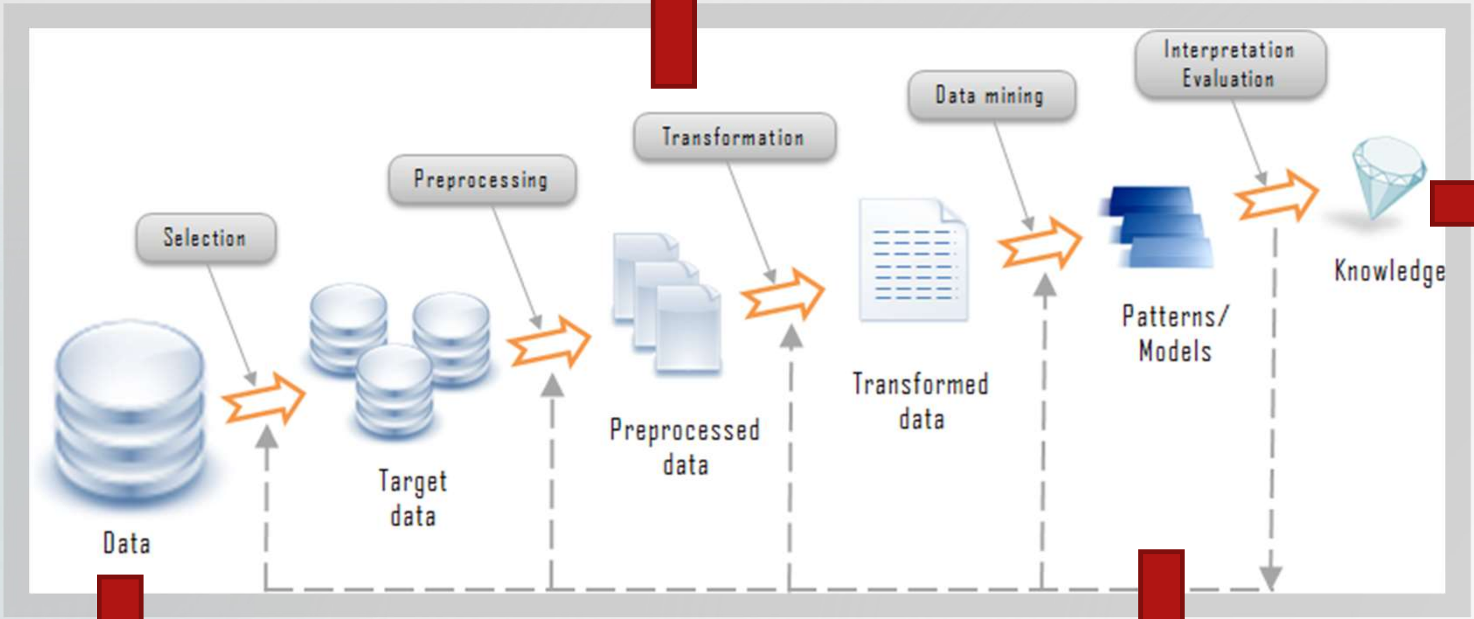
Demographic variables
- Numerical & categorical variables (strings)
- 9 missing values: 8 NIV_EDUC & 1 SX

Bank variables
- Wide ranges
- Different scale
- No missing values
- Numerical

Binary Label

Methodology: when the accuracy is not everything…

String to numerical, missing values, standardization



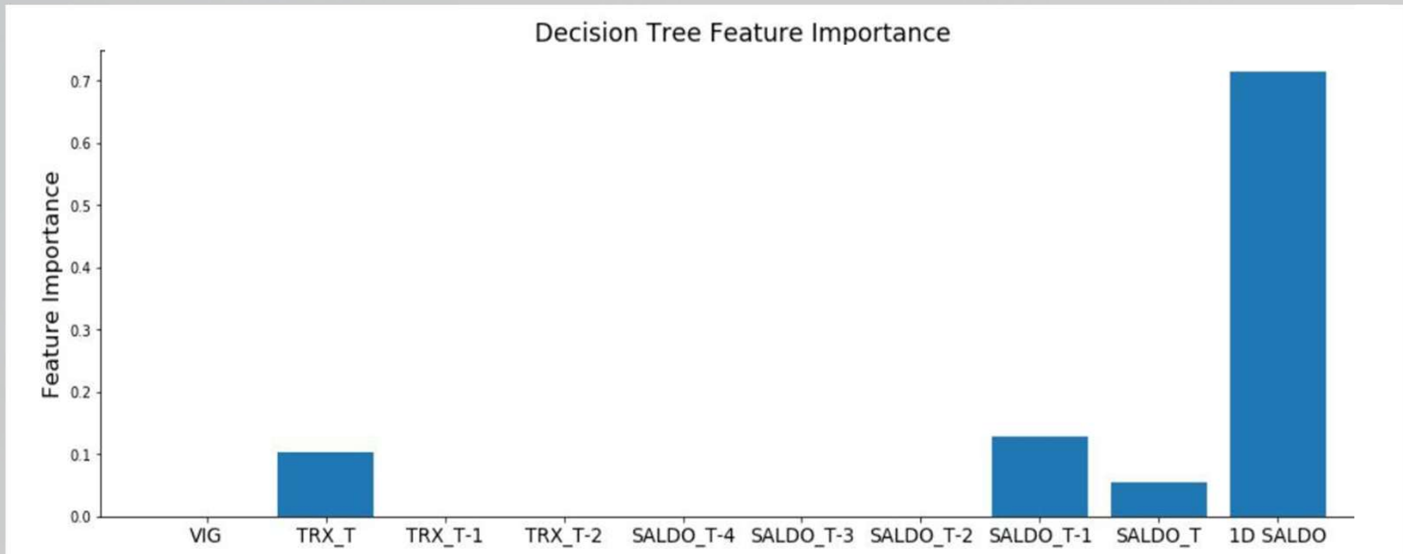
List of clients to contact

Preliminary analysis

Trade-off: Accuracy vs cost
Lift charts

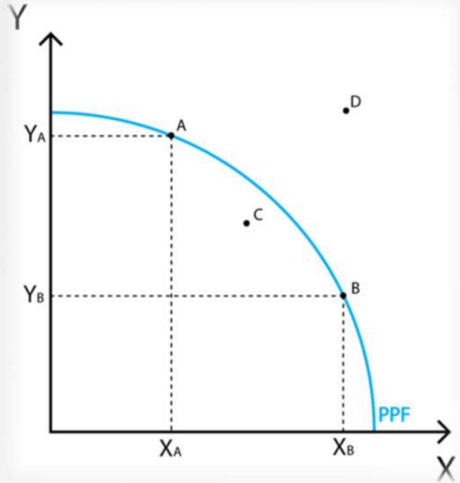
The Instance: **Preliminary analysis**

- Correlation with label (strongest)
  - TRX_T $\rho = -0.39$
  - SALDOT −1 $\rho = -0.32$
  - TRX_T −1 $\rho = -0.26$
- Feature Engineering
  - 1D SALDO
  - MEAN SALDO
  - 1D TRX
  - PCA analysis (3 components)

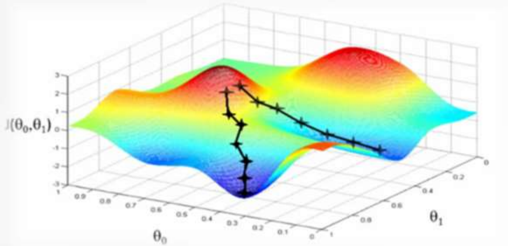


Methodology: when the accuracy is not everything…

- A straightforward accuracy score does not Paint the whole picture.
- A customer predicted to stay that actually leaves is a more costly error.
- The models allow for class weights.



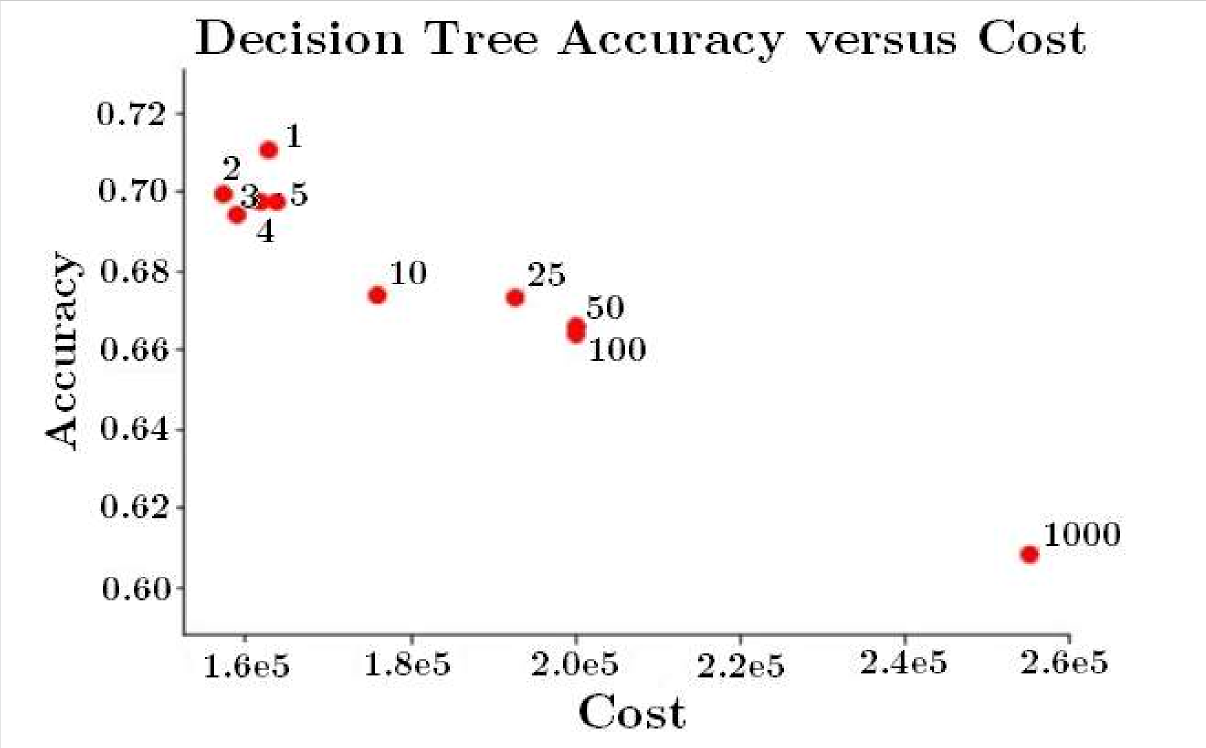
Cost Matrix	Predicted: Stay	Predicted: Leave
Actual: Stay	TN: 0	FP: 350
Actual: Leave	FN: 1000	TP: 250



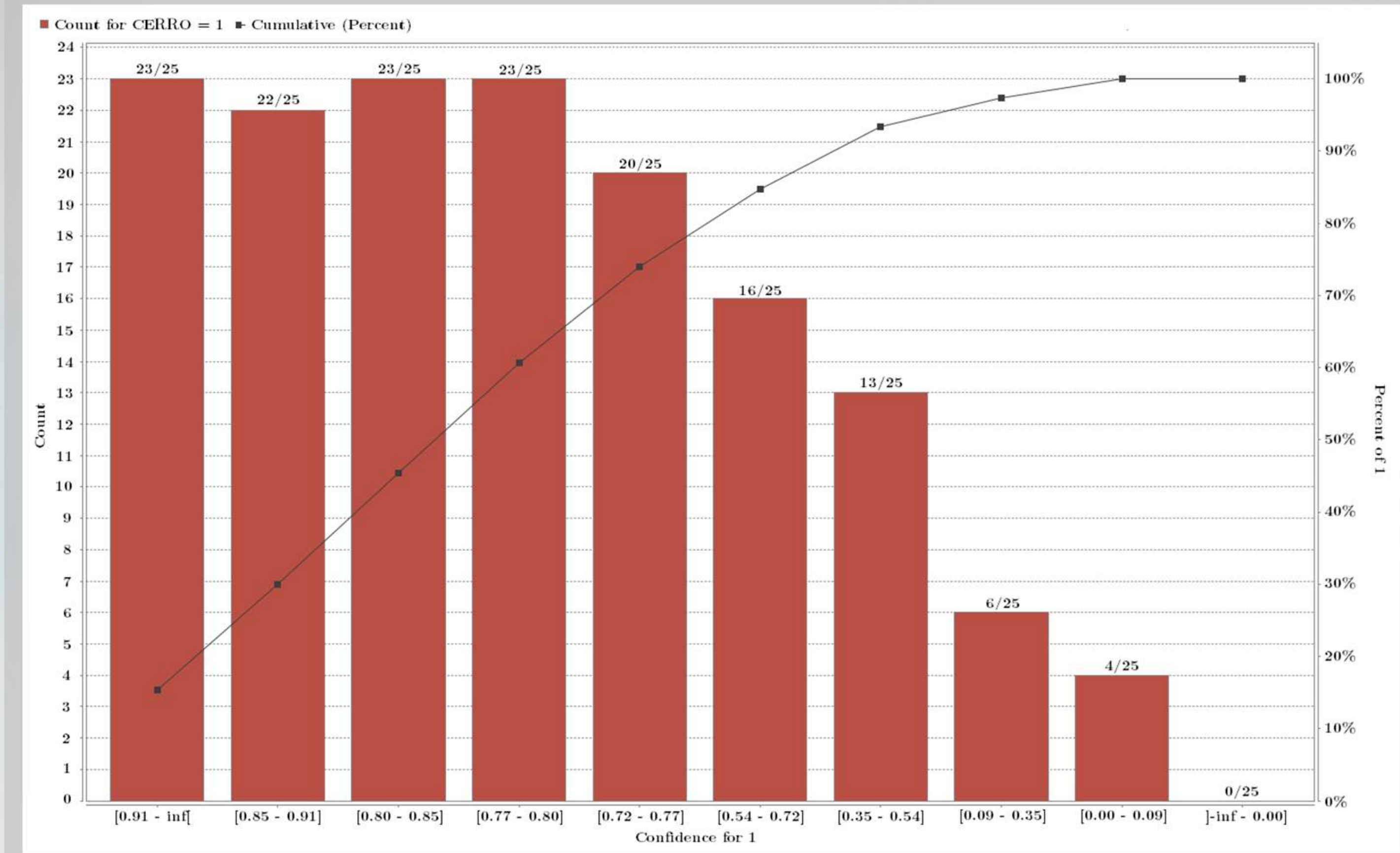
Results and **Discussion:** Models comparison

Model	Decision Tree	Random Forest	Logistic Regression	SVM	Neural Network	K-NN
Training Accuracy	84.9%	88.1%	83.4%	85.3%	80.8%	71.2%
Testing Accuracy	77.6%	83.1%	86.1%	76.5%	73.8%	61.2%
TP	109	116	114	118	87	67
TN	82	90	91	73	96	85
FP	45	37	36	27	40	60
FN	12	5	7	32	25	36
Weight	2	5	4	3	-	-
GAP Cost	6.3%	10.5%	17.4%	2.32%	-	-

Results and **Discussion**: Cost vs Accuracy



Results and **Discussion**: Lift charts



- **The best performance is obtained by the logistic regression model.**
- **Only 9 observations are not well classified among the first 100 classified as abandoning clients, in other words, an error of 9.00%**
- **Customers that are "most likely" to reflect a positive response are identified.**

Results and **Discussion**: Error analysis and interpretation

Classification	AVG SALDO_T	AVG SALDO_T-1	AVG TRX_T	AVG TRX_T-1	AVG Renta	AVG VIG
TP	-0.072	-0.310	20.9	26.9	660.9	43.6
FN	-0.095	-0.160	61.0	42.1	1219.8	76.3
TF	0.261	0.522	62.0	54.9	743.3	73.4
FP	-0.276	-0.318	21.4	22.7	681.2	39.7

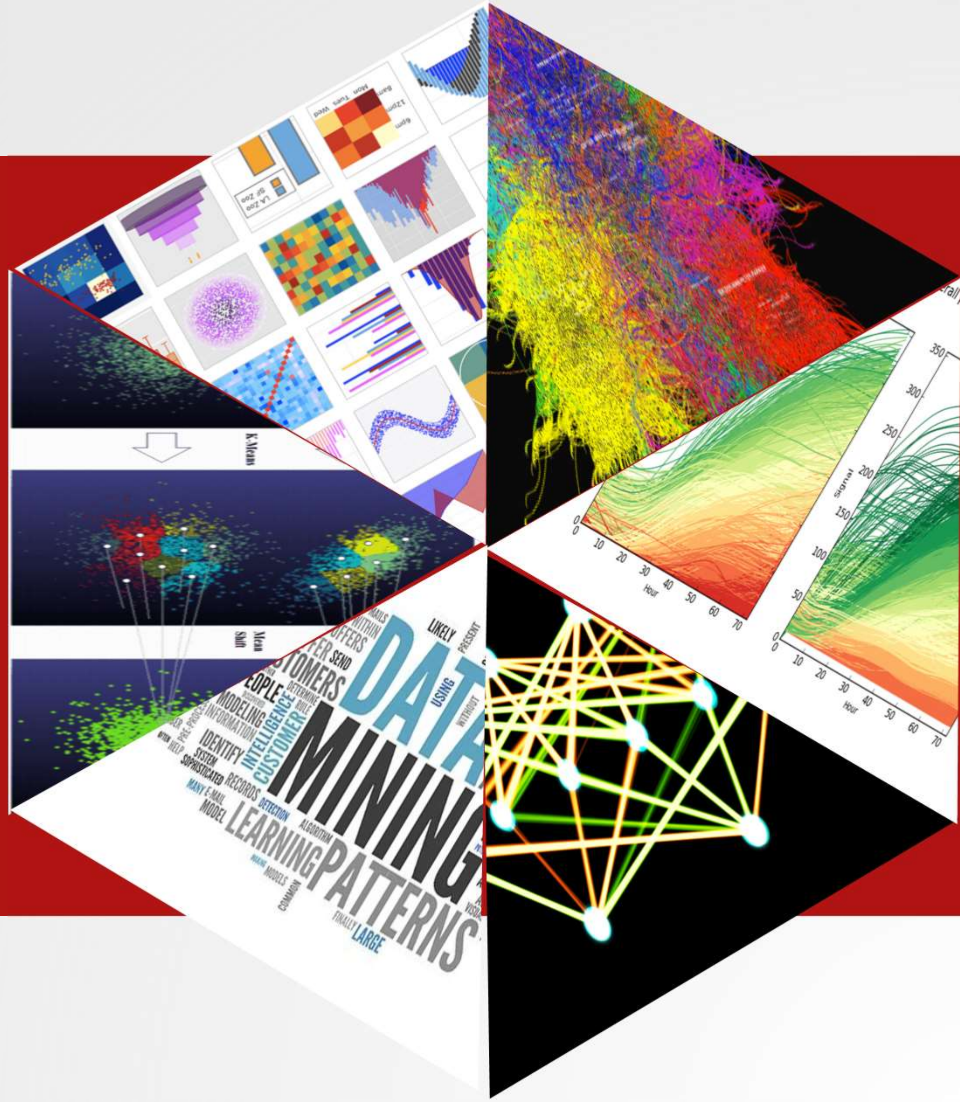
- FN are misclassified due to differences in the number of transactions: 2 types of customer behaviors when closing the account.
- FP observations associated with customers having a loan with the bank.

Conclusions

- The study was successful: tackle a real-world problem while using different approaches from the ones applied during the course.
- Trade-off: complexity of the model vs easy to follow decision rules (Logit vs Decision tree).
- Information about the effectiveness of the offers/MKT campaigns is needed: optimal thresholds of the models.
- Real life implementation: The next challenge of the project consists of implementing it

Thanks for your attention!

Rivers Jenkins
Cristóbal Pais
Yiyu Shi



Berkeley
UNIVERSITY OF CALIFORNIA

UC Berkeley
School of Information

INFO 251

UC BERKELEY, USA – DECEMBER 2017

“Customer attrition prediction model: A bank in Chile, South America”

RIVERS JENKINS, CRISTOBAL PAIS, YIYU-SHI

Berkeley UNIVERSITY OF CALIFORNIA

UC Berkeley School of Information

Appendix: Data Set summary

Table 2: ABC Bank data set summary

Attribute	Variable	Type	Mean/Mode	Range and Frequencies	Missing
Id	Cod_Cliente [ID]	integer	avg = 624.500 +/- 360.411	[1.000 ; 1248.000]	0.0
Label	CERRO [LABEL]	binominal	mode = 0.0 (629) least = 1.0 (619)	0.0 (629)- 1.0 (619)	0.0
Regular	COD_OFI	integer	avg = 74.268 +/- 49.723	[10.000 ; 247.000]	0.0
Regular	COM	integer	avg = 122.481 +/- 91.050	[1.000 ; 516.000]	0.0
Regular	ED [years]	integer	avg = 41.319 +/- 12.013	[21.000 ; 86.000]	0.0
Regular	SX [M male,F women]	binominal	mode = M (839) least = F (408)	M (839)- F (408)	1.0
Regular	NIV_EDUC	nominal	mode = UNV (525) least = BAS (5)	UNV (525), MED (298), TEC(400), BAS (5), EUN (12)	8.0
Regular	RENTA [M CLP/month]	integer	avg = 729.071 +/- 514.604	[250.000 ; 5400.000]	0.0
Regular	E_CIVIL	nominal	mode = CAS (822), least = VIU (21)	CAS (822), SOL (348), VIU (21)SEP (57)	0.0
Regular	VIG [months]	integer	avg = 55.654 +/- 53.425	[3.000 ; 366.000]	0.0
Regular	TRX_T [N°/month]	integer	avg = 38.632 +/- 36.569	[0.000 ; 417.000]	0.0
Regular	TRX_T-1 [N°/month]	integer	avg = 37.557 +/- 34.400	[0.000 ; 390.000]	0.0
Regular	TRX_T-2 [N°/month]	integer	avg = 43.248 +/- 38.089	[0.000 ; 391.000]	0.0
Regular	SALDO_T-4 [$\overline{CLP}$/month]	numeric	avg = 1174103.212 +/- 3481175.370	[-568443.750 ; 60791570.550]	0.0
Regular	SALDO_T-3 [$\overline{CLP}$/month]	numeric	avg = 1148645.272 +/- 3298359.403	[-556992.860 ; 70450696.350]	0.0
Regular	SALDO_T-2 [$\overline{CLP}$/month]	numeric	avg = 1235826.283 +/-3016768.011	[-757777.710 ; 35383283.950]	0.0
Regular	SALDO_T-1 [$\overline{CLP}$/month]	numeric	avg = 3084470.349 +/- 8149275.790	[-1359252.860 ; 98154458.100]	0.0
Regular	SALDO_T [$\overline{CLP}$/month]	numeric	avg = 4327925.344 +/-11252574.346	[-3814728.100;159122938.570]	0.0

Appendix: Feature importance tests

- **Covariance and Chi-Square tests**

Table 4: Covariance Test

Variable	COV
TRX_T	-0.19688583
SALDO_T	-0.16113672
TRX_T-1	-0.13091134
SALDO_T-2	-0.10996098
TRX_T-2	-0.10332789
SALDO_T-4	-0.09200732
VIG	-0.09049295
SALDO_T-3	-0.08612740
SALDO_T	-0.06300822
ED	-0.06297507

Table 5: Chi-Square Test

Variable	Squared-Chi
TRX_T	1
SALDO_T-1	0.607376208
TRX_T-1	0.545469436
SALDO_T-2	0.316205238
TRX_T-2	0.256617819
VIG	0.253020294
ED	0.250801476
SALDO_T	0.179847769

- **Gini index analysis**

Table 6: Gini index test 1

Variable	Gini index
SALDO_T-1	1
SALDO_T-2	0.540096249
TRX_T	0.527380709
SALDO_T-3	0.39007159
SALDO_T-4	0.369913197
TRX_T-1	0.273530545
SALDO_T	0.251854852
TRX_T-2	0.166606362

Table 7: Gini index test 2

Variable	Gini Index
VIG	0.09999627
ED	0.06936445
COM	0.01351656
COD_OFI	0.01148478
E_CIVIL	0.00928331
NIV_EDUC	0.0081042
RENTA	0.00741076
SX	0

Customers lifetime & Profit

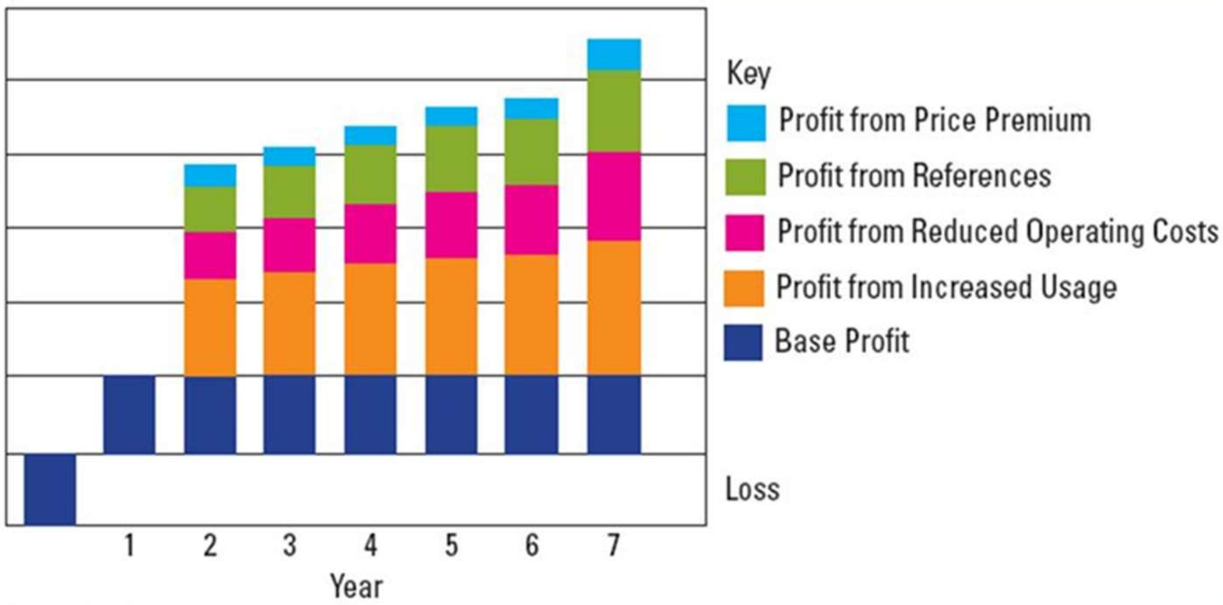


Figure : Why customers become profitable over time.

SOURCE

Reprinted by permission of *Harvard Business Review*. From Zero defections: quality comes from services. By Reichheld, F.R. & Sasser Jr., W.E.(September-October), p.108. Copyright © 1990 by the Harvard Business School Publishing Corporation; all rights reserved.

- New client is more risky than an "old" client
- 5 – 6 times more expensive than actually keeping an old client
- Total utility from a client is proportional to his/her lifetime