Matrix CPM SOLUTIONS

R Ladies

Cross selling analytic model with Sparklyr

Carenne Ludeña

Vienna, April 29, 2021

About us

Matrix CPM Solutions specializes in data based solutions. We are certified partners of Pentaho, Tableau, Vertica, Cloudera, AWS and IBM with experience in support, consulting and training in all of Latin America.

Experience

15+ years

Offices

Caracas, Bogotá and CDMX

Experts

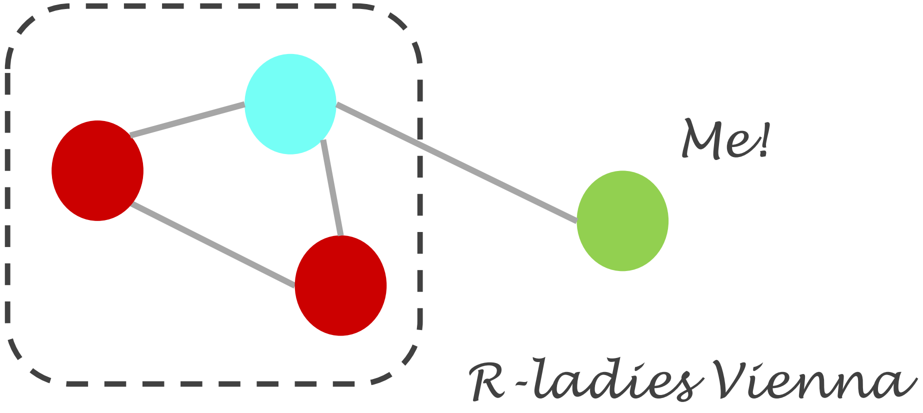
Certified consultants in BI and Data Science

Projects

- BI, Big Data and Data Science
- Training

About me

I am a mathematician, statistician, data scientist (sort of in that order).
I lived most of my life in Venezuela and am now in Colombia, after some years in Ecuador. I worked with academia for many years doing research in mathematical statistics (and teaching R!) until changing to my current role. My interests include statistics, ML, DL, data mining, text mining, graphs-networks and in general analyzing data.



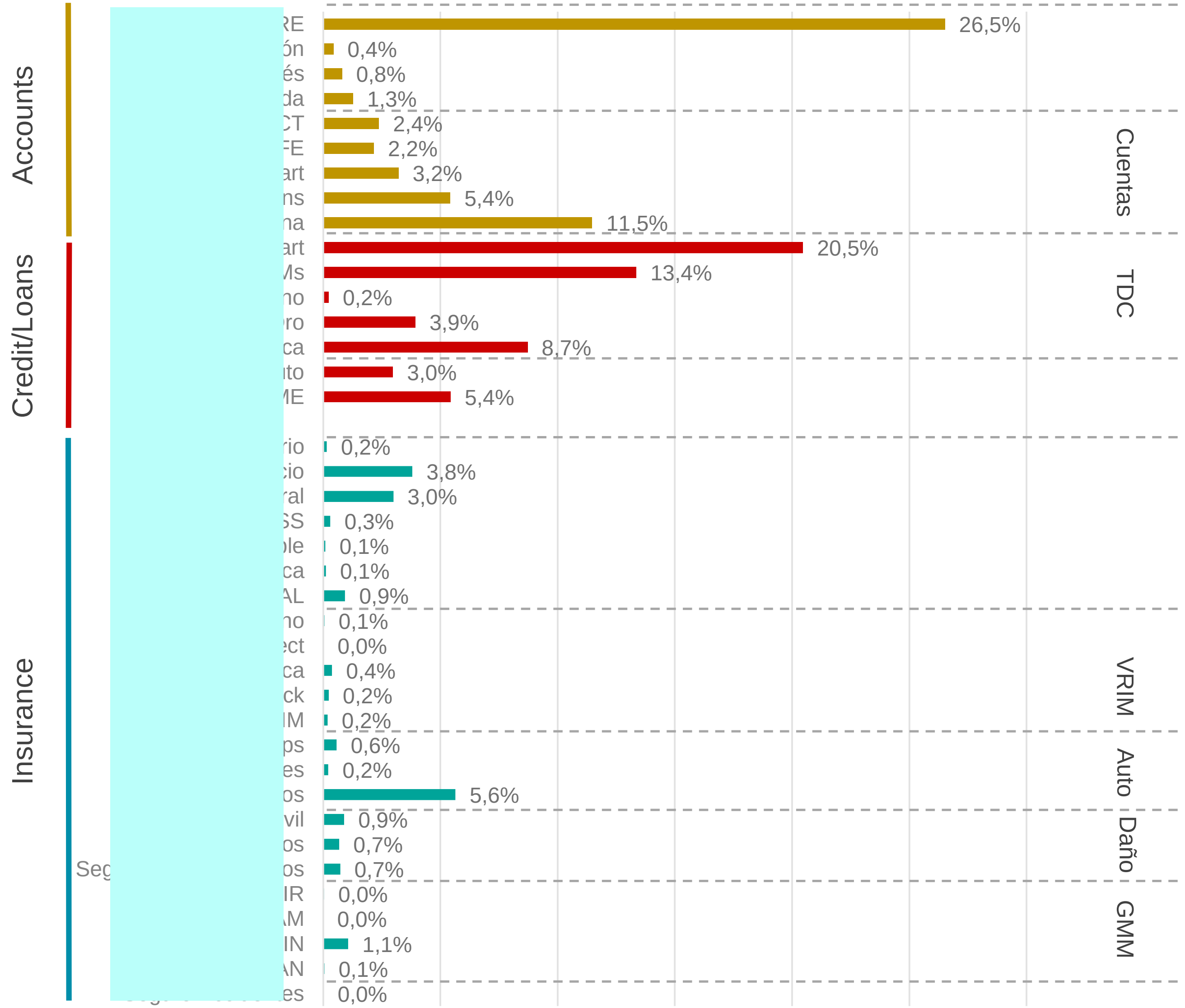
Agenda

- BUSINESS CASE
- SOLUTION
  - MODELS
  - TECH SPECS

BUSINESS CASE

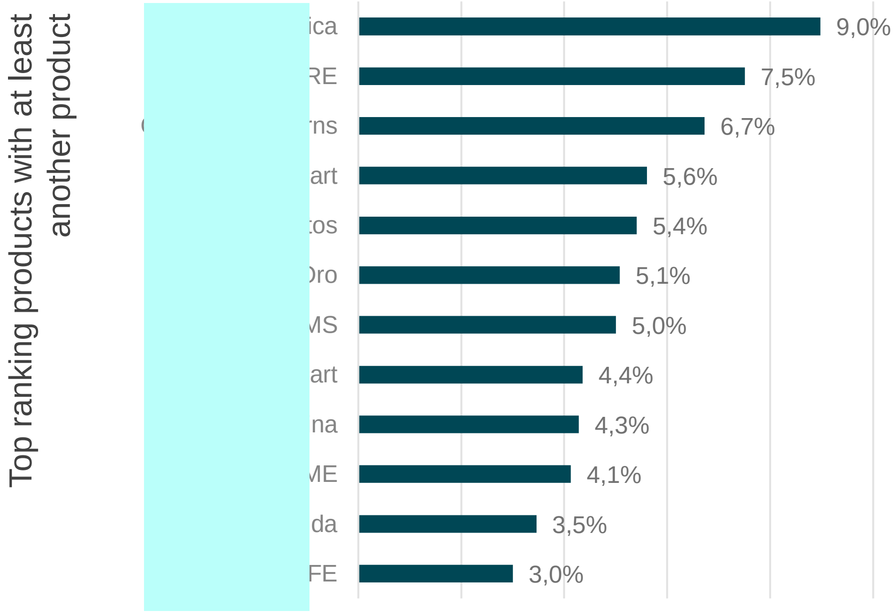
- Top 5 bank in Mexico
- Goal:
  - Big Data project
  - Early win analytics case
  - Most clients: only one product with bank
  - Optimize selling strategy: based on agents

Basic data



Total clients ~ 4 million

Average products per client 1.34



Solution

So…we proposed

- Determine best new products to offer based on client data and bank policies.
- Determine best selling strategies: Rank agents according to best offer and assigned clients.

Tech Specs

- Hadoop (Hortonworks)
- ETL based on Pentaho
- Spark
  - Spark SQL
  - Spark ML
- R

Business specs

- Increase sales
- Empower recently created Big Data/Analytics group
- Joint creation Project (knowledge transfer)

Roadmap

(1)Choose “campaigns” : best choices of have A offer B that will also have a significant impact for the Bank.

(2)Choose clients with higher probability of being interested in such an offer

(3)Understand clients needs: cluster and profile

(4)Optimize campaign strategy: rank agents

Frequent Patterns (basket analytics)

Model

Rules

antecedent A => consequent B

Based on:

Support: frequency of A and B

p(A and B)

Confidence: conditional prob of B given A

p(A and B) / p(A)

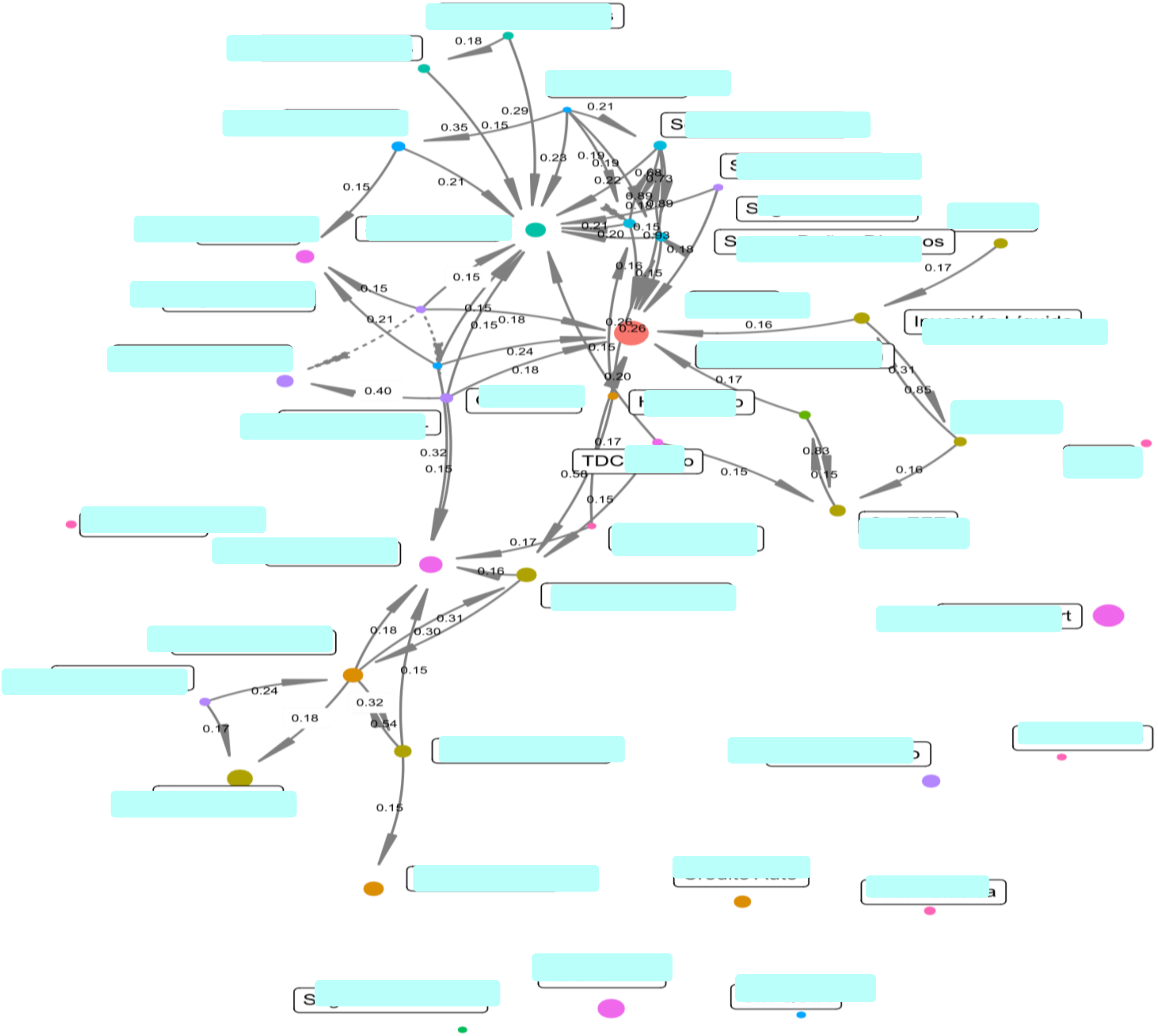
Lift: How much more as related to B

p(A and B) / (p(A) p(B))

Frequent patterns to define campaigns

- FPGrowth algorithm over 4 million client data base
- Frequent patterns over
  - Accounts
  - Insurance
  - Credits and loans
- High support and confidence rules were selected
- Correlation graphs help understand patterns

Correlation graphs



- Small or non existent correlations
- Very few clusters
- CC, Car insurance and a savings account are some of the products with higher frequency that relate to other products

- Other products show a high (unused) potential: supermarket card

Opportunities based on conditional probs

- Green points: easy to choose a good prospect/less available prospects
- Red points: great opportunity if good prospect (harder to find)/ more available prospects

	Auto	Clásica		Platino		Walmart	Nómina	Sanborns	Walmart	Líquida				Inversión	
tes															-1
AN		1	1		-1	-1		1		1	1				
RIN		1	1	1	-1	-1	-1	1	-1	1	1	1		1	-1
AM															
UIR			1		-1	-1		1							-1
sos		1	1	1	-1	-1	-1	1		1	1	1		1	-1
dios		1	1	1	-1	-1	-1	1		1	1	1		1	-1
Civil	1	1	1	1	-1	-1		1		1	1	1		1	-1
tos	-1	1	1	1	-1	-1	-1	1	-1	1	1	1	1	1	-1
hes		-1	1		-1	-1					1				-1
ups			1		-1	-1	-1	1			1	1			-1
RIM					-1	-1	-1								-1
ack		-1			-1	-1	-1								-1
ica		-1			-1	-1	-1								-1
ect		1													
ino					-1	-1	-1								-1
TAL		1	1	1	-1	-1	-1	1	-1	1	1	1	1	1	-1
uca		1	1	1	-1	-1	-1	1		1	1	1		1	-1
ible		1	1	1	-1	-1	-1	1		1	1	1		1	-1
SS		-1			-1	-1	1								-1
oral	-1	-1	-1		-1	-1	-1	-1	-1						-1
icio	-1	-1	-1		-1	-1	-1	-1	-1	-1	-1				-1
ario		1	1	1	-1	-1	-1	1	1	1	1	1			-1
ina		1	1		-1	-1	1	1	1	1	-1	-1			-1
ME	-1	1	1		-1	-1	-1	1	1	-1	-1	-1			-1
uto		-1	-1		-1	-1	-1	-1	-1	-1	-1				-1
ica	-1		-1		-1	-1	-1	1	1	1	1				-1
Oro	-1	-1		1	-1	-1	-1	1	1	1	1	1		1	-1

Opportunities based on conditional probs

Could want

If already has

	Hipo	Nomina	PYME	Auto					Walmart
IN	346	991	799		4,841	5,527	719	1,731	915
Seguro… os	1,515		1,006		2,875	2,689	495	1,243	872
Se… os	1,502		999		2,745	2,603	484	1,209	855
Se… vil	393	1,213	1,160	3,024	3,707	2,938	452	1,582	1,212
os		5,673	3,851	2,670	23,170	17,215	1,364	9,421	6,938
Se… es					453	507		379	171
ps		426	458			1,792		979	501
AL	283		876		4,773	3,882	447	1,789	1,106
SS		2,192			164			64	128
Se… ral		4,328	2,106	879	4,834	3,370		2,747	1,899
tio			1,273	867	3,097	1,957		2,698	2,617

For each combination the number of potential clients was estimated weighing by opportunities (green and red points) : five campaigns were selected.

Campaigns

Data

- Demographics
- Transactions
- Product groups
  - Car insurance
  - Life insurance
  - Damage insurance
  - Accident insurance
  - Accounts
  - Cards
  - Housing loans
  - Personal loans

A campaign is selected if....

- It has a high business impact
- High support and confidence
- High opportunity impact based on conditional probabilities

Which clients are more receptive to a given campaign?

- Are all clients equal?
- Very unbalanced classes
- Cluster & classify

Selecting potential clients

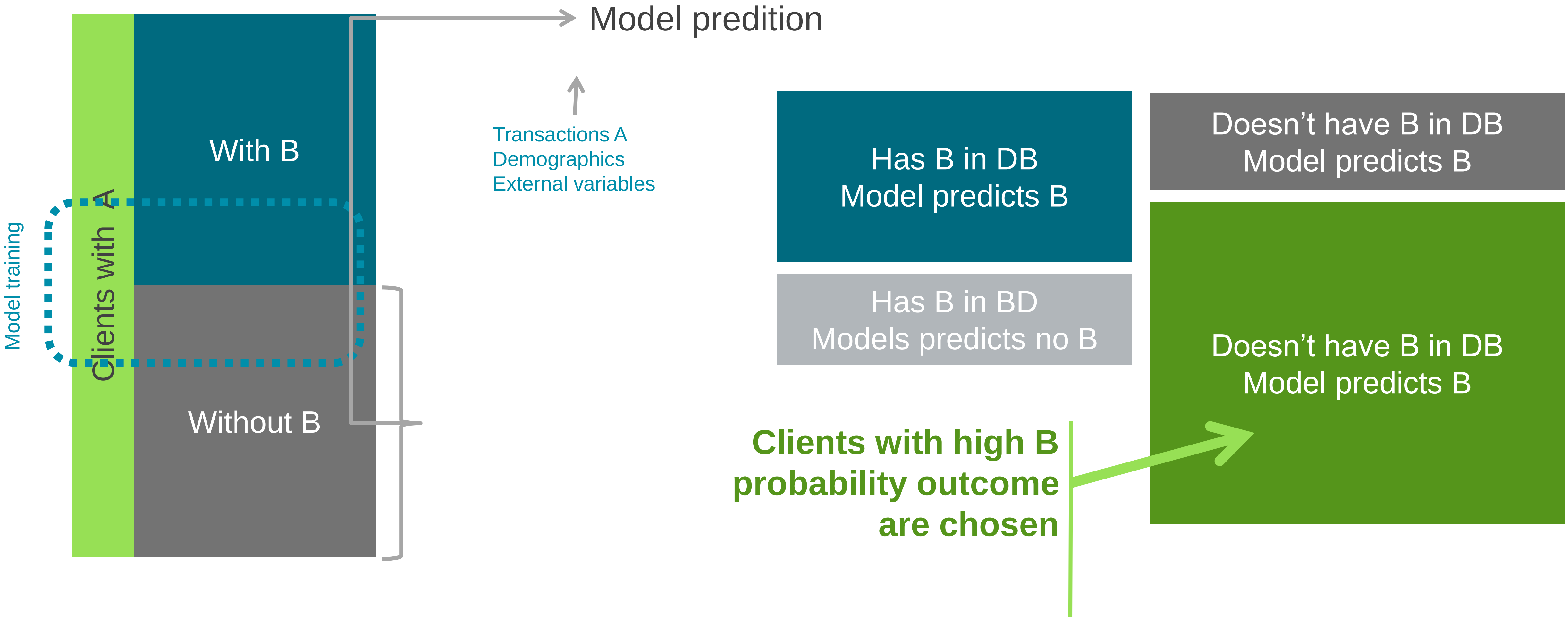
Data

- Demographics
- Transactions
- Product groups
    - Car insurance
    - Life insurance
    - Damage insurance
    - Accident insurance
    - Accounts
    - Cards
    - Housing loans
    - Personal loans

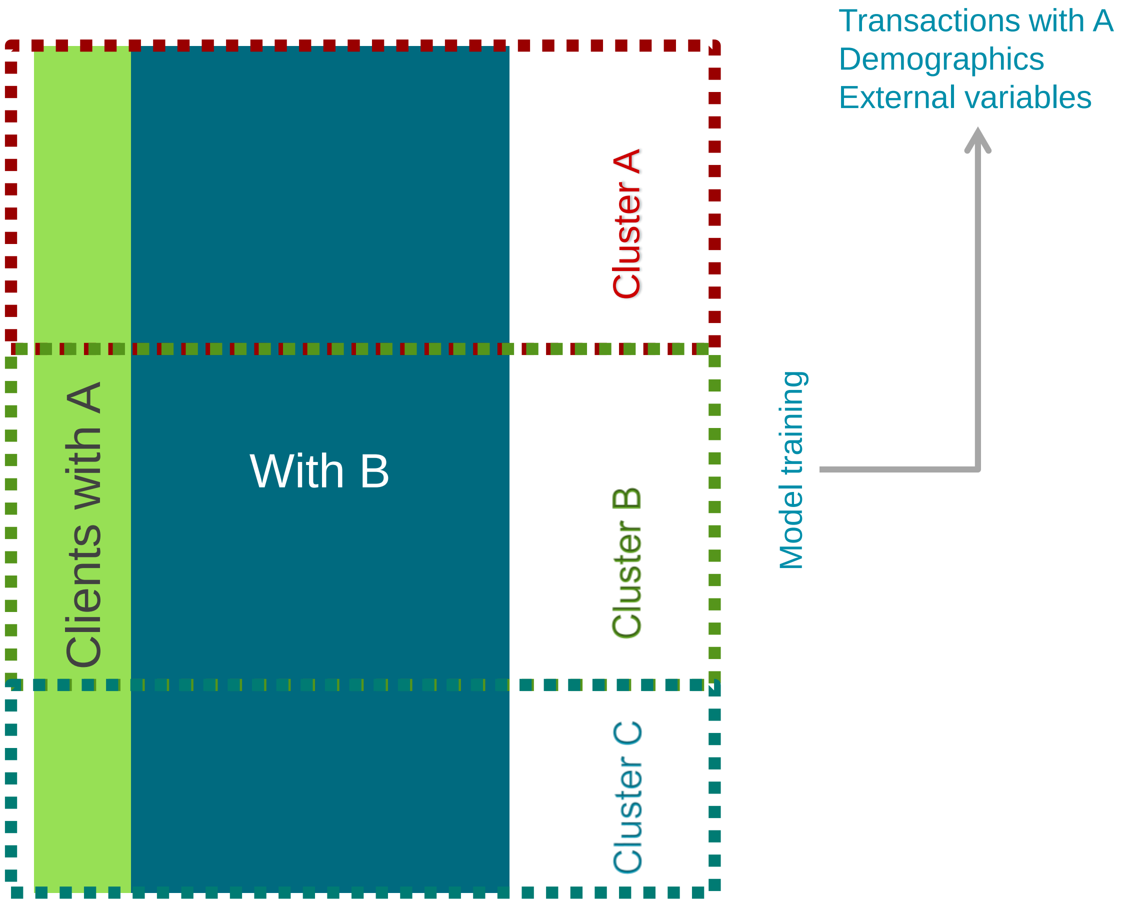
Potential clients

- Selection based on following criteria
    - High probability of offer acceptance.
    - Client profile
    - Use profile info to improve offers: credit cards, loans, insurance risk.

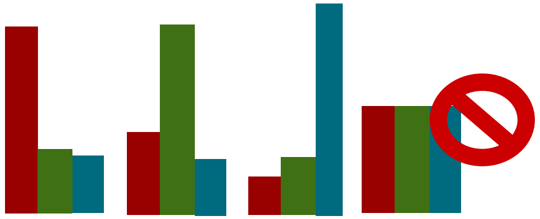
Step 1- classification



Step 2- clustering & classification



Clusters are associated to client profiles in each group



Clients with high probability outcome of belonging to one of the clusters are chosen

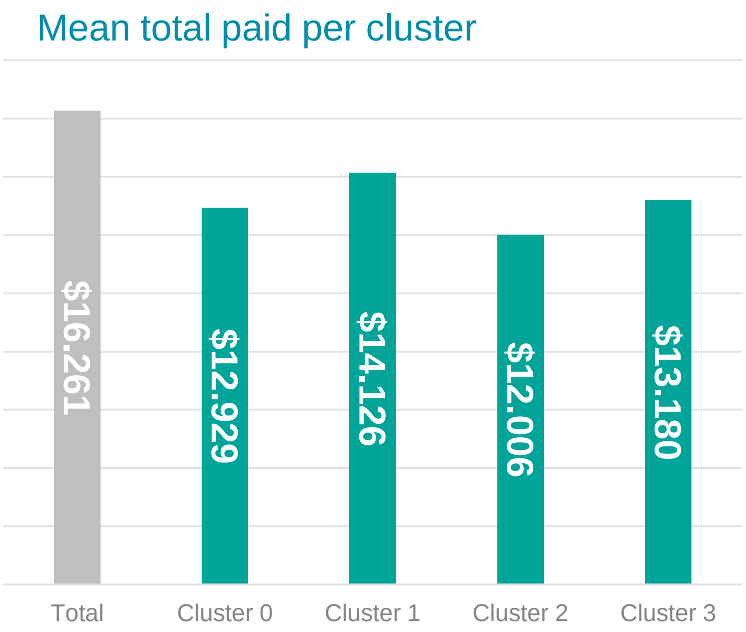
Example: prospects for car insurance already having a credit card

Vía clustering

We found 4 clusters with diffrent risk levels

Having CC: offer Car insurance

Pred Cluster	Potential n	Max Prob	min Prob	Sugg discount
0	75,058	64%	51%	7%
1	82,006	64%	51%	0%
2	1,767	68%	51%	13%
3	50	64%	51%	2%
	158,881			

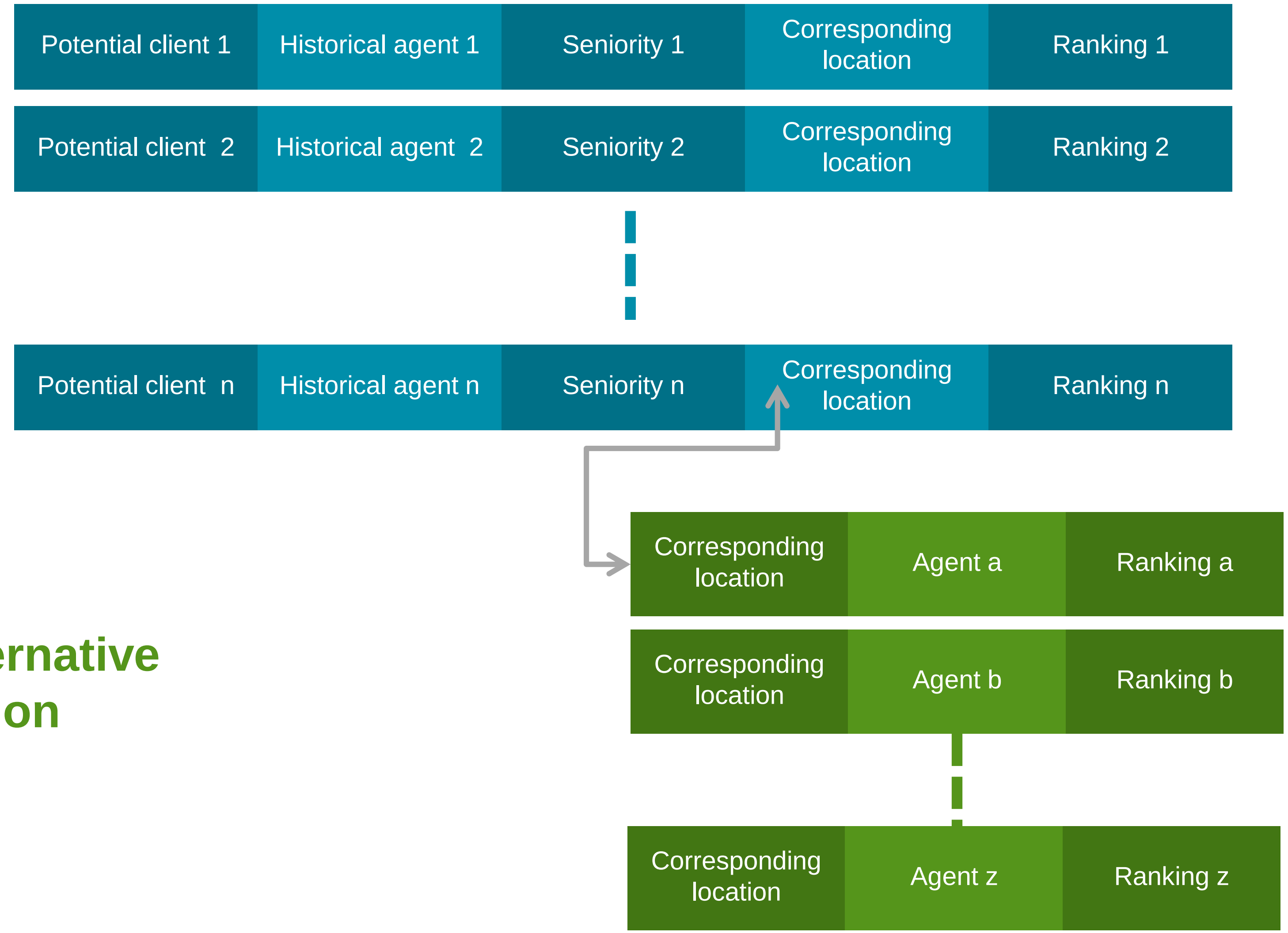


Classification model has 74% accuracy, classifying 49% of all potential clients. Insights regarding potential risks were not available beforehand (not correlated to usual risk scores).

Agent selection

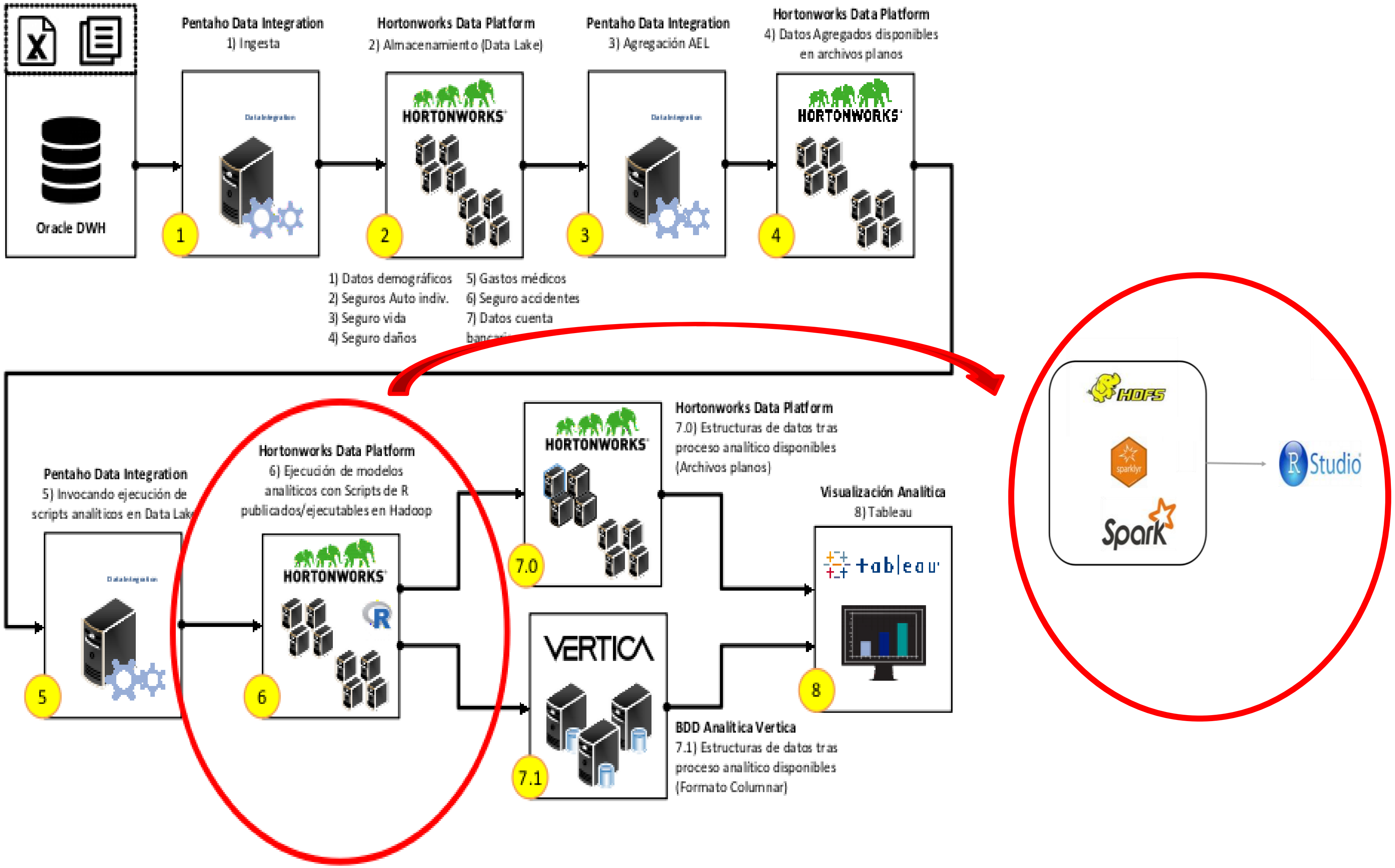
Factors

- Product
- Location
- Seniority
- Historical agent for each client
- Index ranking by product/location



Tables with historical and alternative agents per product/location

How: general architecture



How

- Pre-processing: data was loaded into HDFS and aggregated using Spark SQL
- Aggregated tables were processed in R (calling Sparklyr)
- Models and clusters were trained with R+ Spark with Sparklyr using inbuilt functions
  - ml_fpgrowth
  - ml_bisecting_kmeans
  - ml_gbt_classifier
- The team: a data engineer, two data scientists and the data science group at the bank.

```
source("../connect.R")

#### Load Table ####

ptf_agg <- tbl(sc,
  sql(paste(
      "select distinct id_cliente_comercial, productos ",
      "from f_analisis_portafolio_array lateral view explode(productos) c as m ",
      "where m>=300 and m<400", ## extrae los id de todos los que tienen captaciones: 3xx
      "and id_cliente_comercial is not null"
      )
  ))


#### Patrones Frecuentes ####

fpg <- ml_fpgrowth(ptf_agg,items_col = "productos",
                   min_confidence = 0.5,
                   min_support = .001)

fp <- fpg %>% ml_freq_itemsets() %>% collect() %>% as.data.frame()
ar <- fpg %>% ml_association_rules() %>% collect() %>% as.data.frame()
```

Summing up

Effectivity boost

- Sales up 10%+
- More potential clients
- Better agent/client match
- Improved rentability


Questions?
THANKS!!!