Table of Contents

Introduction

How can Amazon recommend the right product whilst the customers are buying? How can an organization know which customers are interested in leaving their services? Why do we receive fair deals when we are interested in buying a product? The key behind these questions is the effective deployment of Customer Analytics by these organizations, that means the efficient and effective use of customer data.

With the advent of social media and the democratization of the Internet, it has emerged a new type of customer: more informed, instrumented and connected. A customer that interacts with a company using all the available channels (offline, online, mobile and social). A new customer that is eager to feel personalized experiences.

As a result, intuition is no longer enough to understand the customer, data is needed. Organizations are being forced to focus on the client and transform their marketing, communication and sales experiences that must be grounded in the needs and preferences of customers.

This book aims to train professionals in the context of the analysis of customer data, also known as Customer Analytics, with the goal to start developing efficient strategies within its own organization.

The efficiency is achieved through the combination of multiple technologies (from Business Intelligence to Big Data) and the use of algorithms to understand and predict customers' behaviour and their future interactions.

This book complements professor Curto's **Customer Analytics** course at IE Business School, UOC and KSchool. The course, that consists of several sessions, is designed to introduce Customer Analytics from the perspective of the data scientist. That means, it is a practical and quantitative course.

The reader will learn the concept of Customer Analytics and some of the more relevant techniques used in the real world by companies around the world. Basic knowledge in statistics, mathematics, programming, R and RStudio is expected to be able to follow this book.

Methodology

Each technical chapter is explained from the business problem point of view and it includes the following sections:

- The problem
- Algorithm to solve the business problem
- Main concepts (related to the technique)
- Implementation Process
- Benefits
- Use cases
- How to implement the algorithm using R
- References

Tools

In this book will use R and RStudio. These programs can be downloaded from:

- R
- RStudio

It is recommended to install first R and then RStudio.

Fundamentals: Statistics, R & RStudio

During this book, we will refer to R and Rstudio features and statistical concepts. We recommend these additional references:

- The Elements of Statistical Learning
- Introduction to R
- Introduction to RStudio
- Rstudio cheatsheets
- Wickham, H. Ggplot2
- Wickham, H. Advanced R
- Wickham, H. R Packages
- Wickham, H. R for Data Science
- R Graph Catalog
- An Introduction to Statistical Learning with Applications in R

What is Customer Analytics

In this chaper we will introduce what is Customer Analytics.

Origins of Customer Analytics

	Data Dispersion	Data Organization	Data Ownership	Data Collaboration
Evolution Phase	Early 90s. Independent solutions: SFA, Call Center,…	Late 90s. CRM is understood as marketing, sales and support	2000 - 2010. CRM as a global strategy	2010 – 2020. Beginning of Social CRM
Characteristics	Technical solutions are developed for customer interactions	Customer Data is used in the organizacion	Customer Data is a critical asset. Loyalty programs are developed	Use of external data y data analytics
Focus	Technical Solution	Tactical and operational solution	Corporate culture and estrategy	Value co-creation and customer experience
Objective	Improve sales and support efficiency	Improve customer retention and cost savings per customer interaction	Cost savings and Revenue Growth. Predict customer behaviour	Value creation based on customer. Customer Centricity

CRM as the origins of Customer Analytics. Authors: Hannu Saarijärvi, Heikki Karjaluoto y Hannu Kuusela

The new consumer

A new breed of customers has emerged. This new consumer is:

- **Informed**: Information about products and services is available to consumers.
- **Instrumented**: They have devices and channels to accees information.
- **Connected**: They have pervesive access to information. Anywhere, anytime.

- **Less Loyal**: They are open to try new services and products.
- **More demanding**: they expect more value, honesty and transparency from companies.
- **ROBO** (Research Online, Buy Offline/Online) or **ROPO** (Research Online, Purchase Offline/Online): They have different and non-linear purchasing patterns.

Definition

We need a definition. We will use the one from (Wharton Customer Analytics Initiative):

> Customer Analytics refers to the collection, management, analysis and strategic leverage of a firm's granular data about the behavior(s) of its customers

But need to remember that:

> Essentially, all models are wrong, but some are useful (George Box).

Characteristics

Customer Analytics can be characterized as:

- **INHERENTLY GRANULAR**: must be individual-level
- **FORWARD-LOOKING**: orientation towards prediction not just description
- **MULTI-PLATFORM**: combining behaviors from multiple measurement systems
- **BROADLY APPLICABLE**: consumers, donors, physicians, clients, brokers, etc.
- **MULTIDISCIPLINARY**: marketing, statistics, computer science, information systems, operations research, etc.
- **RAPIDLY EMERGING**: starting to take on its own unique identity as a "standalone" area of analysis and decision making
- **BEHAVIORAL**: customer analytics' primary focus is on observed behavioral patterns
- **LONGITUDINAL**: It's ALL about how these behaviors manifest themselves over time

What does it mean to generate value?

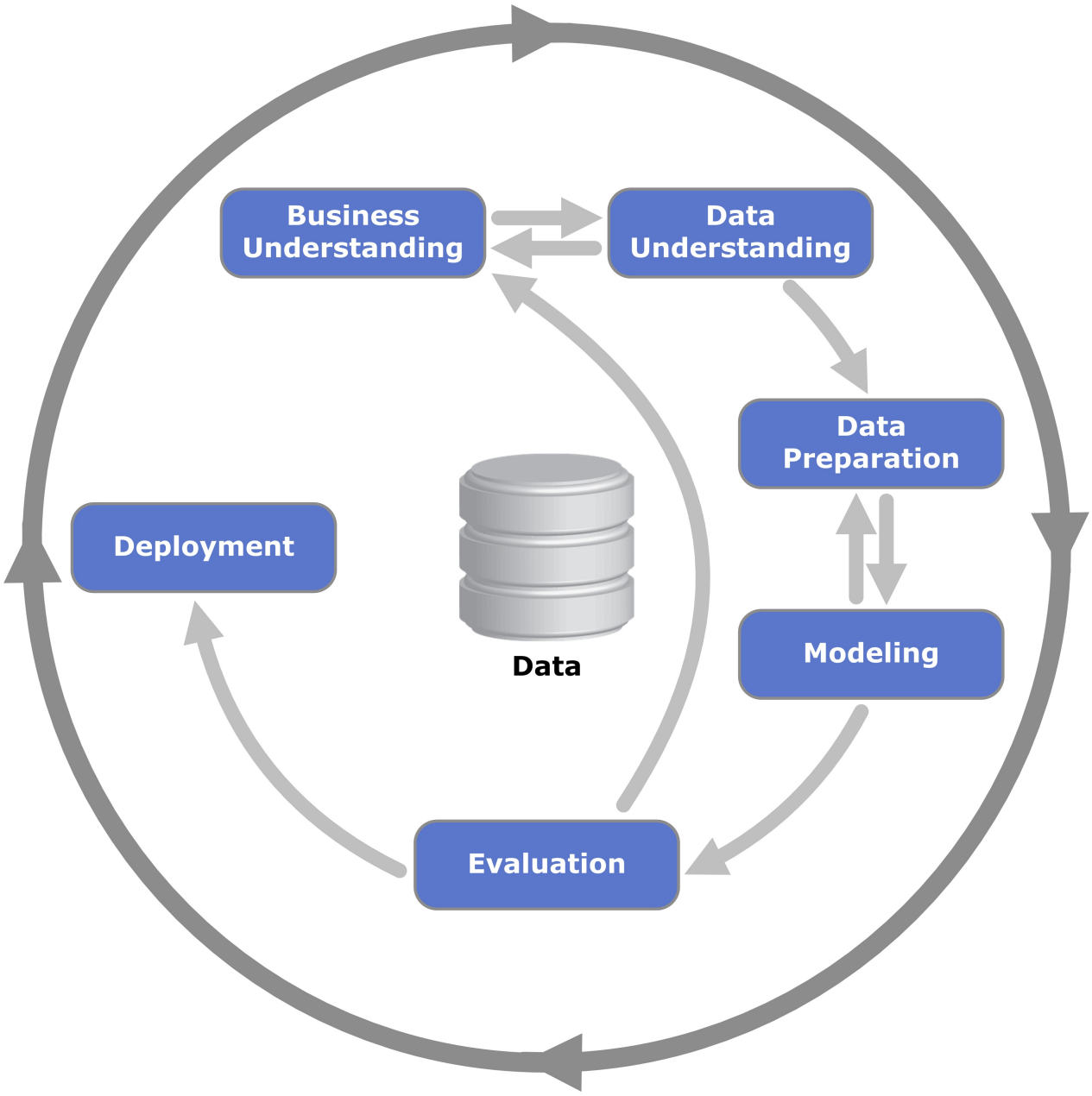
- **Customer Lifetime Value (CLV)** is a prediction of the net profit attributed to the entire future relationship with a customer..
- **Customer equity** is the total of lifetime values of all your current and future customers – the sum total of all the value you'll ever realize from customers.
- **Value Proposition** is a business or marketing statement that summarizes why a consumer should buy a product or use a service. This statement should convince a potential consumer that one particular product or service will add more value or better

solve a problem than other similar offerings.
- **Customer Satisfaction** is a marketing term that measures how products or services supplied by a company meet or surpass a customer's expectation.
- **Customer delight/sacrifice** is the value that is added or subtracted to the customer value proposition as a surprise.
- **Switching Cost** are the negative charges associated with the change of supplier, brand or product by a customer..
- **Customer Loyalty** is the result of consistently positive emotional experience, physical attribute-based satisfaction and perceived value of an experience, which includes the product or services.

Types of strategies

- **Customer Acquisition**: the organization wants to increase the number of customers.
- **Customer Development**: the organization wants to increase customer profitabity and/or loyalty.
- **Customer Retention**: The organization aims to prevent customers leaving the service either because they stop consuming/buying or because they are switching to other suppliers.
- **Acquisition-Retention Optimization**: The organization aims to balance customer acquisition and retention strategies and to avoid side effects such as customer churn propensity.

Methodology



- **[BU]** Business Understanding
  - Determine business objectives
  - Assess situation
  - Determine data mining goals
  - Produce project plan
- **[DU]** Data Understanding
  - Collect initial data
  - Describe data
  - Explore data
  - Verify data quality
- **[DP]** Data Preparation
  - Select data
  - Clean data

    - Construct data
    - Integrate data
- **[M]** Modeling
    - Select modeling technique
    - Generate test design
    - Build model
    - Assess model
- **[E]** Evaluation
    - Evaluate results
    - Review process
    - Determine next steps
- **[D]** Deployment
    - Plan deployment
    - Plan monitoring and maintenance
    - Produce final report
    - Review project

It must be noted that:

> We must find the balance between analytical results, the business needs and operational constraints.

Without taking this into account, Customer Analytics initiatives will fail. This is true for any analytical initiative.

How to obtain customer data

In order to understand the customer we need data. Companies use several resources:

- **Internal resources:** data from information systems such as CRM, ERP, Call Center, e-commerce, etc.
- **External resources:** Cookies; Super Cookies -HTTP Strict Transport Security (HSTS)-; mobile devices ID (iOS Identifiers for Advertisers or Android Advertising ID); beacons; HTML5 storage; geolocalization and wifi (IP); Fingerprinting; Adobe Flash, Applets Java and ActiveX Controllers; Plug-ins, toolbars and spyware; Etags, Google Data; Whatsapp and Facebook; Windows 10 Telemetry; Deep Packet Inspection (DPI); third parties data; security breaches.

Increasing the analytical maturity of the organization

- **Descriptive analysis**: the organization is able to understand what happened in every customer interaction.
- **Diagnostic Analysis**: the organization is able to understand the reasons why interactions with customers happen.
- **Predictive analysis**: the organization is able to predict certain customer interactions.
- **Prescriptive analysis**: the organization is able to make decisions related to customer interactions based on scenarios.
- **Preventive Analysis**: the organization is able to act in advance of customer needs.

References

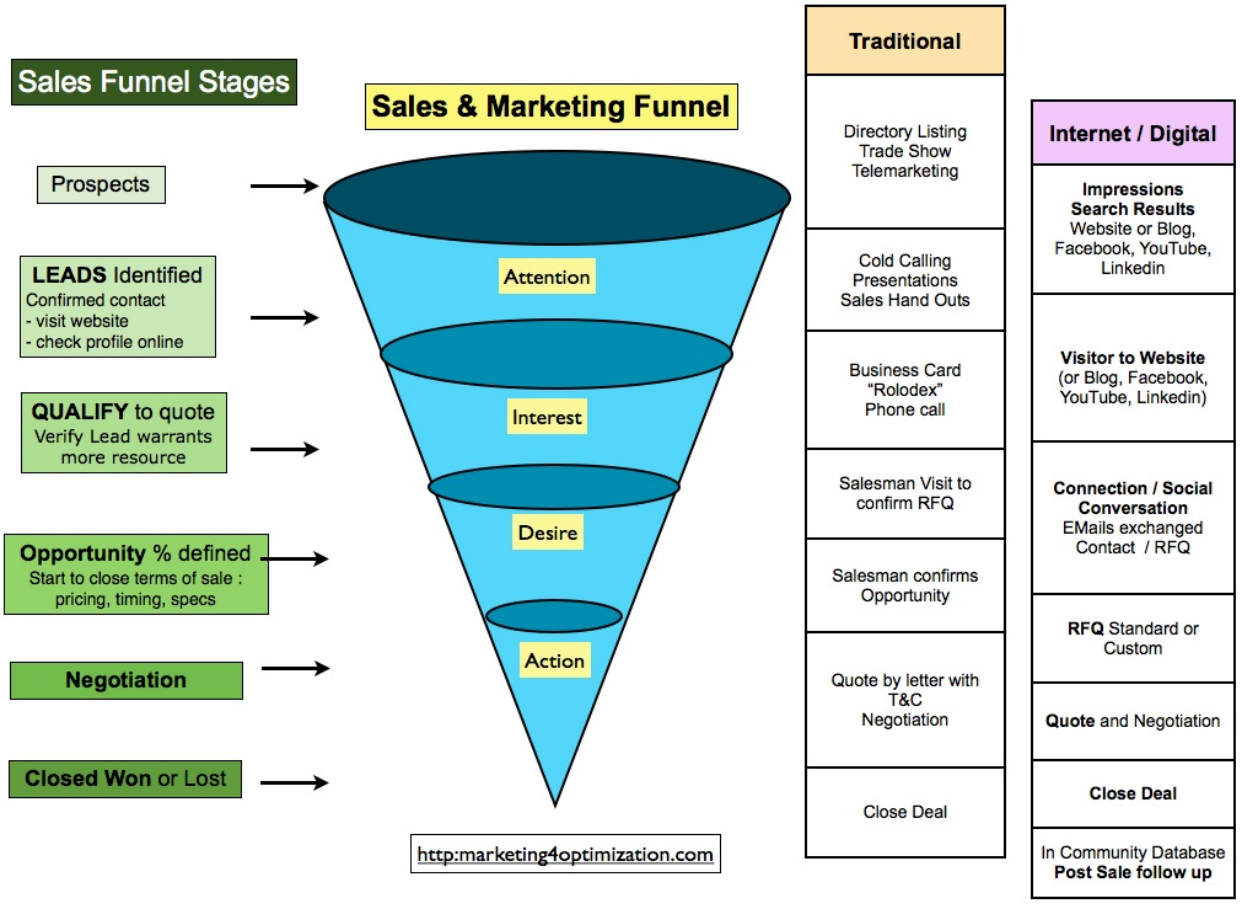
- Curto, J. & Braulio, N. (2015) Customer Analytics. Editorial UOC
- Wharton Customer Analytics Initiative
- Wharton Customer Analytics Initiative Research Paper Series
- Hannu Saarijärvi , Heikki Karjaluoto , Hannu Kuusela , (2013) Customer relationship management: the evolving role of customer data, Marketing Intelligence & Planning, Vol. 31 Iss: 6, pp.584 - 600
- Kwong, K. K., & Yau, O. H. (2002). The conceptualization of customer delight: A research framework. Asia Pacific Management Review, 7(2)
- Rüdiger Wirth (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining 29-39
- Value Proposition Canvas
- Radiant
- Netflix: Recommending for the world

Methodologies, Technology and Techniques

In this chapter we will introduce some methodologies, technology and techniques that can be used to analize customer data.

Customer Pipeline

How we want customers to behave:



Customer Journey

How customers behave:

Rail Europe Touchpoints by Channel

Stage / Channels	Research & Planning	Shopping	Booking	Pre-Travel (Documents)	Travel	Post-Travel
Website	Maps Test intineraries Timetables Destination Pages FAQ General product & site exploration	Schedule look-up Price look-up Multi-city look-up Pass comparison	Web booking funnel - Pass - Trips - Multiple Trips	Select document option (from available options) - station e-ticket - home print e-ticket - mail ticket	Contact page for email or phone	
Call Center	Order brochure Planning (Products) Schedules General questions	Site navigation help	Automated booking payment Cust. Rep booking Site navigation help	Call re: ticket options Request ticket mailed Reslove problems (info, payment, etc.)	Call with questions regarding tickets General calls re: schedules, strikes, documents	
Mobile	Trip ideas	Schedules	Mobile trip booking		Access itinerary Look up schedules Buy additional tickets	
Communication Channels (social media, email, chat)	Chat for web nav help	FB Comparator Email questions Chat for website nav help	Chat for booking support	Email confirmations Email for general help Hold ticket	Ask questions or resolve problems re: schedules and tickets	Complaints or compliments Survey
Customer Relations						Request for refund, escelation from call center.
Non-REI Channels	Trip Advisor Travel blogs Social Media General Google searching	Airline comparison Kayak Direct rail sites	Expedia		Travel Blogs Direct rail sites Google searches	Trip Advisor Review sites Facebook

Non-linear, no time restrictions — *Linear process* — *Non-linear, but time based*

Technologies

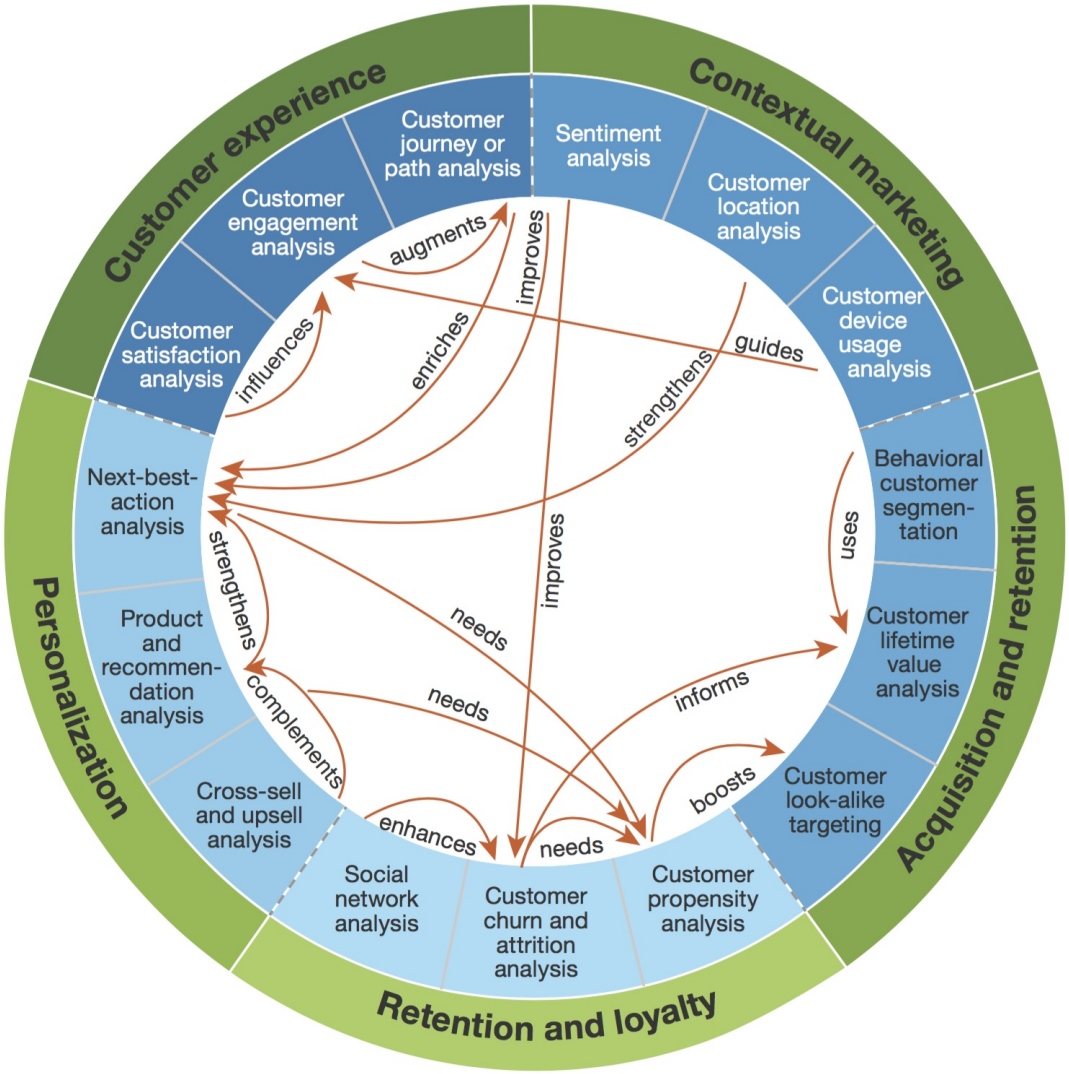
- **Business Intelligence**: includes strategies, technologies and information systems aiming to improve decision-making based on past performance using reporting, OLAP, Dashboards, Balanced Scorecard, Scoreboards, Data Visualization, Data Storytelling, and others.
- **Data Analytics**: includes strategies, technologies and information systems aiming to identify patterns and predict trends and behaviours using Data Mining, Text Mining, Machine Learning, Artificial Intelligence, Cognitive Systems and others.
- **Big Data**: includes strategies, technologies and information systems aiming to capture, process, store, analyse and visualize complex data sets using batch processing, streaming processing, NoSQL, HPC, MPP, In-Memory and others.
- **Data Management**: includes strategies, technologies and information systems aiming to provide information management capabilities using Data Governance, Data Quality, Master Data Management and others.
- **Data Brokerage**: includes strategies, technologies and information systems aiming to create data marketplaces for raw data or insights.

Types of analysis

Strategy	Type of analysis
Customer understanding	Homogeneous and heterogeneous clustering, Demographic (and others) profiling, trends and patterns,...
Customer acquisition	Propensity analysis, factor analysis, affinity models,...
Customer Loyalty and profitability	Propensity analysis, cross-selling and up-selling analysis, Affinity analysis, Social Analytics, Sentiment Analytics, RFM Modeling, Price elasticity, customer satisfaction analysis,...
Customer retention	Churn Analysis, Response models, What-if Analysis,...
Efficiency	Marketing Mix models, customer journey analysis, customer location & device usage analysis,..

Types of analysis - How they relate

According to Forrester:



Understanding your customer

What does it mean understanding your customer? There are different **perspectives**:

- **Behavioral perspective**. Punctuality (payment). Risk Index. Purchasing patterns. Affinity analysis. Propensity analysis. Punctuality profiling. Risk profiling. Event patterns.
- **Profitability perspective**. Current profitability. Future profitability. Wallet share. Profitability Profiling. Profitability Conversion.
- **Lifecycle perspective**. CLV. CLV profiling
- **Loyalty perspective**. Recency. Frequency. Monetary value. Churn. Acquisition. Retention. Increase. RFM profiles. Retention profiles. Growth profiles.
- **Interest perspective**. Response rate. Response modeling. Response analysis. Response profiles
- **Campaign perspective**. Response rate. ROI. Estimated value added. Lifting. Effectiveness of events. Profitability Cannibalization. Events Cannibalization. Event profitability. Cross-channel effectiveness.

Types of analysis and Techniques

Knowledge

Analysis	Method	Techniques	Capabilities
Punctuality	Querys/Reporting	Structured Procedures	Business Intelligence
Risk index	Querys/Reporting	Structured Procedures	Business Intelligence
Purchasing patterns	Querys/Reporting	OLAP	Business Intelligence
Affinity Analysis	Descriptive Analysis	Basket Case Analysis	Business Analytics
Propensivity Analysis	Descriptive Analysis	Affinity Analysis	Business Analytics
Punctuality Profiling	Descriptive Analysis	Decision Trees, Induction Rules	Business Analytics
Risk Profiling	Predictive Analysis	Multi-regression	Business Analytics
Events patterns	Predictive Analysis	OLAP, Association Rules	Business Analytics

Profitability

Analysis	Method	Techniques	Capabilities
Current Profitability	Querys/Reporting	Structured Procedures	Business Intelligence
Potential Profitability	Predictive Analysis	Multi-regression	Business Analytics
Future Profitability	Predictive Analysis	Neural Networks	Business Analytics
Profitability Profile	Predictive Analysis	Decision Trees, Induction Rules	Business Analytics
Profitability Conversion	Predictive Analysis	Lineal/Logistic regression	Business Analytics
Wallet share	Predictive Analysis	Lineal/Logistic regression	Business Analytics

Life-cycle

Analysis	Method	Techniques	Capabilities
CLV	Predictive Analysis	Statistics	Business Analytics
Potential CLV	Predictive Analysis	Multi-regression	Business Analytics
CLV Profiling	Predictive Analysis	Supervised Clustering	Business Analytics

Loyalty

Analysis	Method	Techniques	Capabilities
RFM	Querys / Reporting	Structured Procedures	Business Intelligence
Churn	Predictive Analysis	Decision Trees / Clasificación	Business Analytics
Acquisition	Querys / Reporting	Structured Procedures	Business Intelligence
Retention	Querys / Reporting	Structured Procedures	Business Intelligence
Growth	Querys / Reporting	Structured Procedures	Business Intelligence
RFM profiling	Predictive Analysis	Decision Trees, Induction Rules	Business Analytics
Acquisition Modeling	Predictive Analysis	Neural Networks	Business Analytics
Retention profiling	Predictive Analysis	Decision Trees, Induction Rules	Business Analytics
Growth profiling	Predictive Analysis	Decision Trees, Induction Rules	Business Analytics
Loyalty convertion	Predictive Analysis	Neural Networks	Business Analytics

Interest

Analysis	Method	Techniques	Capabilities
Response	Predictive Analysis	Neural Networks	Business Analytics
Response Modeling	Predictive Analysis	Neural Networks	Business Analytics
Response Analysis	Predictive Analysis	Neural Networks	Business Analytics
Response Profiling	Predictive Analysis	Decision Trees, Induction Rules	Business Analytics

Campaigns

Analysis	Method	Techniques	Capabilities
Response Index	Predictive Analysis	Statistics	Business Analytics
ROI	Querys / Reporting	Structured Procedures	Business Intelligence
Expected Value-added	Predictive Analysis	Regression	Business Analytics
Up-Selling	Querys / Reporting	Structured Procedures	Business Intelligence
Profitability Cannibalization	Querys / Reporting	Structured Procedures	Business Intelligence
Sales Canibalization	Querys / Reporting	Structured Procedures	Business Intelligence
Event Profitability	Querys / Reporting	Structured Procedures	Business Intelligence
Cross-Channel Effectivity	Querys / Reporting	Structured Procedures	Business Intelligence
Channel Effectivity	Querys / Reporting	Structured Procedures	Business Intelligence

References

- Peter Verhoef, Edwin Kooge, Natasha Walk. (2016) Creating Value with Big Data Analytics. Routeledge
- Peter Fader. (2012) Customer Centricity. Wharton Executive Essentials

Customer Lifetime Value

The business problem

- **Problem**: we don't understand our customer. We don't know if a specific customer is relevant for our organization.
- **Goals**:
  - We want to know how much value a customer is generating for the organization.
  - We want to know how the generated value is going to evolve.
  - We want to segment customers based on value.
- **Why?** To perform specific actions for different customer groups

Technique to solve the business problem

We have many options to understand customer's value. One of the most relevant is CLV. CLV is the acronym of **Customer Lifetime Value**. We need a definition:

> CLV is a prediction of all the value a business will derive from their entire relationship with a customer.

There is a tough part in this definition: how to estimate future customer interactions.

The calculation of CLV can be based on:

- ARPU/ARPA (Historic CLV)
- RFM (only applicable to the next period)
- CLV Formula (historic CLV)
- Probability/Econometrics/ Persistence Models/Machine Learning/Growth and dissemination models such as:
  - Moving Averages, Regressions, Bayesian Inference, Pareto/NBD (Negative Binomial Distribution)

The value generated by all our customers is called **Customer Equity** and is used to evaluate companies (with no significant income yet).

Main Concepts

There are many factors that affect customer's value and must be considered in the CLV calculation:

- **Cash flow**: the net present value of the income generated by the customer to the organization throughout their relationship with the organization
- **Lifecycle**: the duration of the relationship with the organization
- **Maintenance costs**: associated costs to ensure that the flow of revenue per customer is achieved
- **Risk costs**: risk associated with a client
- **Acquisition costs**: costs and effort required to acquire a new customer.
- **Retention costs**: costs and effort required to retain a new customer.
- **Recommendation value**: impact of the recommendations of a customer in its sphere of influence on company revenue
- **Segmentation improvements**: value of customer information in improving customer segmentation models

CLV formula

It is:

$$ CLV = \sum^{T}*{n=0} *frac{(p*{t}-c_{t})r_t}{(1+i)^t} - AC

$$

where

$$p_t =$$ price paid by the customer in time $$t$$,

$$c_t$$ = direct costs for customer service in time $$t$$,

$$i$$ = discount rate or cost of money for the firm,

$$r_t$$ = probability that the client returns to buy or is alive in time $$t$$,

$$AC$$ = acquisition cost, and

$$T$$ = time horizon to estimate $$CLV$$.

A simplification

Let's consider that $$p$$ y $$c$$ are constant values and $$r$$ is a decresing fucntion in time, then the formula becomes:

$$ CLV = \sum^{\infty}_{n=0} \frac{(p-c)r^t}{(1+i)^t} =\frac{(p-c)r}{1+i-r}

$$

Final consideration

- Generally we will have a dataset for each of the concepts of the formula
- We'll have to estimate the rest
- Therefore, it is important to remember that:

> Our ability to predict the future is limited by the fact that to some extent is contained in the past.

- What mathematically means we are under some continuity conditions (or the hyphotesis are true)

Implementation Process

- **[BU]** Discuss whether CLV fits as a metric in our business
- **[DP]** Identification and understanding of sources and metadata
- **[DP]** Extract, transform, clean and load data
- **[M]** Choose CLV method
- **[M/E]** Analyze results and adjust parameters
- **[D]** Present and explain the results

Benefits

This technique provides the following benefits:

- Ability to create business objectives
- Better understanding of customers
- Having a common numerical analysis criteria
- Ability to have an alert system
- Improved management of the sales force

Use cases

- Create market strategies based on CLV
- Customer segmentation based on CLV
- Forecasting and customer evolution per segment
- Create different communication, services and loyalty programs based on CLV
- Awake "non-active" customers
- Estimated the value of a company (startup, in the context of acquisition)

How to implement CLV using R

ARPU/ARPA as CLV aproximation

Average Revenue per User (ARPU) and Average Revenue per Account (ARPA) can be used to calculate historical CLV. Process:

- calculate the average revenue per customer per month (total revenue ÷ number of months since the customer joined)
- add them up
- and then multiply by 12 or 24 to get a one- or two-year CLV.

Suppose Josep and Laura are your only customers and their purchases look like this:

Customer Name	Purchase Date	Amount
Josep	January 1, 2015	$150
Josep	May 15, 2015	$50
Josep	June 15, 2015	$100
Laura	May 1, 2015	$45
Laura	June 15, 2015	$75
Laura	June 30, 2015	$100

Suppose today is July 1, 2015. Average monthly revenue from Josep is

$$(150 + 50 + 100)/6 = 50$$

and average monthly revenue from Laura is

$$(45 + 75 + 100)/2 = 110$$.

Adding these two numbers gives you an average monthly revenue per customer of $160/2 = $80. To find a 12-month or 24-month CLV, multiply that number by 12 or 24.

The benefit of an ARPU approach is that it is simple to calculate, but it does not take into account changes in your customers' behaviors

Let's understand the CLV formula

```
# Install packages
install.packages("readxl")
install.packages("ggplot2")
```

```
# Load packages
library(ggplot2)
library(readxl)

# Load data into a dataframe
df <- read_excel("data/s3.xlsx", sheet = "Ex2")

# Summary
summary(df)

# Question: What we can say about the summary?

# Graph: how the number of customers is evolving
ggplot(df, aes(x = t, y = active)) +
  geom_line() + ggtitle("Active Customer Evolution") +
  ylab("Customer") + xlab("Period") +
  theme(plot.title = element_text(color="#666666", face="bold",
size=20, hjust=0)) +
  theme(axis.title = element_text(color="#666666", face="bold",
size=14))

# Graph: how the retentio ratio is evolving
ggplot(df, aes(x = t, y = r)) +
  geom_line() + ggtitle("Retention Ratio Evolution") +
  ylab("Customer") + xlab("Period") +
  theme(plot.title = element_text(color="#666666", face="bold",
size=20, hjust=0)) +
  theme(axis.title = element_text(color="#666666", face="bold",
size=14))

# Partial CLV
df$CLV <- (df$p-df$c)*df$r^df$t/(1+df$i)^df$t

# Now we have a new column
df

#  Graph: how the partial CLV is evolving
ggplot(df, aes(x = t, y = CLV)) + geom_line() + ggtitle("Partial
CLV evolution") + ylab("CLV") + xlab("Period") +
theme(plot.title = element_text(color="#666666", face="bold",
```

```
size=32, hjust=0)) +
  theme(axis.title = element_text(color="#666666", face="bold",
size=22))

# Question: What do we observe?

# Final step
CLV <- apply(df, 2, sum)
CLV[7]

# What does it mean this value?

# Question: What happens if retention ratio has a constant value
of 0.80?
```

References

- Modeling Customer Lifetime Value
- CLTV, Harvard
- The 16 business benefits of Customer Lifetime Value
- Dominique Hanssens's Research
- BTYD package
- Kinshuk Jerath, Peter S. Fader, Bruce G.S. Hardie, Customer-base analysis using repeated cross-sectional summary (RCSS) data, European Journal of Operational Research, Volume 249, Issue 1, 16 February 2016, Pages 340-350, ISSN 0377-2217, http://dx.doi.org/10.1016/j.ejor.2015.09.002.
- Can the negative binomial distribution predict industrial purchases? John W Wilkinson , Giang Trinh , Richard Lee , Neil Brown. Journal of Business & Industrial Marketing 2016 31:4
- Rajkumar Venkatesan, V. Kumar (2004) A Customer Lifetime Value Framework for Customer Selection and Resource Allocation Strategy. Journal of Marketing: October 2004, Vol. 68, No. 4, pp. 106-125

RFM Analysis and Modelization

The business problem

- **Problem**: we don't understand our customer. We don't know if a specific customer is relevant for our organization.
- **Goals**:
  - We want to know how much value a customer is generating for the organization.
  - We want to know how the generated value is going to evolve.
  - We want to segment customers based on value.
- **Why?** To perform specific actions for different customer groups

But attention, our organization is starting the customer analytics journey and we can not propose really sophisticated techniques.

Technique to solve the business problem

RFM Analysis is a customer segmentation method based on:

- Recency
- Frequency
- Monetary

It is easy to implement and understand. It is based on pareto principle: **20% of customers generate 80% of revenue.**

When to use it: Used for non-contractual purchasing.

Main Concepts

The main concepts are:

- **Recency**: the number of days that have passed since the customer last purchased - How recently did the customer purchase?
- **Frequency**: number of purchases in a specific period (for example, last 12 months) - How often do they purchase?
- **Monetary**: value of orders from a given customer in the specific period - How much do they spend?

Implementation Process

- **[BU]** Discuss whether RFM fits as a metric in our business
- **[DP]** Identification and understanding of sources and metadata
- **[DP]** Extract, clean and load data
- **[M]** Create R,F,M variables per customer
- **[M]** Assign RFM percentile (use from 5 -offline- up to 10 -online- percentile segments)
- **[E]** Analyze results
- **[D]** Present and explain the results

Benefits

This technique provides the following benefits:

- It is a customer segmentation easy to calculate and understand
- Good fit for direct marketing: Which are the best segments? Who do I need to contact?
- Helps to improve CLV
- Helps to reduce churn (instead using other more complex techniques)
- Can be extended with location dimension (RFML)
- Provides some kind of insight about loyalty

Use cases

This technique is used in different use cases:

- Forecasting and Sales reporting based on RFM
- RFM profiling
- Churn Analysis
- Marketing actions: mailing, call center,...

How to implement RFM using R

```
# Install packages
install.packages("readr")
install.packages("dplyr")
install.packages("DataExplorer")

# Load packages
```

```
library(dplyr)
library(ggplot2)
library(readr)
library(DataExplorer)

# Load data
sales_data <- read.csv("data/s4.csv",stringsAsFactors=FALSE)

# Review data
str(sales_data)
View(sales_data)
summary(sales_data)

# Create recency
sales_data$Recency <- round(as.numeric(difftime(Sys.Date(),
                            sales_data[,3],units="days")) )

# Creation of Recency, Frequency y Monetary
salesM <-
aggregate(sales_data[,2],list(sales_data$idCustomer),sum)
names(salesM) <- c("idCustomer","Monetary")
salesF <-
aggregate(sales_data[,2],list(sales_data$idCustomer),length)
names(salesF) <- c("idCustomer","Frequency")
salesR <-
aggregate(sales_data[,4],list(sales_data$idCustomer),min)
names(salesR) <- c("idCustomer","Recency")

# Combination R,F,M
temp <- merge(salesF,salesR,"idCustomer")
salesRFM <- merge(temp,salesM,"idCustomer")

#  Creation of R,F,M rank
salesRFM$rankR <- cut(salesRFM$Recency,5,labels=F)
salesRFM$rankF <- cut(salesRFM$Frequency,5,labels=F)
salesRFM$rankM <- cut(salesRFM$Monetary,5,labels=F)

# Review top 10
salesRFM <- salesRFM[with(salesRFM, order(-rankR, -rankF, -
rankM)), ]
```

```
head(salesRFM, n=10)

# Analysis
groupRFM <- count(salesRFM, rankR, rankF,rankM)
groupRFM <- salesRFM$rankR*100 + salesRFM$rankF*10 +
salesRFM$rankM
salesRFM <- cbind(salesRFM,groupRFM)

# Plot
ggplot(salesRFM, aes(factor(groupRFM))) +
  geom_bar() +
  ggtitle('Customer Distribution by RFM') +
  labs(x="RFM",y="# Customers") +
  theme(plot.title = element_text(color="#666666", face="bold",
size=16, hjust=0)) +
  theme(axis.title = element_text(color="#666666", face="bold"))
```

References

- Shweta Singh , Sumit Singh , (2016) "Accounting for risk in the traditional RFM approach", Management Research Review, Vol. 39 Iss: 2, pp.215 - 234
- Fader, Peter S., Bruce G.S. Hardie, and Ka Lok Lee (2005), "RFM and CLV: Using Iso-value Curves for Customer Base Analysis." Abstract PDF
- Fader, Peter S., Bruce G.S. Hardie, and Ka Lok Lee (2004), "'Counting Your Customers' the Easy Way: An Alternative to the Pareto/NBD Model." Abstract PDF Associated Excel spreadsheet and note
- Tkachenko, Yegor. Autonomous CRM Control via CLV Approximation with Deep Reinforcement Learning in Discrete and Continuous Action Space. (April 8, 2015). PDF

Customer Satisfaction

The business problem

- **Problem**: we don't know if our customers are happy with our products and services.
- **Goals**:
  - We want to know if customers are loyal to our brand
  - We want to know if the customers are satisfied with our services and products.
  - We want to segment customers in three groups: promoters, passive and detractors
- **Why?** Perform specific actions to improve our services and products

Technique to solve the business problem

Customer Satisfaction is a complex problem. Discovering the real factors requires time and the combination of several techniques. As a starting point, we will introduce NPS (Net Promoter Score).

Main Concepts

What is Customer Satisfaction

We need a definition

> We define customer satisfaction as the number of clients (or the percentage of the total) whose experience with the products or services of the organization exceed a specific goal of satisfaction.

This concept is often used in marketing and is closely linked to expectations that a customer has for the brand.

Understanding Customer Satisfaction requires to discover the factors that influence satisfaction. Most of the times, we used surveys to discover the factors. Likert scale is used to evaluate the characteristics of a product or service.

> The Likert scale is a psychometric scale that specifies the level of agreement or disagreement with a statement.

What is NPS

- **Net Promoter Score®** (NPS) is a method created by Freid Reichheld (Bain/SatMetrix) in 2004.
- This metric measures the customer loyalty and allows to discover if they would recommend the services/products of a company
- According to Satmetrix, NPS is not just a customer satisfaction survey
- It is a measure that identifies the extent to which a company is increasing customer loyalty and why
- Types: Competitive benchmark, Customer Experience, Customer Relationship

Promoters, Passive and Detractors

NPS allows to segment the customer in three groups:

- **Promoters**: Those who respond with a score of 9 or 10 are called Promoters, and are considered likely to exhibit value-creating behaviors, such as buying more, remaining customers for longer, and making more positive referrals to other potential customers. They are happy and loyal customers. They "represent" 80% of revenue.
- **Passive**: Responses of 7 and 8 are labeled Passives, and their behavior falls in the middle of Promoters and Detractors. They are neither happy neither unhappy customers.
- **Detractors**: Those who respond with a score of 0 to 6 are labeled as Detractors, and they are believed to be less likely to exhibit the value-creating behaviors. They are unhappy customers. They can impact on branding, employee morale and customer acquisition/retention.

Pros & Cons

Pros

- Easy to calculate
- Easy to compare and explain
- Easy to use as benchmarking for companies and industries

Cons

- We lose information by definition
- His interpretation is not unique
- It is very sensitive to small variations
- It is just a diagnostic metric
- It is based on an attitude, intention to recommend, so it depends on personal factors.
- The criteria for selecting the cutoffs between segments (Promoters, Neutrals and Detractors) is diffuse

- The emotional component of a recommendation is not measured

Implementation Process

NPS is based on one question:

What is the likelihood that you would recommend Company X to a friend or colleague?

Let's imagine that we have 200 answers:

- 120 promoters = 60% will recommend
- 50 passive = they don't care
- 30 detractors = 15% won't recommend and probably will critizice
- NPS = 60 – 15 = 45

Procedure

- **[BU/DU]** Determine products/services/department to analyze
- **[DP]** Prepare and validate questions and execute survey
- **[M]** Calculate Promoters, Passive and Detractors
- **[M]** Every product/service/department will have a NPS*
- **[E]** Analyze results
- **[D]** Present and explain the results

Benefits

This technique provides the following benefits:

- Create actions linked to groups
- Understand loyalty and satisfaction
- Discover what products/services should be reviewed
- Investigate the causes of low loyalty
- It can help but must be combined with other metrics

Use cases

This technique is used in different use cases:

- **Retention rate**: Promoters usually drop out at lower rates than other customers, which means they have longer and more profitable relationships with a company.

- **Annual Expenditure and participation in the portfolio**: Promoters increase their purchases faster than detractors, as they tend to consolidate their purchases with their preferred provider. They are more interested in new offerings and brand extensions than detractors.
- **Pricing**: The promoters are usually less price sensitive than other customers, particularly detractors.
- **Cost efficiency**: Promoters often require less expenses in sales, marketing and advertising than other customers. In addition, the average size of orders is typically larger, leading to lower transaction costs per unit of revenue. Generally they have fewer complaints and represent less credit losses. Their positive attitudes have an effect difficult to quantify, but important in raising morale and employee productivity.
- **Word of mouth**: Promoters generate 80 percent to 90 percent of referrals. Quantify (by survey if necessary) the proportion of new customers who selected their company or product because of the reputation or referral. Conversely, detractors represent more negative word of mouth.

Alternatives

Customer satisfaction score (CSAT) is a management tool used to measure a customer's satisfaction for the service received.

- Question: **How would you rate your over all satisfaction with the service/product you received?**

$$ \frac{\hbox{Number of satisfied customers}}{\hbox{Number of total responses}} = \% \hbox{Happy customers}

$$

Customer Effort Score (CES) tries to access how much effort the customer had to put into a particular interaction with the company.

- Question: **How much effort did you personally have to put forth to handle your request?**
- Likert Scale: from 1 (very low effort) to 5 (very high effort).

How to implement this algorithm using R

```
# Load the library
library(NPS)

# Generate a data set
x <- sample(0:10, prob=c(0.02, 0.01, 0.01, 0.01, 0.01, 0.03,
0.03, 0.09, 0.22, 0.22, 0.35), 1000, replace=TRUE)

# Data Exploration
summary(x)

# Frequency Table
prop.table(table(x))

# Let's draw the histogram
hist(x, breaks=-1:10, col=c(rep("red",7), rep("yellow",2),
rep("green", 2)))

# We can use the basic graph library
barplot(prop.table(table(x)),col=c(rep("red",7),
rep("yellow",2), rep("green", 2)))

# NPS Calculation
nps(x) # equivalente nps(x, breaks = list(0:6, 7:8, 9:10))

# Standard Error
nps.se(x)

# Variance
nps.var(x)
```

References

- The One Number You Need to Grow
- EFQM User Guide: Net Promoter Score
- A Longitudinal Examination of Net Promoter and Firm Revenue Growth
- Measuring Customer Satisfaction and Loyalty: Improving Net-Promoter Score - Identification of Key Drivers of Net Promoter Score Using a Statistical Classification Model

Association Analysis

The business problem

Many companies accumulate large quantities of customer purchasing data from their day-to-day operations. This dataset represents an opportunity to understand customer purchasing preferences.

- **Problem**: we don't know which products are purchased by customers, which products are purchased jointly and which regularities can be found between products and customers.
- **Goals**:
  - We want to understand customer purchasing preferences
- **Why?** Perform specific actions to improve our services and products (such as placement, recommendations, etc.)

Technique to solve the business problem

We will use **association analysis**:

- It is a technique that helps to detect and analyse the relationships in registered transactions of individuals, groups and objects.
- One the most know analysis is the market basket analysis aiming to understand the relationship between acquired products.
- We will use the Apriori algorithm

Main Concepts

- **Association rules**, described by lhs => rhs:
  - **Lhs** refers to the left element in the rule and it is the acronym of *left hand side*. It is an itemset.
  - **Rhs** refers to the right element in the rule and it is the acronym of *right hand side*. It is an itemset.
- **Support**: This says how popular an itemset is, as measured by the proportion of transactions in which an itemset appears.
- **Confidence**: This says how likely item Y is purchased when item X is purchased, expressed as {X -> Y}. This is measured by the proportion of transactions with item X, in

which item Y also appears.
- **Lift** This says how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is.

$$ \frac{support(lhs \cup rhs)}{support(lhs) * support(rhs)} $$

The best predictor is *lift*. We should start our analysis with this parameter.

- If $$lift<1$$, lhs presence does not imply rhs presence.
- If $$lift=1$$, lhs and rhs are independent.
- If $$lift>1$$, lhs presence increases the probability of rhs presence in the transaction.

Implementation Process

- **[BU/DU]** Determine whether Association Analysis fits in our business
- **[DP]** Identification and understanding of sources and metadata
- **[DP]** Extract, clean and load data
- **[DP]** Purchasing Transaction identification
- **[M]** Choose the right algorithm (Apriori, Eclat, FP-growth, etc.)
- **[M]** Choose starging values for **lift**, **support** y **confidence**
- **[M]** Every product/service/department will have a NPS*
- **[E]** Analyze results and adjust parameters
- **[D]** Present and explain the results

Benefits

This technique provides the following benefits:

- Understand purchasing patterns. Example: Who is buying what? Which products are the best ones by region, channel, account,...? Which is the profile?
- Affinity Analysis. Frequent items analysis.
- Propension Analysis. Example: Who will buy what? Which is the profile of those customers?

Use cases

This technique is used in different use cases:

- Online: automated recommendation system, discounts, promotion, placement, cross-selling, etc.

- Offine: product placement, supplier management, store design, product catalog design, bundles, discounts, promotions, etc.
- Customer Retention: it can be used to create compelling arguments (using, for example, product bundles) to retain customers.
- Web usage mining, Intrusion detection, Continuous production, Bioinformatics, etc.

How to implement this algorithm using R

We have several R libraries related to Association Analysis:

- Arules: library to find frequent transactions and associations using Apriori and Eclat algorithms
- ArulesNBMiner: java version
- ArulesSequences: transaction manipulation and cSpade algorithm
- ArulesViz: visualization library for arules

The apriori principle can reduce the number of itemsets we need to examine. Put simply, the apriori principle states that if an itemset is infrequent, then all its subsets must also be infrequent.

- Computationally Expensive. Even though the apriori algorithm reduces the number of candidate itemsets to consider, this number could still be huge when store inventories are large or when the support threshold is low.
- Spurious Associations. Analysis of large inventories would involve more itemset configurations, and the support threshold might have to be lowered to detect certain associations. However, lowering the support threshold might also increase the number of spurious associations detected.

```
# Install packages
install.packages("arules")
install.packages("arulesViz")

# Load packages
library("arulesViz")
library("arules")

# Load data
transaction_data <-
read.csv("data/chapter6.csv",stringsAsFactors=FALSE)

# Review data
summary(edata)

# Consider only
ldata <- unique(edata)

# Separamos los datos
i <- split (ldata$producto, ldata$id_compra)

# Transform dataframe into transaction
txn <- as(i,"transactions")

# Apply apriori algorithm
basket_rules <- apriori(txn,parameter = list(sup = 0.005,
conf=0.01,target='rules'))

# Review outcome
inspect(basket_rules)

# Plot outcome as scatterplot
plot(basket_rules)

# Plot outcome as a graph
plot(basket_rules,
     method="graph",
     measure="confidence",
     shading="lift", control=list(type="items"))
```

```
# Refining our analysis
basket_rules_refined <- apriori(txn,parameter = list(sup = 0.05,
conf = 0.2,target="rules"))
inspect(basket_rules_refined)
plot(basket_rules_refined)
plot(basket_rules_refined,
     method="graph",
     measure="confidence",
     shading="lift", control=list(type="items"))
```

References

- Fast Algorithms for Mining Association Rules
- Association Analysis
- Arules Package
- Apriori Algorithm
- Algorithmic Features of Eclat
- Association Mining with R

Customer Segmentation

- **Problem**: we don't know if we have different types of customers and how to approach them
- **Goals**:
  - We want to understand better our customers
  - We want to have clear criteria to segment our customers
- **Why?** To perform specific actions to improve the customer experience

Technique to solve the business problem

We need a formal definition

> **Customer segmentation** is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests and spending habits.

The most common forms of customer segmentation are:

- **Geographic segmentation**: considered as the first step to international marketing, followed by demographic and psychographic segmentation.
- **Demographic segmentation**:based on variables such as age, sex, generation, religion, occupation and education level.
- **Firmographic**: based on features such as company size (either in terms of revenue or number of employees), industry sector or location (country and/or region).
- **Behavioral segmentation**: based on knowledge of, attitude towards, usage rate, response, loyalty status, and readiness stage to a product.
- **Psychographic segmentation**: based on the study of activities, interests, and opinions (AIOs) of customers.
- **Occasional segmentation**: based on the analysis of occasions (such as bieng thirsty).
- **Segmentation by benefits**: based on RFM, CLV, etc.
- **Cultural segmentation**: based on cultural origin.
- **Multi-variable segmentation**: based on the combination of several techniques.

Main Concepts

Customer Segmentation Techniques

- Single discrete variable (CLV, RFM, CHURN)
- Clustering: K-means, Hierarchical
- Latent Class Analysis (LCA)
- Finite mixture modelling (ex. Gaussian Mixture Modelling)
- Self Organizing maps
- Topological Data Analysis
- PCA
- Spectral Embedding
- Locally-linear embedding (LLE)
- Hessian LLE
- Local Tangent Space Alignment (LTSA)
- Random forests, Decision Trees

Implementation Process

- **[BU]** Determine business needs
- **[DU]** Sourcing, Cleaning & Exploration
- **[DP]** Feature Creation (Extract additional information to enrich the set)
- **[DP]** Feature Selection (Reduce to a smaller dataset to speed up computation)
- **[M]** Select Customer Segmentation Technique (test and compare some of them)
- **[M]** Applied Selected Customer Segmentation Technique
- **[E]** Analyze results and adjust parameters
- **[D]** Present and explain the results

Benefits

This technique provides the following benefits:

- Customer profiling
- Targeted marketing actions
- Targeted operations

Use cases

This technique is used in different use cases:

- Reporting
- Commercial actions: Retention offers, Product promotions, Loyalty rewards
- Operations: Optimise stock levels, store layout

- Pricing: price elasticity
- Strategy: M&A, new products,...

How to implement this algorithm using R

K-means

Given a set of observations (x1, x2, …, xn), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k (≤ n) sets S = {S1, S2, …, Sk} so as to minimize the within-cluster sum of squares (WCSS) (sum of distance functions of each point in the cluster to the K center). In other words, its objective is to find:

$$ \underset{\mathbf{S}} {\operatorname{arg\,min}} \sum*{i=1}^{k} \sum*{\mathbf x \in S_i} \left| \mathbf x - \boldsymbol\mu_i \right|^2

$$

where $$μ_i$$ is the mean of points in $$S_i$$.

Case

We consider the dataset: Wholesale customers Data Set. Abreu, N. (2011). Analise do perfil do cliente Recheio e desenvolvimento de um sistema promocional. Mestrado em Marketing, ISCTE-IUL, Lisbon

This dataset has the following attributes:

- FRESH: annual spending (m.u.) on fresh products (Continuous);
- MILK: annual spending (m.u.) on milk products (Continuous);
- GROCERY: annual spending (m.u.) on grocery products (Continuous);
- FROZEN: annual spending (m.u.) on frozen products (Continuous)
- DETERGENTS_PAPER: annual spending (m.u.) on detergents and paper products (Continuous)
- DELICATESSEN: annual spending (m.u.) on and delicatessen products (Continuous);
- CHANNEL: customers Channel - Horeca (Hotel/Restaurant/Café) or Retail channel (Nominal)
- REGION: customers Region of Lisbon, Oporto or Other (Nominal)

```
# Install packages
install.packages("NbClust")

# Load packages
```

```
library(NbClust)

# Load data
data <- read.csv('data/chapter7.csv', header = T,sep=',')

# Review data structure
str(data)

# Review data
summary(data)

# Scale data
testdata <- data
testdata <- scale(testdata)

# Determine number of clusters. Option 1: visual rule
wss <- (nrow(testdata)-1)*sum(apply(testdata,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(testdata,
                                     centers=i)$withinss)
plot(1:15, wss, type="b", xlab="Number of Clusters",
     ylab="Within groups sum of squares")

# Determine number of clusters. Option 2: more frequent optimal
number
res <- NbClust(data, diss=NULL, distance = "euclidean",
min.nc=2, max.nc=12,
             method = "kmeans", index = "all")

# More information
res$All.index
res$Best.nc
res$All.CriticalValues
res$Best.partition

# K-Means Cluster Analysis (based on the proposed number by
NbCluster)
fit <- kmeans(testdata, 3)

# Calculate average for each cluster
aggregate(data,by=list(fit$cluster),FUN=mean)
```

```
# Add segmentation to dataset
data <- data.frame(data, fit$cluster)
```

References

- Hwang, H., Jung, T. and Suh, E., 2004. An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert systems with applications, 26(2), pp.181-188.
- Kim, S.Y., Jung, T.S., Suh, E.H. and Hwang, H.S., 2006. Customer segmentation and strategy development based on customer lifetime value: A case study. Expert systems with applications, 31(1), pp.101-107. -Marcus, C., 1998. A practical yet meaningful approach to customer segmentation. Journal of consumer marketing, 15(5), pp.494-504.
- Chan, C.C.H., 2008. Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer. Expert systems with applications, 34(4), pp.2754-2762.
- Teichert, T., Shehu, E. and von Wartburg, I., 2008. Customer segmentation revisited: The case of the airline industry. Transportation Research Part A: Policy and Practice, 42(1), pp.227-242.
- Espinoza, M., Joye, C., Belmans, R. and Moor, B.D., 2005. Short-term load forecasting, profile identification, and customer segmentation: a methodology based on periodic time series. Power Systems, IEEE Transactions on, 20(3), pp.1622-1630.
- Wu, J. and Lin, Z., 2005, August. Research on customer segmentation model by clustering. In Proceedings of the 7th international conference on Electronic commerce (pp. 316-318). ACM.
- Machauer, A. and Morgner, S., 2001. Segmentation of bank customers by expected benefits and attitudes. International Journal of Bank Marketing, 19(1), pp.6-18.

Cohort Analysis

- **Problem**: we don't know how customers' behavior changes over time
- **Goals**:
  - Understanding customers' behaviour evolution
  - Understanding behavior using groups that evolve over time, not individually
- **Why?** Perform specific actions on groups, detect similar temporal patterns on group

Technique to solve the business problem

> **Cohort analysis** is a observational, analytic and longitudinal study. It is a comparison of the evolution of a particular aspect (KPI). Individuals comprising study groups are selected based on the presence of a particular characteristic.

Main Concepts

> **Cohort** is a group of people used in a study who have something (such as age or social class) in common

We can find the origin in ancient times. A cohort was a military unit, one of ten divisions in a Roman legion.

Cohort Analysis is a traditional tool in epidemiology. When we applied this technique in other industries most of the times:

- Metrics are easier to capture and analyse
- Direct: number of customers, revenue, cost
- Derived: retention

Implementation Process

- **[BU]** Determine business questions/needs, measure to study and cohorts of interest
- **[DU]** Data Sourcing, Cleaning & Exploration
- **[DP]** Create cohorts, extract data according to cohorts
- **[M]** Calculate the measure
- **[E]** Analyze results and adjust parameters
- **[D]** Present and explain the results

Benefits

This technique provides the following benefits:

- Understand Customer Lifecycle/Journey: length, value, situation,...
- Identify patterns
- Behavioral/Psychographic analysis

Use cases

This technique is used in different use cases:

- Examine where cashflow is coming from and understand the health of your business
- Easily see how much monthly or quarterly revenue is driven from newer and older cohorts
- Study customer retention patterns to see if they are getting better or worse
- Compare cohorts of users from different segments

How to implement this algorithm using R

```
# Install packages
# install.packages("ggplot2")
# install.packages("dplyr")
# install.packages("readxl")
# install.packages("reshape")

# Load packages
library(ggplot2)
library(readxl)
library(dplyr)
library(reshape2)

# Load data into a dataframe
df <- read_excel("data/s8.xlsx")
df

# How many people complete the MOOC
Finished <- df$Finished[49]
Finished
```

```
# Question: Is this a good result?
Ratio <- df$Finished[49] / df$Started[49] *100
Ratio

# Ratio Evolution
RatioEvolution <- data.frame(day = df$Day, ratio=df$Finished /
df$Started *100)
RatioEvolution

# Ratio Evolution Graph
g1 <- ggplot(RatioEvolution, aes(x = day, y = ratio)) +
  geom_line() + ggtitle("Completition Ratio Evolution") +
  ylab("Ratio") + xlab("Period") +
  theme(plot.title = element_text(color="#666666", face="bold",
size=20, hjust=0)) +
  theme(axis.title = element_text(color="#666666", face="bold",
size=14))
g1

# Ratio Evolution Graph vs MOOC objectives
g1 + geom_vline(xintercept=35, colour="red") +
geom_hline(yintercept=20, colour="red")

# How we research about the evolution
df_finished <- dplyr::select(df, contains("Finished"))
df_finished <- data.frame(day = df$Day, df_finished)
df_finished.chart <- melt(df_finished, id.vars = "day")
colnames(df_finished.chart) <- c('Day', 'Cohort', 'Students')

# Let's create a graph
p <- ggplot(df_finished.chart, aes(x=Day, y=Students,
group=Cohort, colour=Cohort))
p + geom_line() + ggtitle('Students Completition per day and
cohort')

# Question: What we observe?

# Let's create another graph
p1 <- ggplot(df_finished.chart, aes(x=Cohort, y=Students,
```

```
group=Day, colour=Day))
p1 + geom_line() + ggtitle('Students Completition per day and
cohort')

# Let's do the same for the completition ratio
df_finished_ratio <- as.data.frame(apply( df_finished, 2,
function(x) x/df$Started*100 ))
df_finished_ratio$day <- df_finished$day
df_finished_ratio.chart <- melt(df_finished_ratio, id.vars =
"day")
colnames(df_finished_ratio.chart) <- c('Day', 'Cohort', 'Ratio')

# Let's create a graph
p2 <- ggplot(df_finished_ratio.chart, aes(x=Day, y=Ratio,
group=Cohort, colour=Cohort))
p2 + geom_line() + ggtitle('Completition Ratio per day and
cohort')

# Let's create another graph
p3 <- ggplot(df_finished_ratio.chart, aes(x=Cohort, y=Ratio,
group=Day, colour=Day))
p3 + geom_line() + ggtitle('Completition Ratio per day and
cohort')

# Question: What we observe?
```

References

- Bent Nielsen. apc: An R Package for Age-Period-Cohort Analysis. The R Journal, 7(2):52-64, Dec. 2015
- Startup Metrics for Pirates
- Lean Analytics
- Cohort Analysis Cheat Sheet
- Data Analytics for Startups - Tetuan Valley Startup School Fall 2015

Churn Analysis

- **Problem**: One of the main problems a company can face is churn. That means the company is losing customers.
- **Goals**:
  - Understanding customers' behaviour evolution
  - Predict churn
  - Identify churn reasons
- **Why?** Perform specific actions on groups to reduce churn

Technique to solve the business problem

- Churn refers to an existing customer deciding to end the business relationship.
- Customer churn is also known as customer attrition, customer turnover or customer defection.
- **Churn analysis** aims to divide customers in active, inactive and "about to churn".
- Churn models predict probability of churn given influencing factors or key factors
- If action is taken to address the factors that influence churn, the model in turn becomes obsolete and must be rebuilt with new churn data and influencing factors.

Main Concepts

There are two types of churn:

- **Voluntary churn**: occurs due to a decision by the customer to switch to another company or service provider
- **Involuntary churn**: occurs due to circumstances such as a customer's relocation to a long-term care facility, death, or the relocation to a distant location.

Involuntary reasons for churn are excluded from the analytical models as most of the times can not be influenced. Any company has what we can call a **natural churn rate** that is unavoidable.

When we analyse churn all customer data is interesting:

- CRM, device usage, interaction data
- clickstream data
- granular credit card, insurance data
- ...

Companies focus on understanding, predicting, reducing and managing churn.

Reasons for churn

- The hardest part is to identify reasons for churn. Many companies use NPS.
- NPS is not enough. We require to determine the importance and weight of the detected factors.
- This can be done using Structural Modelling Equations (SME) and other predictive modelling techniques.

Implementation Process

- **[BU]** Determine business needs: churn evolution, churn prediction, churn reasons
- **[DU]** Data Sourcing, Cleaning & Exploration
- **[M]** Select specific technique
- **[M]** Applied specific technique
- **[E]** Analyze results and adjust parameters
- **[D]** Present and explain the results

Benefits

This technique provides the following benefits:

- Analyze and establish churn profiles and identify critical moments
- Understand the churn causes
- Combine CLV and churn for profiling
- Awake inactive customers based on specific marketing actions
- Retain customers based on specific marketing actions
- Improve customer experience
- Predict future customer behaviour

Use cases

This technique is used in different use cases:

- **Churn for Telecom Providers**: anticipating subscription cancellations and proposing specific commercial actions to foster loyalty
- **Churn for E-commerce Players**: increasing loyalty and client lifetime value by activating personnalized campaigns to "dormant" clients - pushing the right product at

the right time through the right channel
- **Churn for Banking & Insurance Companies**: predicting life events from behavioral data to anticipate structural changes in the client's consumption profile that may signal churn or upsell / cross-sell opportunities
- **Generic**: create alarms (based on churn changes), create churn critical path (when, where, how)
- **Churn applied to employees**: in this case we speak about Employee Churn.

How to implement this algorithm using R

Once we know that some users have left the services of our company it is time to understand the evolution of the rest of the customer based on previous data. We can understand this using survival analysis.

```
# Install packages
install.packages("survival")
install.packages("ggplot2")
if(!require(devtools)) install.packages("devtools")
devtools::install_github("kassambara/survminer")

# Load packages
library(survival)
library(ggplot2)
library(survminer)

# Load Data
df <- read.csv('data/s9.csv')

# Question: Do we have some insights from data exploration?

# Let's create a survival curve
fit <- survfit(Surv(time, churned) ~ 1, data = df)

# Let's create a graph
ggsurvplot(fit)

# Let's improve the graph
g1 <- ggsurvplot(fit,
           color = "#2E9FDF",
```

```
            ylim=c(.75,1),
            xlab = 'Days since subscription',
            ylab = '% Survival')
g1

# Question: what happens?

# Now with two groups (based on gender)
fit2 <- survfit(Surv(time, churned) ~ female, data = df)
ggsurvplot(fit2)

# Let's improve the graph
g2 <- ggsurvplot(fit2, legend = "bottom",
            legend.title = "Gender",
            conf.int = TRUE,
            pval = TRUE,
            ylim=c(.75,1), lty = 1:2, mark.time = FALSE,
            xlab = 'Days since subscription', ylab = '%
Survival',
            legend.labs = c("Male", "Female"))
g2

# We can add the risk table
g3 <- ggsurvplot(fit2, legend = "bottom",
                 legend.title = "Gender",
                 conf.int = TRUE,
                 pval = TRUE,
                 ylim=c(.75,1), lty = 1:2, mark.time = FALSE,
                 xlab = 'Days since subscription', ylab = '%
Survival',
                 legend.labs = c("Male", "Female"),
                 risk.table = TRUE, risk.table.y.text.col =
TRUE)
g3

# Question: what happens?
```

References

- Zero Defections: Quality comes to service, Reichheld & Sasser. Harvard Business review. 1990.
- Van Den Poel; Lariviére (2004). Customer Attrition Analysis For Financial Services Using Proportional Hazard Models. European Journal of Operational Research 157: 196–217.
- Applying and evaluating models to predict customer attrition using data mining techniques, Tom Au, et al. Journal of Comparative International Management. 1 June 2003
- Mittal, Vikas and Sarkees, Matthew, Customer Divestment (2006). Journal of Relationship Marketing, 5(2/3), 71-85, 2006.
- Mittal, Vikas and Sarkees, Matthew and Murshed, Feisal, The Right Way to Manage Unprofitable Customers (April 1, 2008). Harvard Business Review, Vol. 86, No. 4, 2008.
- Buckinx Wouter, Dirk Van den Poel (2005), Customer Base Analysis: Partial Defection of Behaviorally-Loyal Clients in a Non-Contractual FMCG Retail Setting, European Journal of Operational Research, 164 (1), 252-268.
- Employee Churn

About the author

Josep is a consultant, research analyst, entrepreneur and professor who has helped many companies to create competitive advantages based on data. He believes that the combination of data, models, technology and experience is the ultimate combination to crack everyone's passion and purpose. He is the CDSO (Chief Data Science Officer) of the Institute of Passion.

His expertise comprises data technologies, methodologies and strategies and how they can be applied to solve business problems, improve decision-making, discover hidden insights, validate business hypothesis and create data products. He is the founder and CEO of Delfos Research and collaborates at City_EX. He complements his professional career with a passion for higher education being professor at IE Business School and Universitat Oberta de Catalunya (UOC) among other institutions.

He is author of many articles, academic notes and books related to his professional expertise. His know-how is the result of developing consultancy and research projects for multiples industries. His interests focus on the adoption, implementation and use of data strategies and their impact on organizations, processes, and people.

Born in Spain and married to an Argentinian. He currently lives in UK with his family.

Josep holds a BSc (Hons) in Mathematics, a MSc (Hons) in Business Intelligence, and a MSc (Hons) in IT Management, as well as a top tier MBA. Josep aims to deeply understand how organizations are developing successful big data strategies by conducting research in the field of Technology & Management as part of a PhD.

Josep's motto in life is: "All data models are wrong by definition, some are useful. Let's use them wisely."

Contat Josep at:

- @josepcurto
- LinkedIn
- Web
- Github