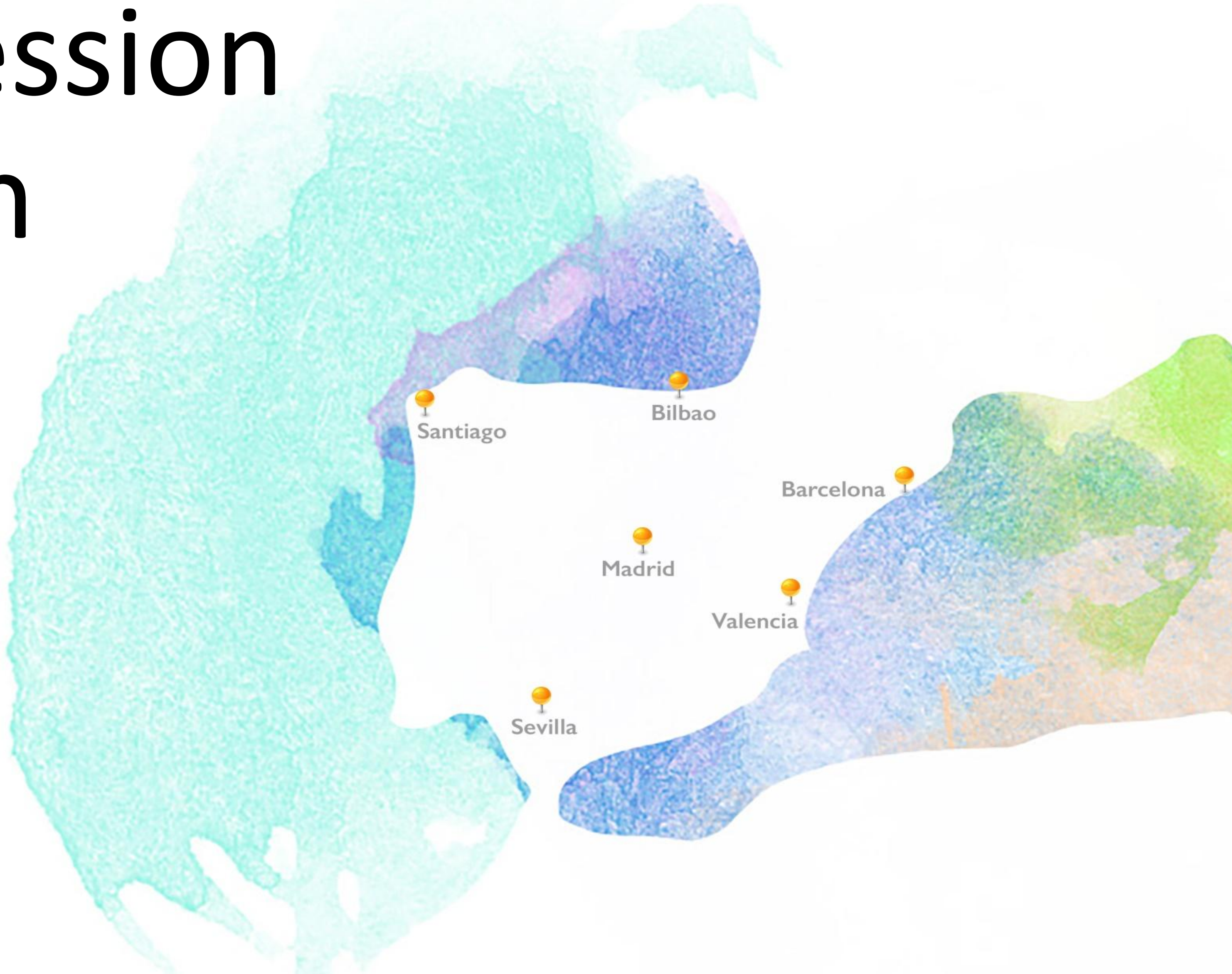
Logistic Regression Meets Fintech

Cristian Pachon, Jelena Mirkovic

Machine Learning Study Meetup
Barcelona, 21 October, 2015



digitalOrigin

Outline

- About us

- PART I – Logistic regression & risk scoring (Cristian)
  - Odds and odds ratio
  - Logistic regression model
    - Interpretation of the parameters and the relationship with OR
    - Goodness of Fit
  - LR in practice: R framework

- PART II – Regression models applied in marketing (Jelena)
  - Generalized linear models
  - Marketing basics for techies
  - Measuring TV marketing campaigns impact

digitalOrigin

Fintech is unbundling the banking offering

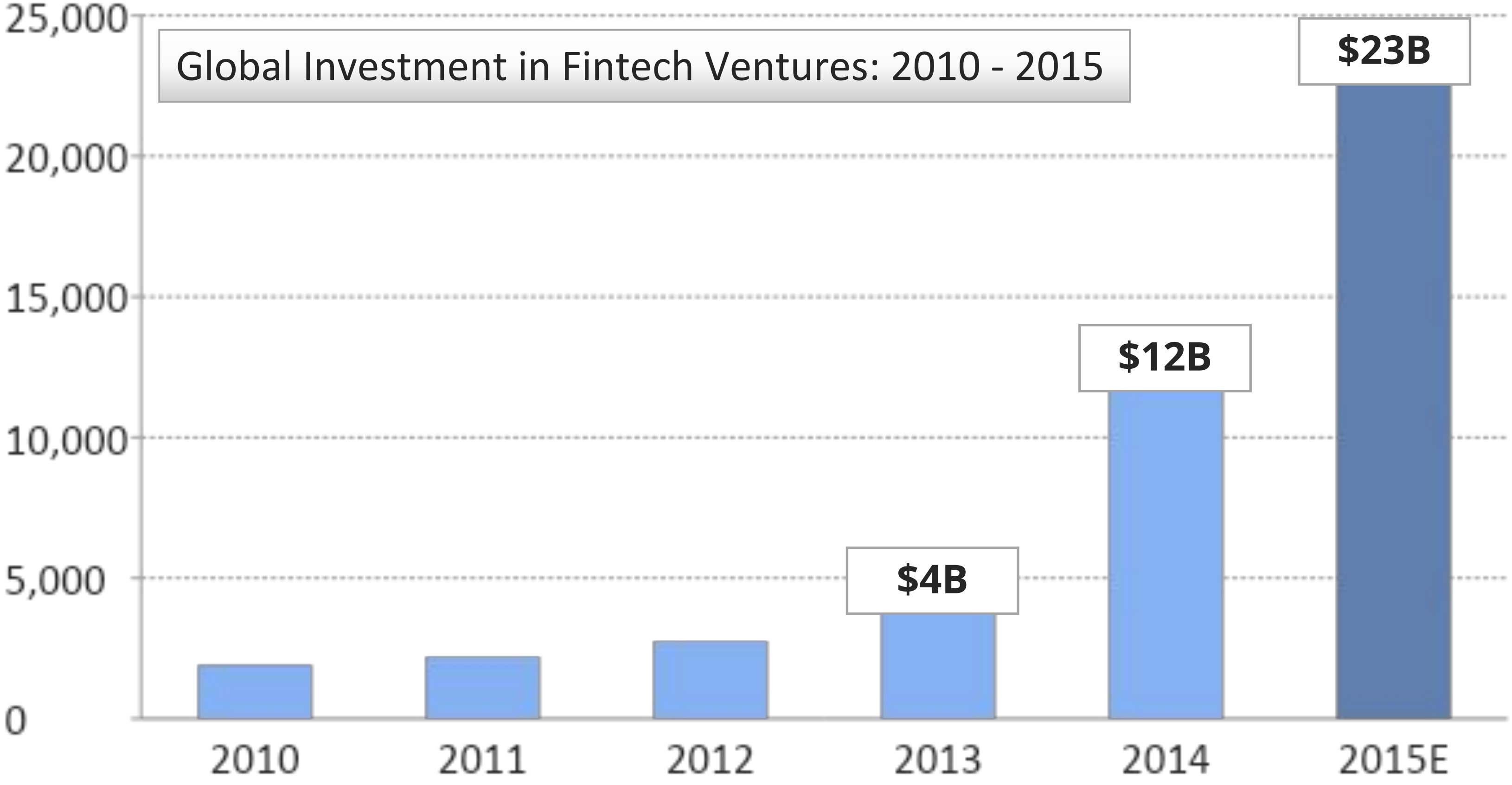
BANK

Foreign exchange	peerTransfer International education payments made easy	TransferWise
Consumer loans	paga +tarde	LendingClub
Business loans	OnDeck	CAN CAPITAL
Transfers	venmo	Snapcash
Payments	Square	Braintree
Wealth management	LEARNVEST	acorns

Source: CBInsights

digitalOrigin

The sector is attracting significant investment…

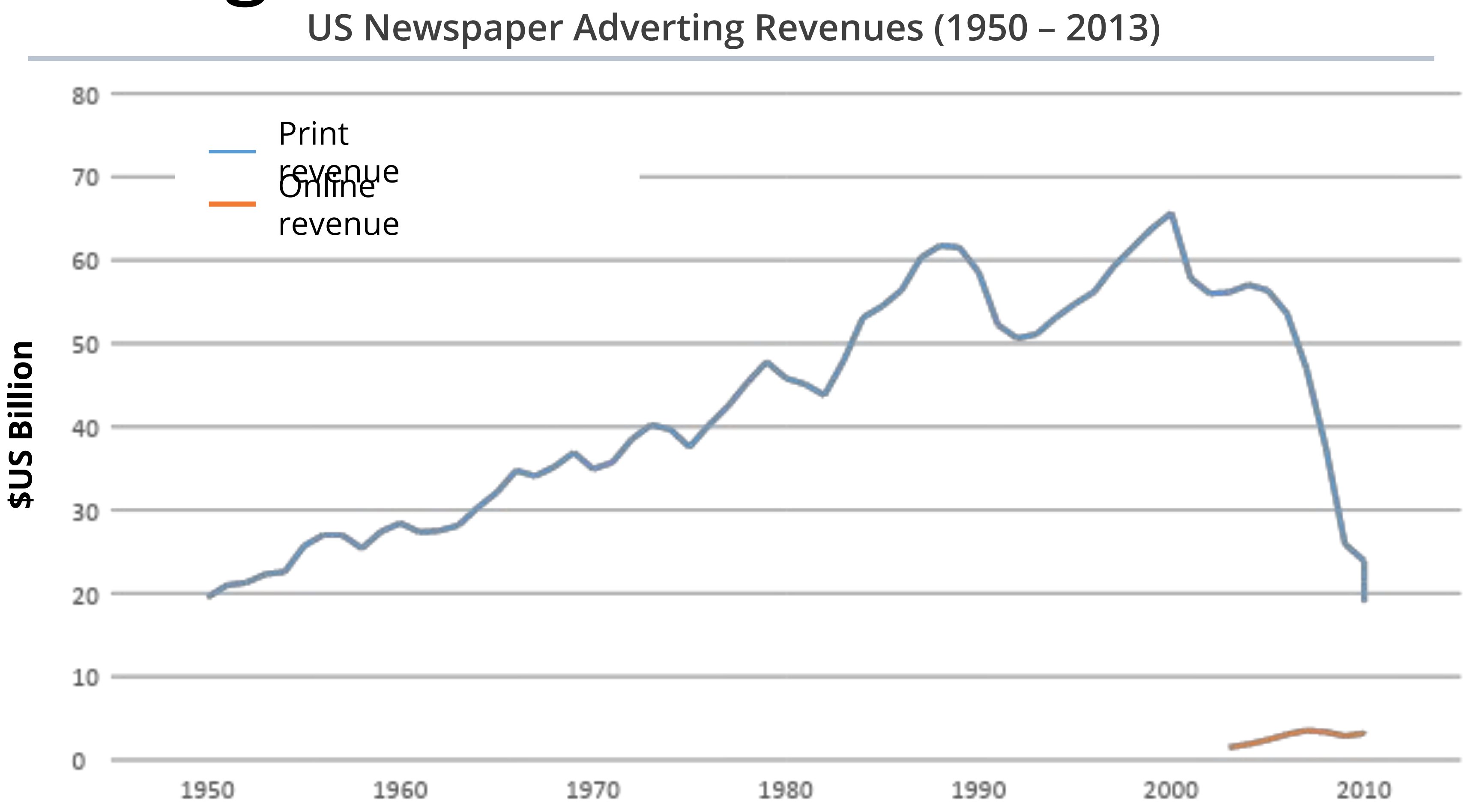


- The biggest issue: regulation!

Source: accenture The Future of Fintech and Banking, March 2015 + CBInsights

digitalOrigin

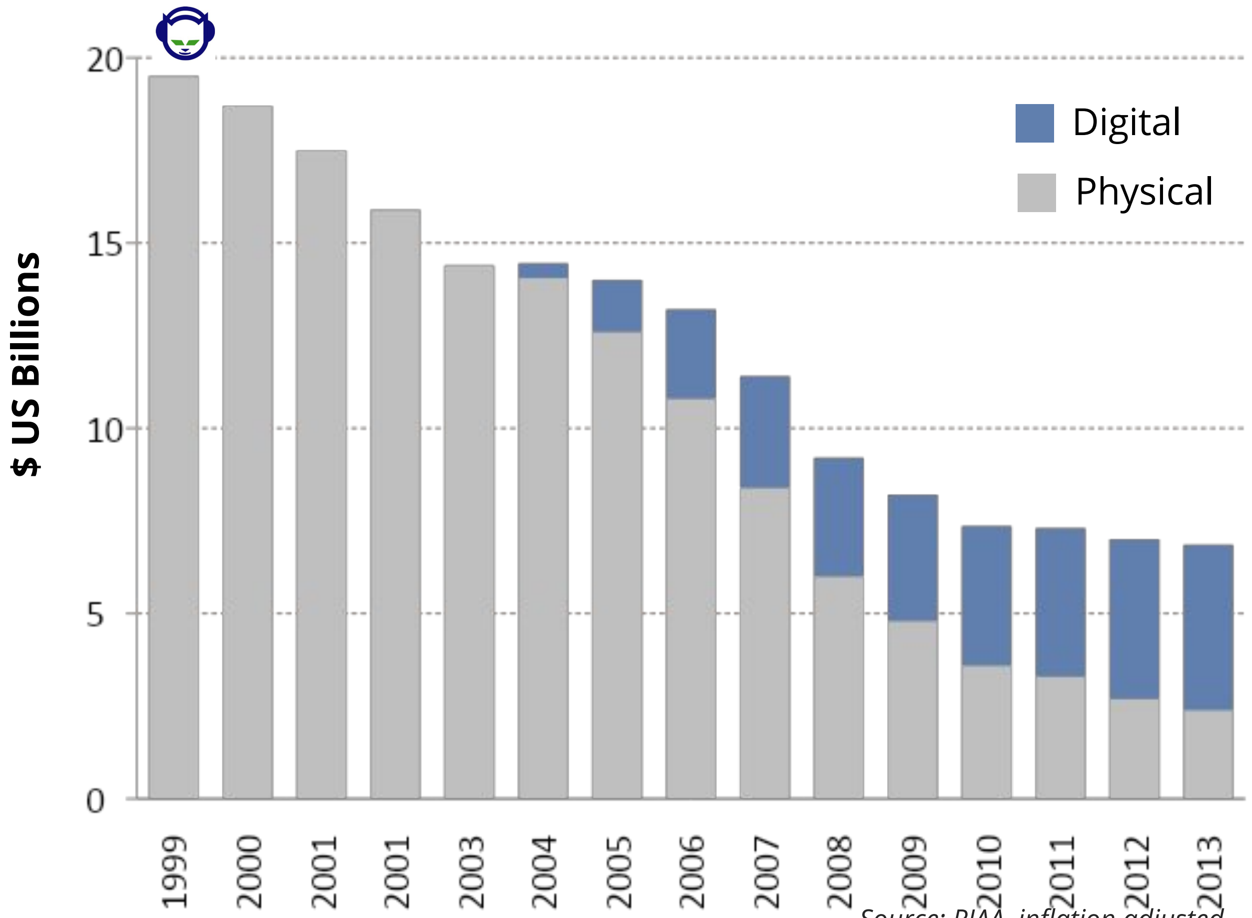
The impact of the Internet can be devastating



Source: Newspaper Association of America, revenue figures adjusted for inflation

…and incumbents do not always win

US Music Industry Revenues (1999 – 2013)



Source: RIAA, inflation adjusted

Spotify (founded in 2011) **more valuable than** Universal Music Group (39% of music industry)*

** Nielsen Soundscan 2012 report*

digitalOrigin

Our Business: Originating Loans 24/7!

- Online, paperless lending
  - QueBueno.es – overdraft solution, microcredits
  - PagaMasTarde.es – installment payments
  - NFC-based mobile payments & more coming soon!

Our Business: Originating Loans 24/7!

- Online, paperless lending
  - QueBueno.es – microcredits, overdraft solution
  - PagaMasTarge.es – installment payments for e-commerce
  - NFC-based mobile payments & more coming soon!
- Risk-based pricing

Outline

- About us

- PART I – Logistic regression & risk scoring (Cristian)
  - Odds and odds ratio
  - Logistic regression model
    - Interpretation of the parameters and the relationship with OR
    - Goodness of Fit
  - LR in practice: R framework

- PART II – Regression models applied in marketing (Jelena)
  - Generalized linear models
  - Marketing basics for techies
  - Measuring TV marketing campaigns impact

digitalOrigin

The Model – Introduction

- A loan is in a Defualt situation when it has not been paid.
- From a Statistic point of view, the Default is a random variable represented by:

$$D = \begin{cases} 1 & \text{if the user has not paid the loan (D)} \\ 0 & \text{otherwise (D)} \end{cases}$$

The Model – Introduction

Goal

- We are interested in finding an expression that allows us to establish a relationship between D and the rest of variables.

Age	Gender	LABORAL	**Default**
19	M	EMPLOYED	**0**
34	M	UNEMPLOYED	**0**
43	F	SELFEMPLOYED	**1**
29	F	STUDENT	**0**
35	F	STUDENT	**0**
56	M	EMPLOYED	**1**
…	…	…	**…**
32	M	SELFEMPLOYED	**0**

The Model – Previous concepts

Odds

- Let D be the outcome of interest (in our case, it is the Default variable) . Then, the odds(D) is defined as:

$$odds(D) = \frac{P(D)}{1 - P(D)}$$

digitalOrigin

The Model – Previous concepts

Odds

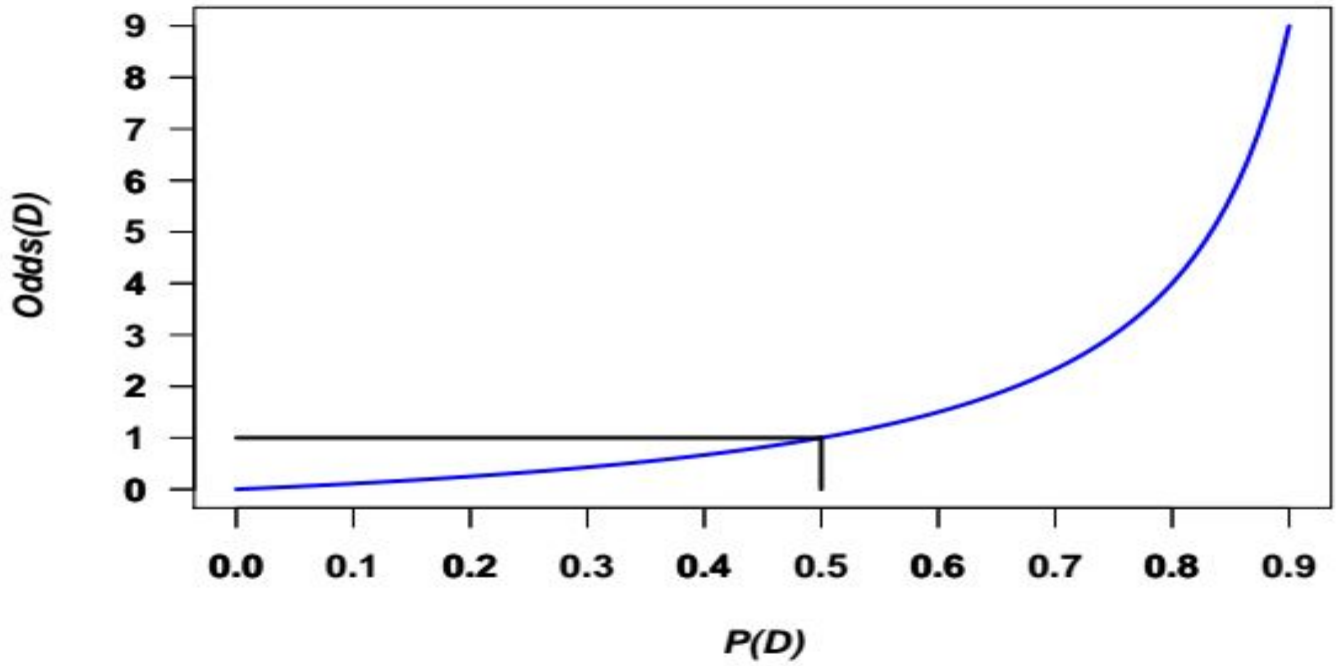
- For example:

- P(D) = 0.2 → odds(D) = 0.2/0.8 = 1/4 (= 1: 4)
- P(D) = 0.5 → odds(D) = 0.5/0.5 = 1 (= 1: 1)
- P(D) = 0.6 → odds(D) = 0.6/0.4 = 3/2 (= 3: 2)

- The odds is often used to describe the chance of winning a game

The Model – Previous concepts

Odds



The Model – Previous concepts

Odds

- Default : $\begin{cases} 1 \text{ (not paid)} \\ 0 \text{ (paid)} \end{cases}$ X: $\begin{cases} 1 \text{ if Gener = Male} \\ 0 \text{ Otherwise} \end{cases}$

	D	$\overline{\text{D}}$	Total
X = 1 (M)	45	105	150
X = 0 (F)	40	160	200
Total	85	265	350

$$Odds(D \mid X=0) = \frac{P(D \mid X=0)}{1-P(D \mid X=0)} = \frac{40/200}{1-40/200} = 0.25$$

$$Odds(D \mid X=1) = \frac{P(D \mid X=1)}{1-P(D \mid X=1)} = \frac{45/150}{1-45/150} = 0.43$$

digitalOrigin

The Model – Previous concepts

Odds Ratio

- Default : $\begin{cases} 1 \text{ (not paid)} \\ 0 \text{ (paid)} \end{cases}$ X: $\begin{cases} 1 \text{ if Gender = Male} \\ 0 \text{ Otherwise} \end{cases}$

$$OR = \frac{odds(D \mid X = 1)}{odds(D \mid X = 0)}$$

digitalOrigin

The Model – Previous concepts

Odds Ratio

- Given the following contingency table, the Odds Ratio is obtained as follows:

	D	$\overline{D}$	Total
X = 1 (M)	a	b	a+b
X = 0 (F)	c	d	c+d
Total	a+c	b+d	a+b+c+d

$$\longrightarrow \quad OR = \frac{a \cdot d}{b \cdot c}$$

digitalOrigin

The Model – Previous concepts

The Odds Ratio

	D	$\overline{D}$	Total
X = 1 (M)	45	105	150
X = 0 (F)	40	160	200
Total	85	265	350

$$OR = \frac{45 \cdot 160}{105 \cdot 40} = 1.71$$

- That is, the odds of the default is 1.71 times higher among males as compared with females.

The Model – Previous concepts

The Odds Ratio

- Comments
  - When OR>1, we say that the variable is a Risk factor.
  - When OR = 1, there is not relation between the variable and the Default.
  - It is important to check the confidence interval of the Odds Ratio to see if 1 is included.

The Model – General Expression

- Let D be the Default

$$D = \begin{cases} 1 & \text{if Loan = Default} \\ 0 & \text{otherwise} \end{cases}$$

- Let X a covariate vector

(age, gender, laboral status ...)

Age	Gender	LABORAL	Default
19	M	EMPLOYED	0
34	M	UNEMPLOYED	0
43	F	SELFEMPLOYED	1
29	F	STUDENT	0
35	F	STUDENT	0
56	M	EMPLOYED	1
...	...	...	...
32	M	SELFEMPLOYED	0

X: Covariate Vector
D: Default

digitalOrigin

The Model – General Expression

Goal

- Our goal is to find an expression that allows us to establish a relationship between D and the covariate vector X

Expression of the logistic regression model

- Let D be the default and X de age. The first idea that we have:

$$D = \alpha + \beta \cdot X$$

- A linear model such as the previous one is not meaningful!!!!

The Model – General Expression

$$P(D=1\mid X) = \alpha + \beta \cdot X$$

$$P(D=1\mid X) \in [0,1] \qquad \alpha + \beta \cdot X \in \Re$$

$$odds(D=1\mid X) = \frac{P(D=1\mid X)}{1-P(D=1\mid X)} = \alpha + \beta \cdot X$$

$$\frac{P(D=1\mid X)}{1-P(D=1\mid X)} \in [0,+\infty) \qquad \alpha + \beta \cdot X \in \Re$$

digitalOrigin

The Model – General Expression

- We solve this problem if we model the log of the odds:

$$\ln[odds(D=1 \mid X)] = \ln\left[\frac{P(D=1 \mid X)}{1-P(D=1 \mid X)}\right] = \alpha + \beta \cdot X$$

General Expression

- logit function of p = P(D=1|X) as a linear combination of $X_1, \ldots, X_{n:}$

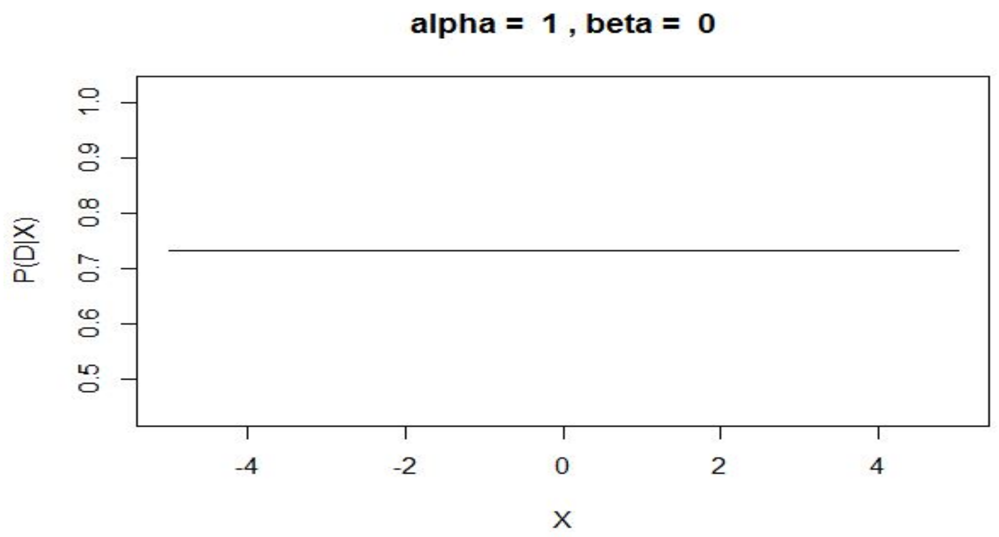
$$\log it(p) = \ln\left[\frac{p}{1-p}\right] = \alpha + \beta_1 \cdot X_1 + \cdots \beta_n \cdot X_n$$

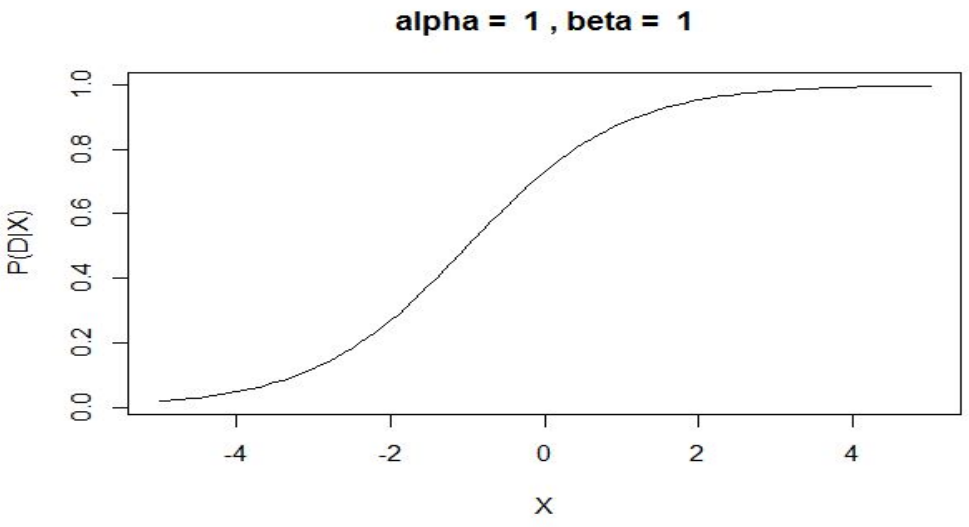
$$p = \frac{\exp(\alpha + \beta_1 \cdot X_1 + \cdots \beta_n \cdot X_n)}{1 + \exp(\alpha + \beta_1 \cdot X_1 + \cdots \beta_n \cdot X_n)}$$

The Model

Logistic Curve

$$P(D \mid X) = \frac{\exp(\alpha + \beta \cdot X)}{1 + \exp(\alpha + \beta \cdot X)}$$



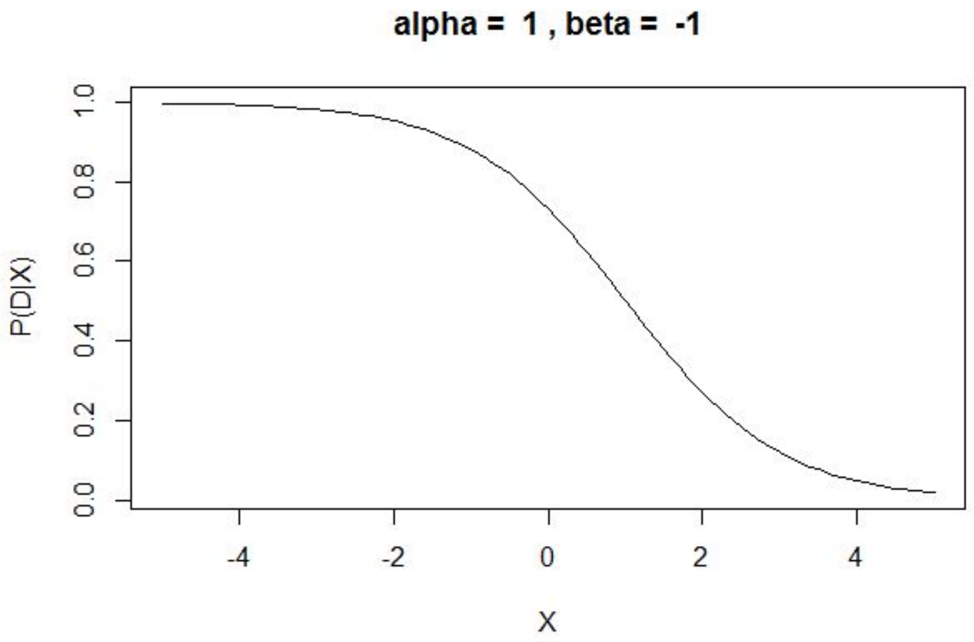


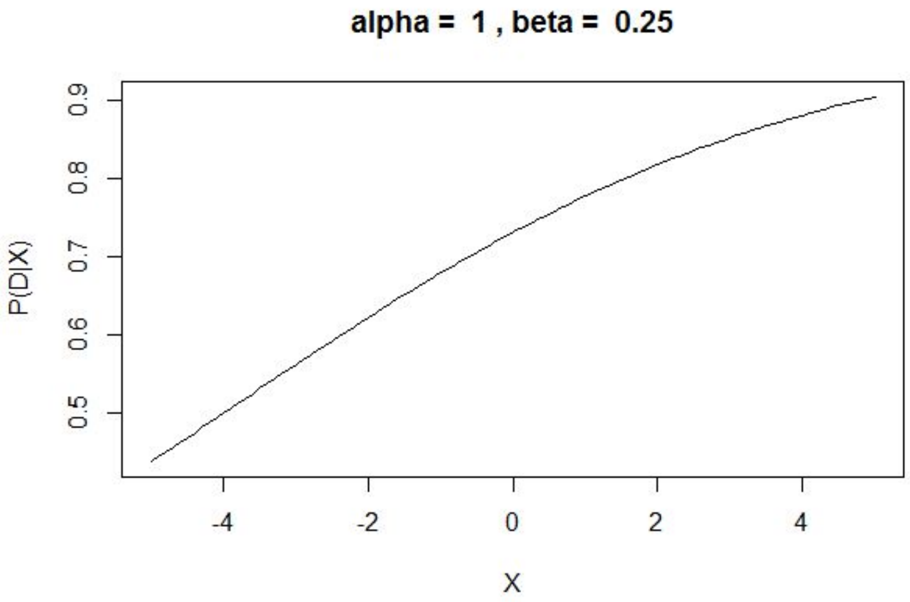
digitalOrigin

The Model

Logistic Curve

$$P(D \mid X) = \frac{\exp(\alpha + \beta \cdot X)}{1 + \exp(\alpha + \beta \cdot X)}$$





The Model

- If we model P(D=0|X) : the sign of the coefficient changes.
- Covariates: categorical variables as well as continuous variables.
- Dummy coding is used when categorical variables are included in a regression model. Let X be the gender:

$$X_1 = \begin{cases} 1 \text{ if Gender is male} \\ 0 \text{ otherwise} \end{cases}$$

$$\ln\left[\frac{P(D=1 \mid X_1)}{1-P(D=1 \mid X_1)}\right] = \alpha + \beta \cdot X_1$$

digitalOrigin

The Model

- In the financial industry, this model is known as scoring model, and the right hand term is called score ($\alpha + \beta \cdot X$)
- Another model, similar to this one, is the probit model. In this case the expression is:

$$\phi^{-1}(p) = \alpha + \beta \cdot X$$

where φ is the probability distribution function of a N(0, 1)

The Model – Interpretation of the parameters

Dichotomic variable

- Let X_k be a dichotomic covariate of a logistic regression model. The odds ratio associated with X_k = 1:

$$OR = \frac{odds(D=1 \mid X_1, \ldots, X_k = 1, \ldots, X_n)}{odds(D=1 \mid X_1, \ldots, X_k = 0, \ldots, X_n)} = \exp(\beta_k)$$

Since:

$$\ln\left[odds(D \mid X_1, \ldots, X_n)\right] = \alpha + \beta_1 \cdot X_1 + \cdots \beta_n \cdot X_n$$

The Model – Interpretation of the parameters

Continuous variable

- Let X_k be a continuous variable. Then, the odds ratio associated with comparing two levels which differ **c** units is:

$$OR = \frac{odds(D=1 \mid X_1,\ldots,X_k = x + c,\ldots,X_n)}{odds(D=1 \mid X_1,\ldots,X_k = x,\ldots,X_n)} = \exp(c \cdot \beta_k)$$

Since:

$$\ln\left[odds(D \mid X_1,\ldots,X_n)\right] = \alpha + \beta_1 \cdot X_1 + \cdots \beta_n \cdot X_n$$

digitalOrigin

The Model – Interpretation of the parameters

The model constant

- The interpretation of the model constant α is related to the probability for D = 1 in case of an individual with zero values in all covariates:

$$p = \frac{\exp(\alpha + \beta_1 \cdot X_1 + \cdots \beta_n \cdot X_n)}{1 + \exp(\alpha + \beta_1 \cdot X_1 + \cdots \beta_n \cdot X_n)} \Longrightarrow p_0 = P(D \mid X = 0) = \frac{\exp(\alpha)}{1 + \exp(\alpha)}$$

Which is the same as: $$\frac{p_0}{1 - p_0} = \exp(\alpha)$$

digitalOrigin

The Model – Interpretation of the parameters

The model constant

- To provide the model constant with a relevant meaning, continuous covariates may be centered, for example, around their means:

  $Y = X - \overline{X}$

$$p_0 = P(D=1 \mid Y=0) = P(D=1 \mid X=\overline{X}) = \frac{\exp(\alpha)}{1+\exp(\alpha)}$$

The Model – Checking Goodness of fit

The Hosmer-Lemeshow (HL) Test

- If the model is well specified, the number of events predicted should be similar to the ones observed.
- The HL test works as follows:
    - It sorts the observations according to the estimated probability.
    - They are divided into 5 to 10 groups of the same size.

The Model – Checking Goodness of fit

The Hosmer-Lemeshow (HL) Test

- Within each of these n group, the observed number of events O_k k = 1, …, n is compared to the expected number E_k

$$E_k = \sum_{j=1}^{R_k} p_j$$

Where Rk is the number of observations of the group k.

- Finally: $$\chi^2_{HL} = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{V_k} \sim \chi_{n-2}, \qquad V_k = \sum_{j=1}^{R_k} p_j (1 - p_j)$$

digitalOrigin

Outline

- About us

- PART I – Logistic regression & risk scoring (Cristian)
  - Odds and odds ratio
  - Logistic regression model
    - Interpretation of the parameters and the relationship with OR
    - Goodness of Fit
  - LR in practice: R framework

- PART II – Regression models applied in marketing (Jelena)
  - Generalized linear models
  - Marketing basics for techies
  - Measuring TV marketing campaigns impact

Generalized Linear Models (GLzM)

- Linear Model: $$Y = \beta X + \varepsilon$$
- GLzM: when *Y* goes not follow the normal distribution:
$$E(Y) = g^{-1}(\beta X)$$
    - g – link function
- The model has very nice properties if *Y* follows a distribution from the exponential family!!
    - Normal, exponential, gamma, Poisson, Bernoulli, Binomial…
    - Canonical link function
- Logistic regression as a special case of GzLM
    - Y follows Bernoulli: link function is logit functioN

Marketing

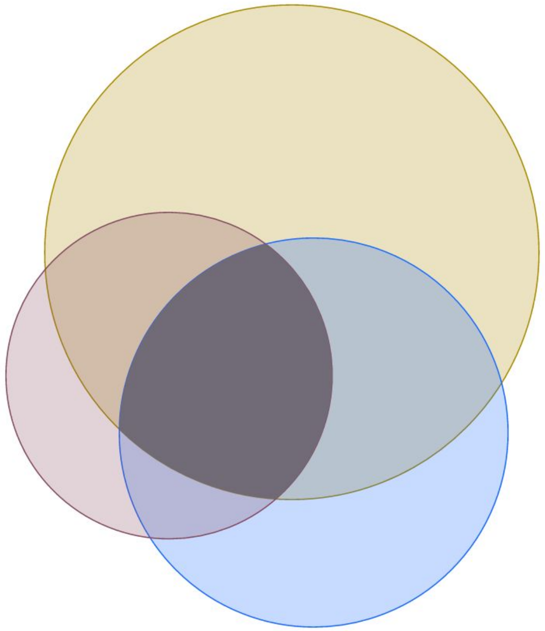
- Online & offline marketing campaigns
  - From the first visit to the conversion
  - Attribution modeling

Multi-Channel Conversion Visualizer

See the percentage of conversion paths that included combinations of the channels below. Select up to four channels.

	Channel	% of total conversions
☑	Direct	74.84%
☑	Paid Search	46.31%
☑	Referral	32.61%
☐	Organic Search	17.06%
☐	Other Advertising	0.18%
☐	Social Network	0.08%

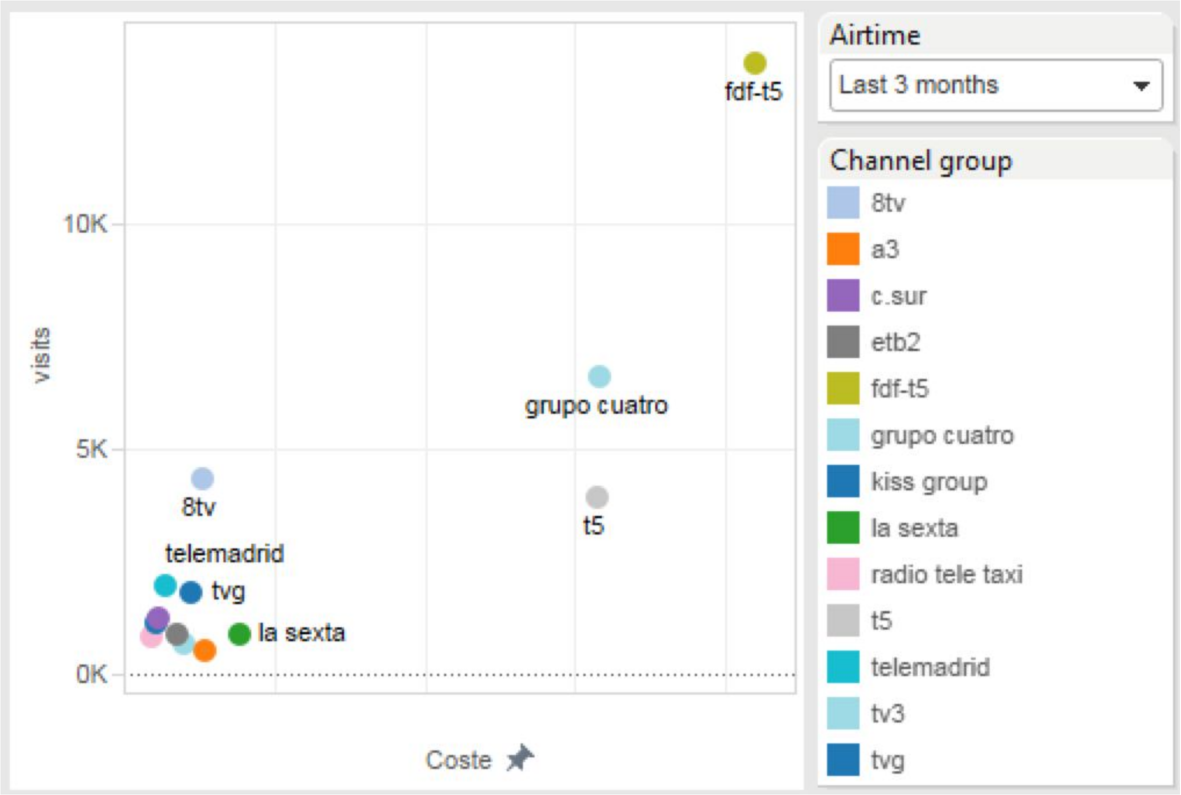
Direct & Paid Search & Referral: 12% (13895)



- How to evaluate offline campaigns?
  - i.e. TV campaigns

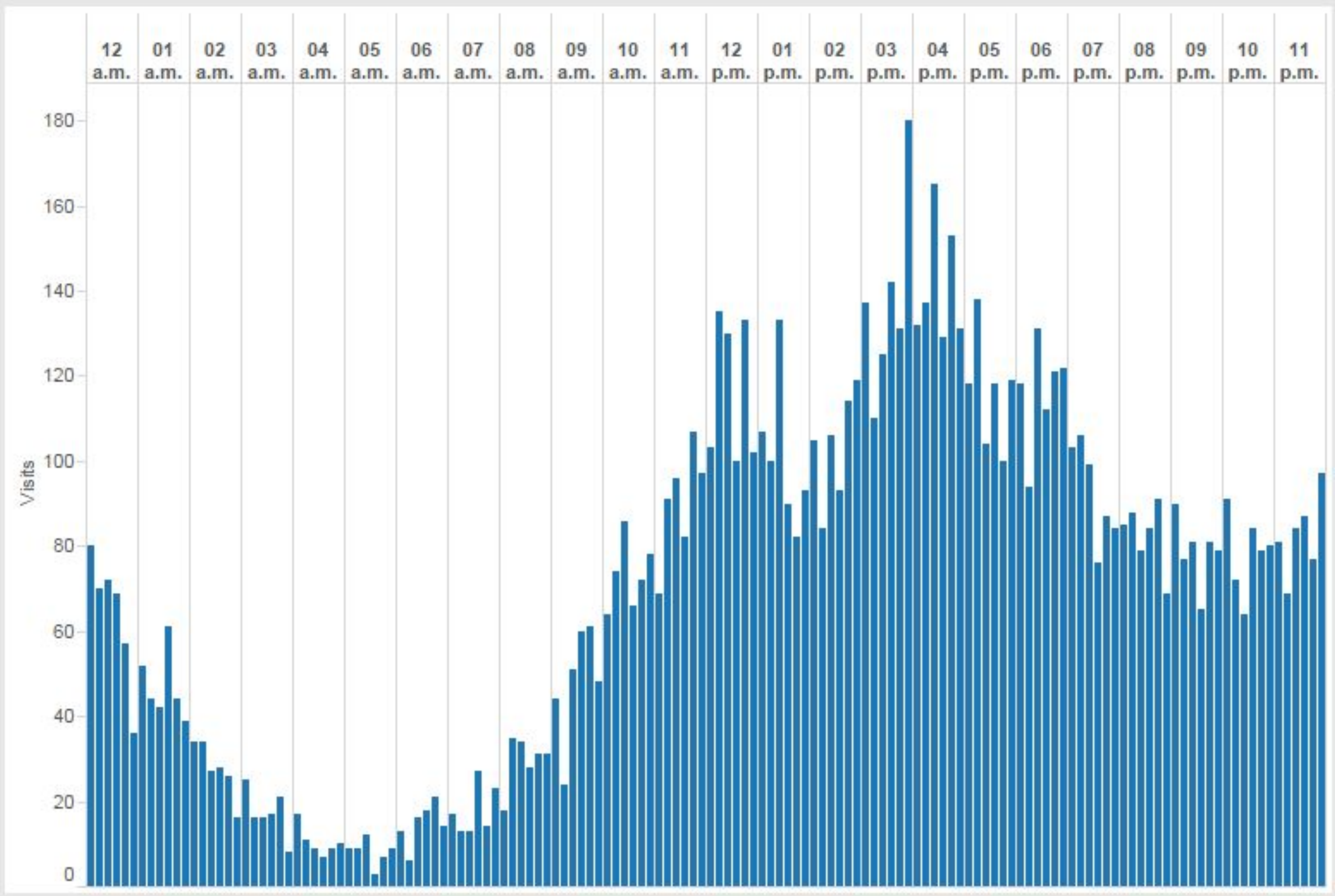
TV Marketing

- Impact of TV campaigns
  - Branding vs. direct response



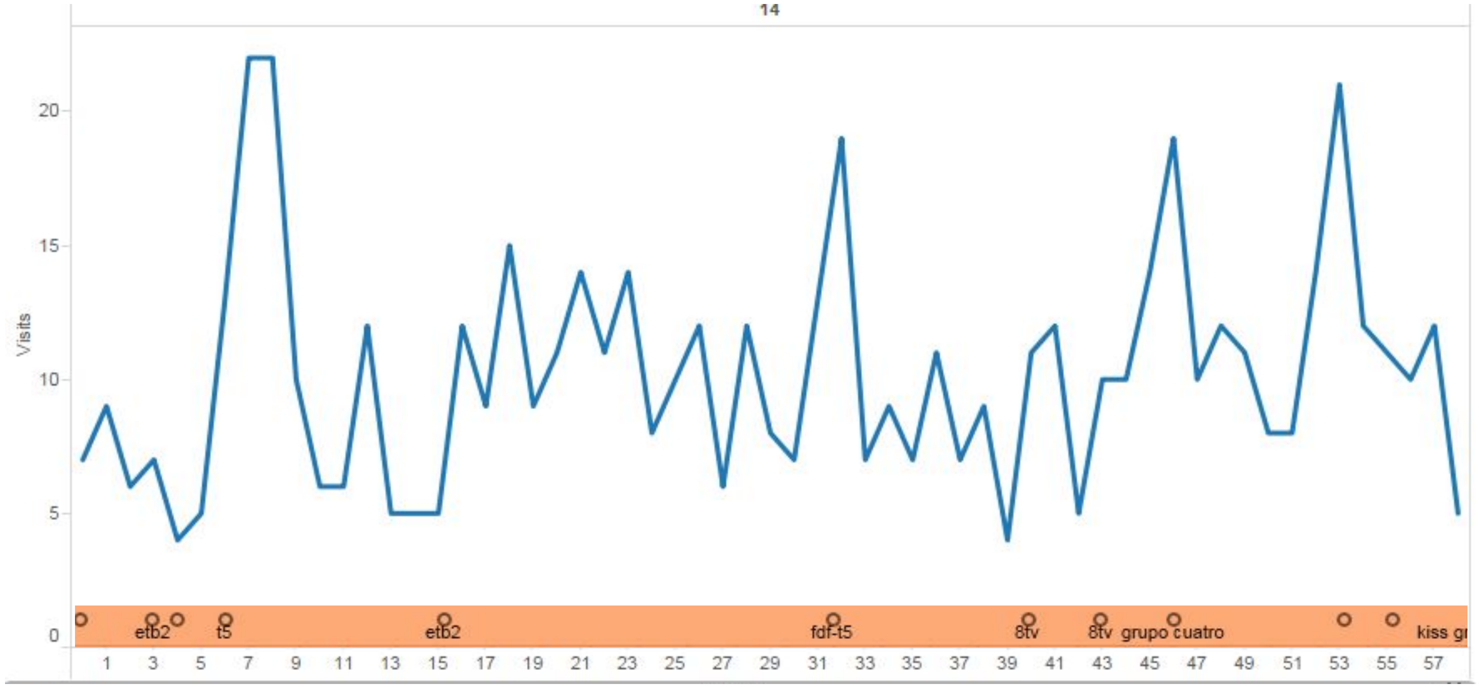
Web Traffic

- Organic + Brand



digitalOrigin

TV Spot Impact



- Questions:
  - What is it impact duration?
  - What is the baseline traffic?
  - Overlapping spots: how to attribute visits to different spots?

digitalOrigin

GzLM for modeling TV visits

- GLzM model: $E(Y) = g^{-1}(\beta X)$
- Y: TV visits
  - Y follows Poisson distribution / from the exponential family
- X: TV spots on different channels, month, day of the month/week,
- β: estimated coefficients, directly corresponding to the impact of an ad
  - Intercept indicates the branding awareness
- Comments:
  - Negative coefficients, small channels, “Autonomicas”: geographically limited impact
- R: glm, nls (non-linear least square)

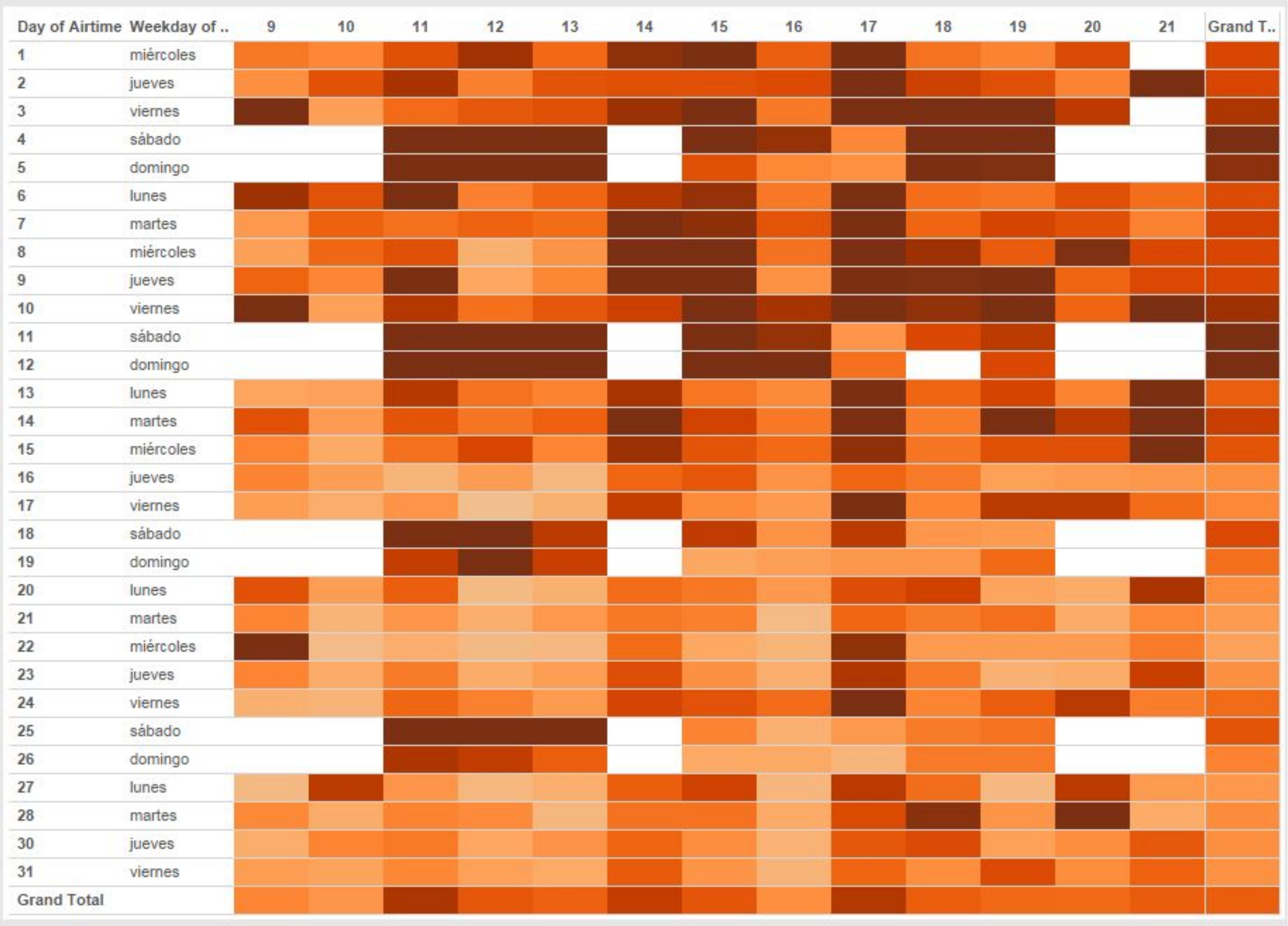
Cost Effectiveness Of TV Spots

Color:

The darker,
the more expensive

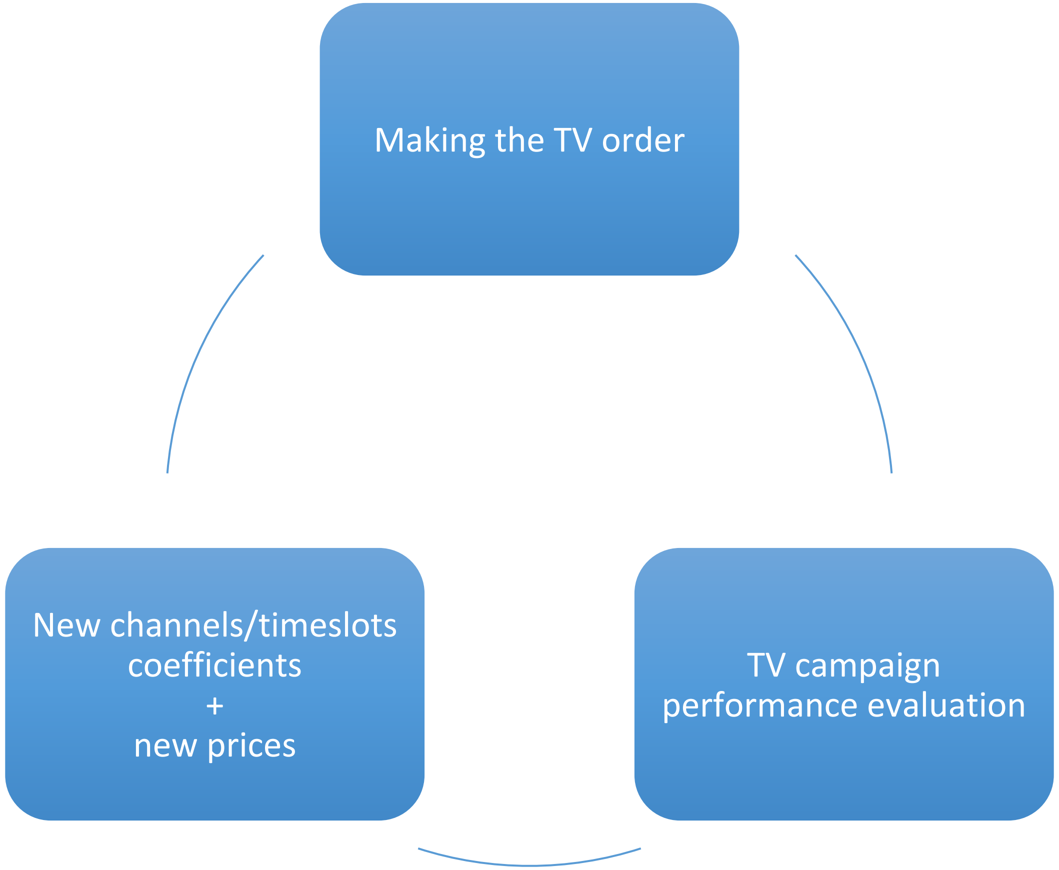
Major factors:

- Day of month
- Hour
- Weekday



digitalOrigin

TV Marketing Optimization Cycle



digitalOrigin

Thank you!

Questions?

cristian@digitalorigin.com
jmirkovic@digitalorigin.com

digitalOrigin