

# Bank Marketing data (with social/economic context)

Maciej Maecki

24 października 2019

## Streszczenie

W pliku Bank Marketing data.csv znajduj si dane charakteryzujce klientw pewnego banku oraz kampanie marketingowe skierowane do tych klientw. Doczone s ponadto wskaniki spoeczne i ekonomiczne. Na podstawie tych danych naley zbudowa model prognozujcy szans, e klient w wyniku prowadzonej kampanii zaoy lokat terminow.

## Spis treści

<b>1</b>	<b>Wprowadzenie</b>	<b>2</b>
1.1	Opis problemu . . . . .	2
1.2	Opis danych . . . . .	2
1.3	Wstpna eksploracja danych . . . . .	3

# 1 Wprowadzenie

## 1.1 Opis problemu

W ramach kampani marketingowej organizowanej przez pewien bank w latach między majem 2008 roku, a listopadem 2010 roku, były zbierane informacje na temat klientów tego banku. Na podstawie tych danych planowane jest przewidzenie, czy i jaki rodzaj klientów kupi lokat terminową w tym banku.

## 1.2 Opis danych

Nasze dane zawierają 21 kolumn danych. Kolumny możemy podzielić na 3 grupy:

### **I: Zmienne związane z danymi klienta bankowego:**

1. Wiek (age): wiek klienta.
2. Praca (job): rodzaj pracy klienta.
3. Stan cywilny (marital): stan cywilny klienta.
4. Edukacja (education): edukacja klienta.
5. Domylnie (default): Klient wcześniej domylnie miał kredyt.
6. Mieszkanie (housing): Klient ma kredyt mieszkaniowy.
7. Poyczka (loan): Klient ma osobistą pożyczkę.

### **II: Zmienne związane z ostatnim kontaktem bieżącej kampanii marketingowej:**

8. Kontakt (contact): Typ komunikacji kontaktowej (telefonicznej lub komrkowej).
9. Miesiąc (month): Ostatni kontakt miesiąca roku.
10. Dzie tygodnia (day of week): dzień ostatniego kontaktu tygodnia.
11. Czas trwania (duration): czas trwania ostatniego kontaktu w sekundach. Jeśli czas trwania wynosi 0, nigdy nie skontaktowaliśmy się z klientem, aby założyć konto lokaty terminowej.
12. Kampania (campaign): liczba kontaktów wykonanych podczas tej kampanii i dla tego klienta
13. Liczba dni (pdays): liczba dni, które upłynęły od ostatniego kontaktu klienta z poprzednią kampanią (warto liczbowa; 999 oznacza, że klient wcześniej się nie skontaktował)
14. Poprzedni (previous): liczba kontaktów wykonanych przed tą kampanią i dla tego klienta (numerycznie)
15. Poutcome: wynik poprzedniej kampanii marketingowej (kategorycznie: porażka, nieistniejąca, sukces)

### **III: Atrybuty kontekstu społecznego i gospodarczego:**

16. Emp.var.rate: wskaźnik zmienności zatrudnienia - wskaźnik kwartalny
17. Cons.price.idx: wskaźnik cen konsumpcyjnych - wskaźnik miesięczny
18. Cons.conf.idx: wskaźnik zaufania konsumentów - wskaźnik miesięczny
19. Euribor3m: stawka 3-miesięczna euribor - wskaźnik dzienny
20. Liczba zatrudnionych (nr employed): liczba pracowników - wskaźnik kwartalny

### **Zmienna wyjściowa (podany cel):**

21. y - czy klient subskrybował lokatę? (dwjkowo: tak, nie)

### 1.3 Wstępna eksploracja danych

Badane dane zawierają 4119 wierszy oraz 21 kolumn o następujących nazwach:

```
## [1] "age"          "job"          "marital"      "education"
## [5] "default"      "housing"      "loan"         "contact"
## [9] "month"        "day_of_week"  "duration"     "campaign"
## [13] "pdays"       "previous"     "poutcome"     "emp.var.rate"
## [17] "cons.price.idx" "cons.conf.idx" "euribor3m"    "nr.employed"
## [21] "y"
```

Struktura danych:

```
str(df_bank)

## 'data.frame': 4119 obs. of 21 variables:
## $ age : int 30 39 25 38 47 32 32 41 31 35 ...
## $ job : chr "blue-collar" "services" "services" "services" ...
## $ marital : chr "married" "single" "married" "married" ...
## $ education : chr "basic.9y" "high.school" "high.school" "basic.9y" ...
## $ default : chr "no" "no" "no" "no" ...
## $ housing : chr "yes" "no" "yes" "unknown" ...
## $ loan : chr "no" "no" "no" "unknown" ...
## $ contact : chr "cellular" "telephone" "telephone" "telephone" ...
## $ month : chr "may" "may" "jun" "jun" ...
## $ day_of_week : chr "fri" "fri" "wed" "fri" ...
## $ duration : int 487 346 227 17 58 128 290 44 68 170 ...
## $ campaign : int 2 4 1 3 1 3 4 2 1 1 ...
## $ pdays : int 999 999 999 999 999 999 999 999 999 999 ...
## $ previous : int 0 0 0 0 0 2 0 0 1 0 ...
## $ poutcome : chr "nonexistent" "nonexistent" "nonexistent" "nonexistent" ...
## $ emp.var.rate : num -1.8 1.1 1.4 1.4 -0.1 -1.1 -1.1 -0.1 -0.1 1.1 ...
## $ cons.price.idx : num 92.9 94 94.5 94.5 93.2 ...
## $ cons.conf.idx : num -46.2 -36.4 -41.8 -41.8 -42 -37.5 -37.5 -42 -42 -36.4 ...
## $ euribor3m : num 1.31 4.86 4.96 4.96 4.19 ...
## $ nr.employed : num 5099 5191 5228 5228 5196 ...
## $ y : chr "no" "no" "no" "no" ...
```

Czy w danych znajdują się wartości typu NaN lub Na ?

```
## [1] FALSE
```

Jednakże wiemy, że w danych występują wartości brakujące i są one opisane "unknown". W danych znajduje się 30 rekordów z wartościami "unknown" rozmieszczonych w 1029 różnych wierszach. To stanowi 24.98% wszystkich wierszy w naszej bazie danych, więc możemy pozwolić na usunięcie tych wszystkich informacji. W tabeli 1 znajdują się informacje na temat liczby nieznanych wartości w każdej z kolumn z osobna.

```
## Error: nie znaleziono obiektu 'Number_of_unknown'
## Error in kable(table_unknown, "latex", booktabs = F, caption = "Liczba nieznanych
wartości w poszczególnych kolumnach."): nie znaleziono obiektu 'table_unknown'
```

Nazwy kolumn