Model klasyfikacyjny: czy klient dokona zgłoszenia?

Maciej Podwojewski

Obiekt claims

- Dane zostały wczytane do programu i zapisane jako obiekt o nazwie “claims”.

```
claims <- read_excel("dane.xls", sheet = "claims")

class(claims)
dim(claims)
str(claims)
```

- Obiekt stanowi ramkę danych zawierającą 10296 wierszy (obserwacji) oraz 34 kolumny (zmienne).
- Zmienne miały typy “character” oraz “integer”. Dane wymagały czyszczenia.

Zmienna celu

- #19 CLM_FLAG - dychotomiczna zmienna celu. Zmienna określająca czy dokonano zgłoszenia w bieżącym okresie ubezpieczeniowym. Zapisana jako „character": "Yes", "No". Zamieniona na zmienną „factor”.
- Pierwotnie podjęto decyzję o zmianie typu zmiennej na “logical”, ale modele lasów losowych nie działały na zmiennej logicznej.

“Character” na “date”

- Zmienne #3, #8, #18 i # 20 zapisane jako „character” zamienione na „date”.
- Wartości zawierają francuskie nazwy miesiacy. Stworzono funkcję **fr_months** zmieniającą te nazwy na odpowiadające im numery (w formacie -XX-).

```
fr_months <- function(data) {
  data <- sub(pattern = "janvier", replace = "-01-", x = data)
  data <- sub(pattern = "février", replace = "-02-", x = data)
  data <- sub(pattern = "mars", replace = "-03-", x = data)
  data <- sub(pattern = "avril", replace = "-04-", x = data)
  data <- sub(pattern = "mai", replace = "-05-", x = data)
  data <- sub(pattern = "juin", replace = "-06-", x = data)
  data <- sub(pattern = "juillet", replace = "-07-", x = data)
  data <- sub(pattern = "août", replace = "-08-", x = data)
  data <- sub(pattern = "septembre", replace = "-09-", x = data)
  data <- sub(pattern = "octobre", replace = "-10-", x = data)
  data <- sub(pattern = "novembre", replace = "-11-", x = data)
  data <- sub(pattern = "décembre", replace = "-12-", x = data)
  data <- as.Date(format(as.Date(data,"%d-%m-%y"), "19%y-%m-%d"), "%Y-%m-%d")
}
```

- 2-cyfrowy format lat powodował problemy (56 zamienialo się na 2056, a nie 1956), dlatego zastosowano funkcje **format**.

“Character” na “integer”

- Zmienne #7, #13, #25 i # 31 zapisane jako „character” zamienione na „integer”.
- Wartości zawierają znak „$”. Zastosowano funkcję **sub** w celu jego usunięcia.

```
claims$BLUEBOOK <- as.numeric(sub(pattern = "\\$", replace =
"", x = claims$BLUEBOOK))
```

- Brakujące wartości zostały zamienione na 0.

```
claims$INCOME[is.na(claims$INCOME)] <- 0
```

- Zestandaryzowano zmienne liczbowe.

```
claims$BLUEBOOK <- scale(claims$BLUEBOOK, center = TRUE,
scale = TRUE)
```

“Character” na “factor”

- Zmienne #5, #11, #26 i # 29 zapisane jako „character” zamienione na „factor”.

```
claims$CAR_USE <- as.factor(claims$CAR_USE)
```

- Zmienne #21, #30 i #33 zapisane jako „character” zamienione na „factor” o uporządkowanych poziomach.

```
claims$AGE <- factor(claims$AGE,ordered = TRUE, levels =
c("16-24", "25-40", "41-60", ">60"))
```

- Brakujące wartości zostały zamienione na “Unknown”.

```
claims$JOBCLASS[is.na(claims$JOBCLASS)] <- "Unknown"
```

Błędne dane

- Zmienna #32 SAMEHOME określająca liczba lat zamieszkania kierowcy w obecnym domu zawierała wartość ujemna.
- Podjęto decyzję o usunięciu tej obserwacji.

Wykluczone zmienne

- #1 ID - numer identyfikacyjny. **Zmienna w pełni zrandomizowana.**
- #6 POLICYNO - numer polisy . **Zmienna w pełni zrandomizowana.**
- #17 CLM_AMT – sumaryczna wartość zgłoszeń w bieżącym okresie ubezpieczeniowym. **Zmienna bezpośrednio związana ze zmienną celu.**
- #18 CLM_DATE – data ostatniego zgłoszenia w bieżącym okresie ubezpieczeniowym. **Zmienna bezpośrednio związana ze zmienną celu.**
- #22 AGE*GENDER – **zmienna zawiera same wartości “0„.**
- #34 YEARQT – kwartał podpisania umowy. **Zmienna tożsama z PLCYDATE**.

Efekty czyszczenia

claims

- Obiekt stanowi ramkę danych zawierającą 10296 wierszy (obserwacji) oraz 34 kolumny (zmienne).
- Zmienne miały typy “character” oraz “integer”.
- Dane wymagały czyszczenia.

claims2

- Obiekt stanowi ramkę danych zawierającą 10295 wierszy (obserwacji) oraz 28 kolumny (zmienne).
- Zmienne miały typy “factor”, “date” oraz “integer”.
- Dane gotowe do analizy.

Regresja logistyczna - przygotowanie

- Cel: budowa modelu klasyfikacyjnego zdolny określić czy dany klient dokona zgłoszenia zdarzenia.
- Dane zostały przetasowane.

```
set.seed(42)
rows <- sample(nrow(claims2))
claims3 <- claims2[rows,]
```

- Dokonano podziału danych na zbiór treningowy i testowy w proporcji 80:20.

```
split <- round(nrow(claims3) * 0.80)
train <- claims3[1:split, ]
test <- claims3[(split+1):nrow(claims3), ]
```

Regresja logistyczna – model

- Zbudowano model regresji logistycznej.

```
model <- glm(CLM_FLAG ~ .,  family = "binomial", train)
```

- Dokonano predykcji.

```
p <- predict(model, test, na.action = na.pass, type =
"response")
T_or_F <- ifelse(p>0.5,"Yes","No")
class(T_or_F)class(test$CLM_FLAG)
p_class <- factor(T_or_F)
confusionMatrix(table(p_class, test[["CLM_FLAG"]])
```

Regresja logistyczna – wyniki model

- Macierz pomyłek i statystyka

```
p_class    No  Yes
     No  1413  308
     Yes  106  232
Accuracy : 0.7989
95% CI : (0.781, 0.8161)
P-Value [Acc > NIR] : 5.131e-11
Sensitivity : 0.9302
Specificity : 0.4296
```

- Problem: Model zawiera aż 27 zmiennych objasniajacych! Podjęto decyzję o odrzuceniu z modelu zmiennych nieistotnych (p > 0,05).

Regresja logistyczna – model 2

- Ponownie zbudowano model z 18 zmiennymi.

```
model <- glm(CLM_FLAG ~ KIDSDRIV + TRAVTIME + CAR_USE + BLUEBOOK +
NPOLICY + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS +
BIRTH + AGE + INCOME + MARRIED + PARENT1 + MAX_EDUC + HOME_VAL +
DENSITY, family = "binomial", train)
```

- Dokonano predykcji.

```
p <- predict(model, test, na.action = na.pass, type =
"response")
T_or_F <- ifelse(p>0.5,"Yes","No")
class(T_or_F)class(test$CLM_FLAG)
p_class <- factor(T_or_F)
confusionMatrix(table(p_class, test[["CLM_FLAG"]]))
```

Regresja logistyczna – wyniki model 2

- Macierz pomyłek i statystyka

```
p_class    No   Yes
     No  1410   301
     Yes  109   239
Accuracy : 0.8009
95% CI : (0.783, 0.8179)
P-Value [Acc > NIR] : 1.228e-11
Sensitivity : 0.9282
Specificity : 0.4426
```

- Zdecydowano o usunięciu zmiennych skorelowanych (cor > 0,3) oraz zmiennej BIRTH, na podstawie której zbudowano dyskretną zmienną AGE.

.

Regresja logistyczna – model 3

- Ponownie zbudowano model z 15 zmiennymi.

```
model <- glm(CLM_FLAG ~ KIDSDRIV + TRAVTIME + CAR_USE + BLUEBOOK +
NPOLICY + CAR_TYPE + CLM_FREQ + REVOKED + MVR_PTS + AGE +  MARRIED
+ PARENT1 + MAX_EDUC + HOME_VAL + DENSITY, family = "binomial",
train)
```

- Dokonano predykcji.

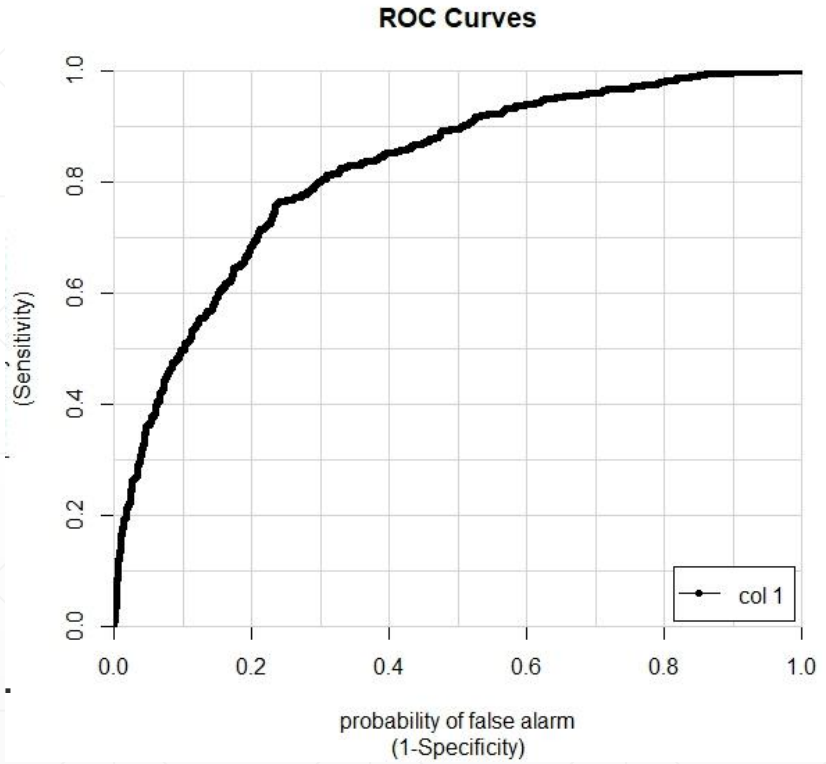
```
p <- predict(model, test, na.action = na.pass, type =
"response")
T_or_F <- ifelse(p>0.5,"Yes","No")
class(T_or_F)class(test$CLM_FLAG)
p_class <- factor(T_or_F)
confusionMatrix(table(p_class, test[["CLM_FLAG"]]))
```

Regresja logistyczna – wyniki model 3

- Macierz pomyłek i statystyka

```
p_class    No   Yes
     No  1411   310
     Yes  108   230
Accuracy : 0.797
95% CI : (0.779, 0.8142)
P-Value [Acc > NIR] : 2.044e-10
Sensitivity : 0.9289
Specificity : 0.4259
```

- Model jest bardzo skuteczny w wykrywaniu kierowców, którzy nie będą dokonywać zgłoszenia.
  W celu dostosowania modelu do potrzeb zadania podjęto decyzję o obniżeniu wartości cut-off-u z .5 do .3.



Regresja logistyczna – wyniki model 3

- Macierz pomyłek i statystyka

```
p_class    No   Yes
     No  1180   150
     Yes  339   390
Accuracy : 0.7625
95% CI : (0.7435, 0.7807)
P-Value [Acc > NIR] : 0.005343
Sensitivity : 0.7768
Specificity : 0.7222
```

- Zmieniając wartość cut-off-u osiągnięto wartość maksymalizującą wrażliwość i swoistość modelu.

Interpretacja modelu

- Analizę współczynników modelu logistycznego dokonuje się przedstawiając je w terminach szans (iloraz prawdopodobieństwa sukcesu do prawdopodobieństwa porażki).

```
CLM_FLAG^ = -0.89 + 0.24×KIDSDRIV - 0.85×CAR_USEPrivate...
```

```
exp(coef(model))
```

- CAR_USEPrivate 0,43 - Szansa, że kierowca samochodu prywatnego zgłosi szkodę jest o 57% mniejsza niż w przypadku kierowcy samochodu służbowego.
- KIDSDRIVE 1,27 - Wraz ze wzrostem liczby liczby dzieci szansa, że kierowca zgłosi szkodę wzrasta (przez standaryzację intuicyjnie określić można jedynie kierunek zmian).

Dziękuję za uwagę!