Małgorzata Pieronek

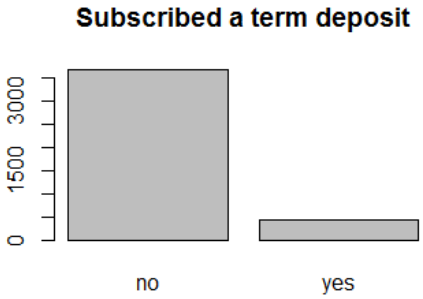
Zbudowanie modelu prognozującego szansę, że klient w wyniku prowadzonej kampanii założy lokatę terminową.

1. Cel i zakres analizy

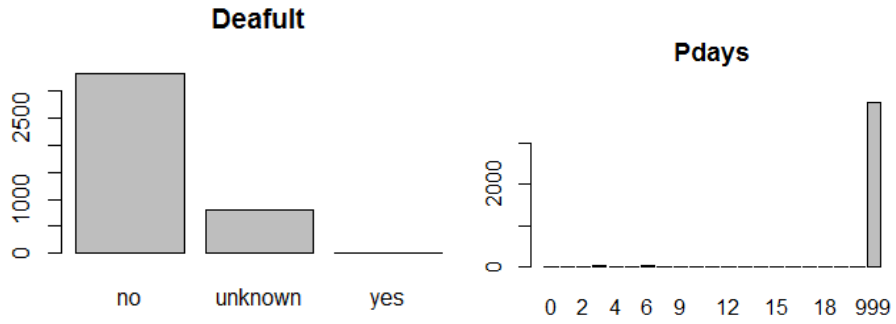
Celem niniejszej pracy jest zbudowanie modelu prognozującego szansę, że klient w wyniku prowadzonej kampanii założy lokatę terminową. Do przeprowadzenia analizy użyłam metody CRISP-DM, która składa się z 6 etapów. Pierwszym z nich jest zrozumienie problemu biznesowego, który ma odzwierciedlenie w danych. Drugi krok to eksploracja danych polegająca na wstępnej ocenie jakości danych i występujących pomiędzy nimi zależności. Kolejny to modyfikacja danych, służąca przygotowaniu ich do modelu, w tym kroku wykonuje się m.in. podział zmiennych na zbiór treningowy i testowy. Po wykonaniu tych czynności można przystąpić do modelowania za pomocą wybranej metody. Ostatnią czynnością przed wdrożeniem jest ewaluacja, czyli ocena jakości zbudowanego modelu.

2. Opis zbioru danych

Dane dotyczą klientów pewnego banku, kampanii marketingowych skierowanych do tych klientów oraz wskaźników społecznych i ekonomicznych. Zbiór zawiera 4119 obserwacji i 21 zmiennych. Zmienna objaśniana (celu) przyjmuje dwie wartości: tak (451 klientów) - jeśli klient zdecydował się na założenie lokaty terminowej lub nie (3 668 klientów) - jeśli nie podjął takiej decyzji.



Na podstawie otrzymanych rozkładów cech podjęto decyzję o nie braniu pod uwagę przy budowaniu modelu zmiennej *default* ponieważ występuje ogromna dysproporcja w liczbie odpowiedzi: „no” (3315 obserwacji) w stosunku do odpowiedzi „yes” która występuje w zbiorze danych jeden raz. Z kolei decyzja o niewłączeniu zmiennej *duration* do modelu została podjęta na podstawie wiedzy eksperckiej. Również zmienna *pdays* nie będzie uwzględniona w modelu, ponieważ 96% klientów nie było wcześniej kontaktowanych (dla takich osób zmienna przyjmowała wartość 999).



Ze zbioru usunięte zostały obserwacje które przynajmniej dla jednej zmiennej przyjmowały wartość „unknown”. Finalnie w dalszej analizie korzystano z 3810 obserwacji. Dla zmiennej *poutcome* ok.85% odpowiedzi to „nonexistent” ale nie jest to brak danych, tylko informacja o tym, że danego klienta poprzednia akcja marketingowa nie dotyczyła.

3. Kroswalidacja i selekcja zmiennych

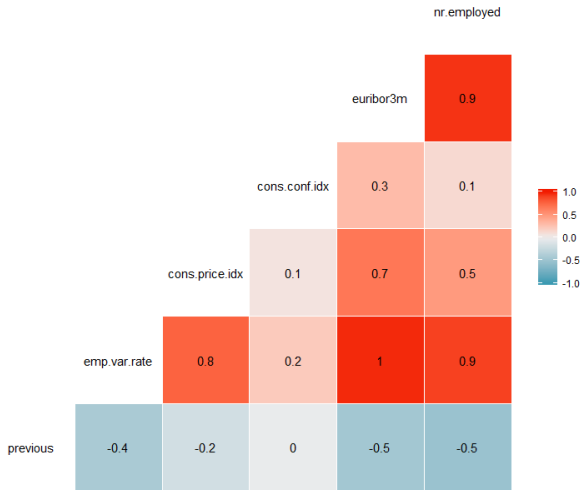
W budowaniu modeli machine learningowych konieczne jest podzielenie zbioru danych na dane uczące i testowe (walidacja prosta). Jest to konieczne w celu weryfikacji jakości modelu (sprawdzenia mocy predykcyjnej na nowym zestawie danych). Zbiór został podzielony na zbiór uczący (60%) i zbiór testowy (40%).

Przed przystąpieniem do budowy modelu należy dokonać wstępnej selekcji zmiennych. Zależność pomiędzy zmienną objaśnianą a zmiennymi kategorialnymi zweryfikowałam za pomocą testu niezależności chi-kwadrat. Test niezależności chi-kwadrat wykonuje się w celu zbadania związku pomiędzy dwoma zmiennymi nominalnymi. Hipoteza zerowa mówi że zmienne są niezależne, hipoteza alternatywna że nie są niezależne. Na poziomie istotności równym 0,05 (po odrzuceniu hipotezy zerowej) przyjąć można że istnieje zależność między skłonnością do założenia lokaty a zmiennymi: *contact, job, month, poutcome*.

Dla danych numerycznych policzony został wskaźnik IV. Ze względu na wartość IV odrzucona zostanie zmienna *age* i *campaign*, ponieważ wartość IV wynosi oscyluje wokół 0,1.

Variable	IV
nr.employed	1.1211967
euribor3m	1.0512336
emp.var.rate	0.9606322
cons.conf.idx	0.6442381
cons.price.idx	0.3648050
previous	0.3331563
age	0.1098200
campaign	0.1012806

Aby zbadać czy nie występuje korelacja między zmiennymi posłużyłam się współczynnikiem korelacji Pearsona. Na podstawie otrzymanych wyników, została podjęta decyzja o dalszym uwzględnianiu jedynie zmiennych *previous, cons.conf.idx*.



Ponieważ zmienne *job* i *month* posiadają liczne kategorie wykonane zostało grupowanie, aby nie było kategorii o bardzo małej liczności a także żeby zmniejszyć liczbę występujących kategorii co ułatwi interpretację modelu. W tą samą kategorię połączone zostały te poziomy które posiadają podobny odsetek pozytywnych odpowiedzi w danej kategorii. Poniższe tabele prezentują sposób przekodowania zmiennych:

Zmienna *job*	
admin, technician	grupa1
blue collar, enterpreneur	grupa2
housemaid, management, self-employed, services	grupa3
retired, student, unemployed	grupa4

Zmienna *month*	
aug, jun, nov	grupa1
apr, dec, mar, oct, sep	grupa2
jul, may	grupa3

Dla zmiennych kategorialnych utworzone zostaną sztuczne zmienne (dummy variables), jedna z tych zmiennych dla każdej kategorii pozostanie zmienną referencyjną.

4. Budowa modelu

Duża dysproporcja w liczności kategorii zmiennej objaśnianej wymagała zastosowania oversamplingu na zbiorze uczącym. Następnie na tak przygotowanych danych zbudowany został model regresji logistycznej, z wykorzystaniem krokowej metody selekcji zmiennych (wstecz i w przód). Na podstawie otrzymanych wyników wybrany został model o najmniejszej wartości AIC, przy jednoczesnym uwzględnieniu jak najodpowiedniejszej wielkość modelu. Ostatecznie w modelu znalazły się takie zmienne jak: *previous, poutcome.failure, contact.telephone, job.grupa4_job, poutcome.nonexistent, job.grupa3_job, job.grupa2_job, cons.conf.idx.*

Po zbudowaniu modelu zbadałam wartość VIF, aby wykryć współliniowość między zmiennymi. Wartości bliskie 10 oznaczają, że dany predyktor jest silnie powiązany z innym i należy się zastanowić nad jego usunięciem lub agregacją z innym predyktorem. W modelu żadna ze zmiennych nie przekroczyła wartości 10 tego wskaźnika.

```
      previous     poutcome.failure     contact.telephone       job.grupa4_job poutcome.nonexistent
           4.5                  4.9                   1.1                  1.1                  8.9
job.grupa3_job       job.grupa2_job        cons.conf.idx
           1.2                  1.2                   1.1
```

Współczynniki otrzymanego modelu zawarte są w poniższym zestawieniu:

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          2.41172    0.39292    6.14  8.4e-10 ***
previous             0.64741    0.13890    4.66  3.1e-06 ***
poutcome.failure    -2.21726    0.22406   -9.90  < 2e-16 ***
contact.telephone   -0.66959    0.08057   -8.31  < 2e-16 ***
job.grupa4_job       0.67141    0.11464    5.86  4.7e-09 ***
poutcome.nonexistent -1.76609    0.27893   -6.33  2.4e-10 ***
job.grupa3_job      -0.46284    0.08952   -5.17  2.3e-07 ***
job.grupa2_job      -0.40418    0.09038   -4.47  7.8e-06 ***
cons.conf.idx        0.01485    0.00694    2.14    0.032 *
```

Interpretacja otrzymanego modelu:

- Wraz ze wzrostem liczby kontaktów z klientem przed kampanią rośnie prawdopodobieństwo zakupienia lokaty terminowej (zmienna *previous*).
- Jeśli poprzednia kampania reklamowa zakończyła się niepowodzeniem w przypadku danego klienta lub w ogóle nie objęła ona danego klienta to jest mniejsze prawdopodobieństwo zakupu lokaty przez tego klienta aniżeli w przypadku gdy poprzednia kampania reklamowa zakończyła się sukcesem (zmienna *poutcome*).
- Osoby wykonujące zawód który został zaklasyfikowany do grupy 2 i 3 (blue collar, enterpreneur, housemaid, management, self-employed, services) wykazują mniejszą skłonność do założenia lokaty aniżeli osoby z grupy 1 (admin, technician). Z kolei dla osób z grupy 4 (retired, student, unemployed) skłonność do założenia lokaty jest większa aniżeli u osób z grupy 1 (zmienna *job*).
- Klienci z którymi kontaktowano się za pomocą telefonu stacjonarnego wykazują mniejszą skłonność do założenia lokaty aniżeli klienci kontaktowani na telefon komórkowy (zmienna *contact*).
- Wraz ze wzrostem wartości zmiennej *consumer confidence index* rośnie skłonność do założenia lokaty (zmienna *cons.conf.idx*).

5. Ocena jakości modelu

Aby zweryfikować jakość otrzymanego modelu wykonałam regresję logistyczną na zbiorze testowym. W efekcie powstał model którego wszystkie zmienne pozostały nadal istotne. Niepotrzebna była więc modyfikacja modelu.

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)            1.212      0.702    1.73    0.084 .
previous               0.503      0.203    2.47    0.013 *
poutcome.failure      -2.163      0.312   -6.93  4.1e-12 ***
contact.telephone     -0.773      0.185   -4.18  2.9e-05 ***
job.grupa4_job         0.547      0.222    2.47    0.014 *
poutcome.nonexistent  -1.948      0.391   -4.98  6.5e-07 ***
job.grupa3_job        -0.395      0.193   -2.04    0.041 *
job.grupa2_job        -0.472      0.206   -2.28    0.022 *
cons.conf.idx          0.032      0.015    2.13    0.033 *
```

Również wartość VIF jest na akceptowalnym poziomie dla wszystkich zmiennych:

```
      previous     poutcome.failure    contact.telephone       job.grupa4_job poutcome.nonexistent
           3.6                  2.4                  1.1                  1.1                  5.3
job.grupa3_job       job.grupa2_job       cons.conf.idx
           1.1                  1.2                  1.1
```

Na podstawie otrzymanych modeli dla każdej obserwacji można wyliczyć prawdopodobieństwo zajścia badanego zjawiska, w tym przypadku założenia lokaty terminowej. Obliczenia te zostały wykonane w programie R, w efekcie dla każdej obserwacji obliczone jest prawdopodobieństwo.

Do oceny jakości dopasowania modelu do danych stosuje się macierz trafności, a na jej podstawie oblicza odpowiednie miary:

		model	
		true	false
real	yes	48	201
	no	23	2028

Precision (precyzja): 67,6%. Miara ta mówi w jakim procencie pozytywne przewidywania w rzeczywistości takie były. Czyli np. ile osób które wytypowaliśmy jako skłonne założyć lokatę rzeczywiście ją założyło.

Recall (czułość): 19,28%. Miara opisująca jaki odsetek pozytywnych przypadków model jest w stanie wykryć.

Specificity (swoistość): 98,88%. Miara wskazująca jaki odsetek negatywnych przypadków model jest w stanie wykryć.

Model z ogromną dokładnością wykrywa negatywne wartości. Blisko 99% przypadków które w rzeczywistości przyjmowały wartość „no" rzeczywiście przez model zostały tak zaklasyfikowane. Jednak odsetek poprawnie rozpoznanych jako „yes" wśród wszystkich w rzeczywistości pozytywnych odpowiedzi wynosi zaledwie 19,28%. Dlatego aby otrzymać wynik uwzględniający zdolność modelu do wykrywania danego zjawiska jak również wykrywania braku tego zjawiska oblicza się miarę accuracy (dokładność) W przypadku tego modelu wynosi 90,26% co świadczy o dobrej mocy predykcyjnej tego modelu.

Inną miarą pozwalającą ocenić jakość modelu jest krzywa ROC, która pokazuje zależność między FPR a TPR. Im większe pole pod krzywą tym model jest lepiej dopasowany. Miernikiem powiązanym z tym wykresem jest AUC (Area Under Curve) czyli pole pod krzywą. AUC może przyjmować wartości z zakresu <0,1>. Wartości (0; 0,5) świadczą o tym że model jest gorszy od losowego. AUC=0,5 klasyfikator losowy, AUC=1 klasyfikator idealny. W przypadku tego modelu AUC wynosi 0,71.