Sprawozdanie z projektu Data Mining

Temat nr 1: Klasyfikacja

Spis treści

Opis danych i zagadnienia klasyfikacyjnego

W repozytorium *UCI Machine Learning Repository* znaleźliśmy zbiór Bank Marketing Data Set, który zawiera nieco ponad 41 tysięcy rekordów, oraz 20 atrybutów. Dane są informacjami o klientach jednej z portugalskich instytucji bankowych i pochodzą z kampanii marketingu bezpośredniego, polegającej na telefonicznym zachęcaniu klientów do skorzystania z usługi lokaty bankowej.

Zagadnieniem klasyfikacyjnym, z jakim przyjdzie nam się zmierzyć w obliczu tego zbioru danych, jest przewidzenie czy klient zdecyduje się wykupić lokatę terminową czy nie. W realizacji tego zadania skorzystamy z drzew klasyfikacyjnych.

Opis atrybutów zbioru

W zbiorze znajduje się 20 zmiennych, po 10 kategorycznych i 10 liczbowych:

Podstawowe bankowe informacje:

1. *age* – (liczbowa) wiek klienta
2. *job* – (kategoryczna) zawód wykonywany przez klienta, architekci bazy danych wyodrębnili 12 typów wykonywanych profesji, z uwzględnieniem własnej działalności oraz braku zatrudnienia ['admin.','blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown']
3. *marital* – (kategoryczna) stan cywilny klienta ['divorced','married','single','unknown'], przy czym do statusu 'divorced' zaliczają się również osoby owdowiałe
4. *education* – (kategoryczna) informacja o wykształceniu klienta ['basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown']
5. *default* – (kategoryczna) informacja o tym czy klient ma niespłacony kredyt w banku ['no', 'yes', 'unknown]
6. *housing* – (kategoryczna) informacja o tym czy klient ma kredyt hipoteczny w banku ['no', 'yes', 'unknown]
7. *loan* – (kategoryczna) informacja o tym czy klient ma pożyczkę osobistą w banku ['no', 'yes', 'unknown]

Informacje związane z bieżącą kampanią marketingową:

8. *contact* – (kategoryczna) rodzaj komunikacji w kampanii ['cellular', 'telephone']
9. *month* – (kategoryczna) miesiąc w którym był ostatni kontakt w związku z kampanią ['jan', 'feb', 'mar', ..., 'nov', 'dec']
10. *day_of_week* – (kategoryczna) dzień tygodnia, w którym był ostatni kontakt w związku z kampanią ['mon', 'tue', 'wed', 'thu', 'fri']
11. *duration* – (liczbowa) czas trwania ostatniego kontaktu w związku z kampanią. Ta informacja może być mocno związana z naszą wartością przewidywaną (np. gdy jest równa bądź bliska 0, to wiadomo, że klient nie zdecydował się na usługę) jest czymś w rodzaju wycieku danych, stąd nie może być wykorzystana do modelu, a jedynie do walidacji. Dodatkowo nie będzie znana przed wykonaniem telefonu do klienta, stąd nie powinna być używana do procesu uczenia.

Inne informacje o kampanii:

12. *campaign* – (liczbowa) ilość kontaktów z klientem wykonanych w ramach bieżącej kampanii
13. *pdays* – (liczbowa) ile dni minęło od ostatniego kontaktu, przy czym 999 znaczy, że nie było wcześniej kontaktu
14. *previous* – (liczbowa) ilość kontaktów z klientem wykonanych przed bieżącą kampanią

15. *poutcome* – (kategoryczna) wynik poprzedniej kampanii marketingowej ['failure', 'nonexistent', 'success']

Informacje dotyczące społecznych i ekonomicznych wskaźników, w momencie, gdy kontaktowano się z klientem:

16. *emp.var.rate* – (liczbowa) wskaźnik zmienności zatrudnienia w ujęciu kwartalnym
17. *cons.price.idx* – (liczbowa) indeks kosztów konsumenckich w ujęciu miesięcznym
18. *cons.conf.idx* – (liczbowa) wskaźnik ufności konsumenckiej w ujęciu miesięcznym
19. *euribor3m* – (liczbowa) wskaźnik euribor z 3 miesięcy w ujęciu dziennym. Jest to międzybankowa stopa procentowa, podstawowa stopa procentowa stosowana do udzielania pożyczek między bankami na rynku międzybankowym Unii Europejskiej. Jest stosowana jako punkt odniesienia przy ustalaniu stopy procentowej innych pożyczek.
20. *nr.employed* – (liczbowa) liczba pracowników w ujęciu kwartalnym

Zmienna opisywana

Naszą zmienną przewidywaną jest zmienna *y*, która jest zmienną kategoryczną o wartościach [‘yes’, ‘no’]. Informuje ona o tym czy klient zdecydował się wykupić lokatę terminową.

Dane pochodzą z bazy, w której telemarketerzy, po odbyciu rozmowy telefonicznej, wprowadzali informacje o tym czy klient zdecydował się na zakupienie usługi czy nie. Celem naszego projektu jest przewidzenie wyniku rozmowy telefonicznej zanim zostanie ona przeprowadzona.

Wśród 41 188 klientów, do których zadzwoniono, 4640 zdecydowało się na zakup lokaty – stanowi to 11,3% zbioru. W zmiennej opisywanej nie ma pustych wartości, zbiór jest kompletny.

Eksploracja danych

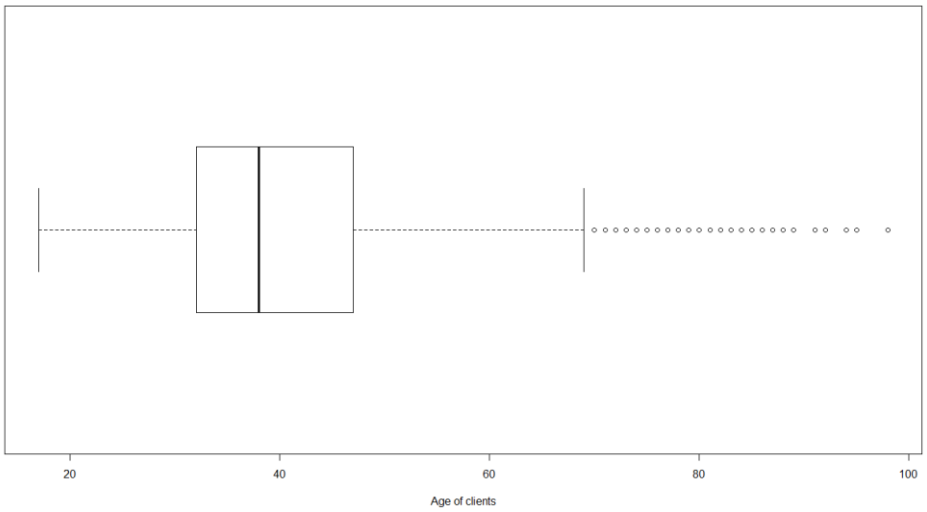
Do środowiska statystycznego R wczytano dane zawarte w pliku *bank-additional-full.csv*, który znajduje się we wspomnianym wcześniej repozytorium. Repozytorium zawiera cztery pliki, zdecydowaliśmy się na skorzystanie z zestawu o największej liczbie rekordów oraz największej liczbie atrybutów (starszy zestaw zawiera 17 zamiast 20 atrybutów).

Dane pochodzą z okresu Maj 2008 – Listopad 2010.

Podstawowe bankowe informacje o kliencie

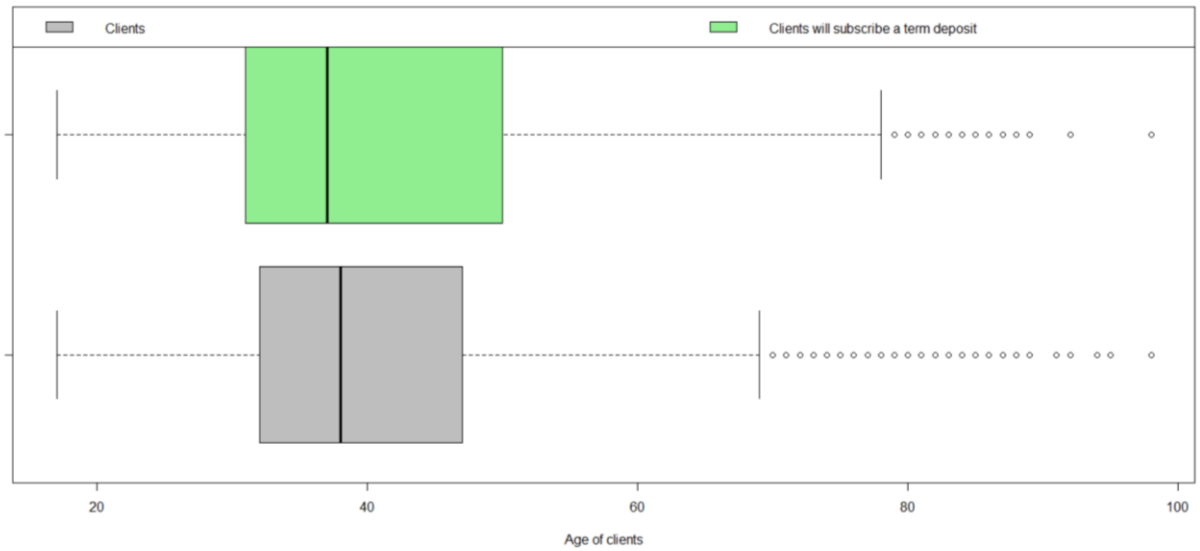
Wiek

Rozstęp międzykwartylowy wieku klientów, do których wykonano połączenie zawierał się w przedziale od 32 do 47. „Wąsy” na wykresie pudełkowym sięgają blisko wieku 70 lat, natomiast jest również około 20 klientów powyżej tego progu.

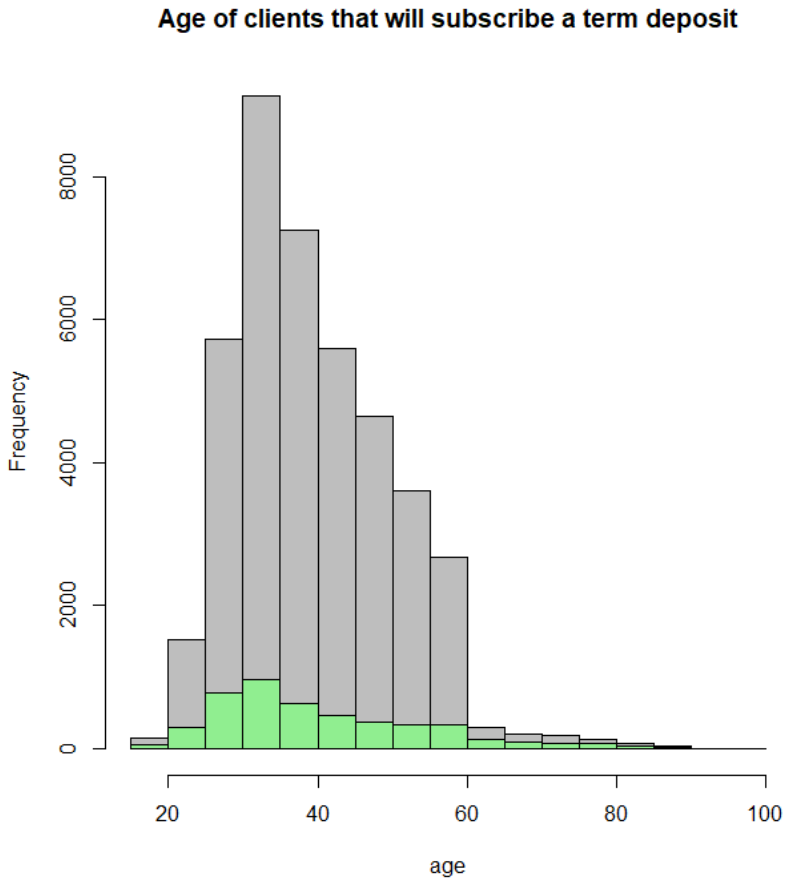


Ich należy potraktować jako obserwacje odstające, między innymi klienta, który doczekał nawet wieku 98 lat.

Porównano wykres pudełkowy wieku klientów, którzy wykupili lokatę, że wszystkimi klientami:



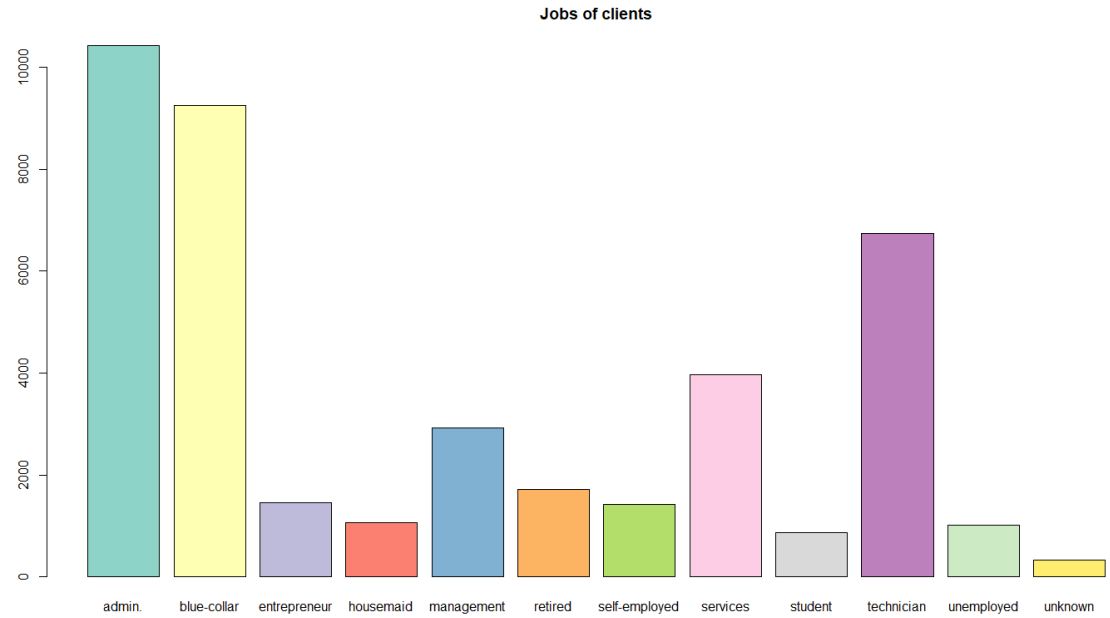
Widać, że rozstęp międzykwartylowy jest szerszy niż dla wszystkich klientów, a także, że wąsy wydłużyły się w kierunku wyższego wieku. Może to świadczyć o tym, że wśród starszych klientów zainteresowanie lokatą jest wyższe.



Widać to również na powyższym histogramie, na którym zestawiono klientów, którzy wykupili lokatę, na tle wszystkich klientów. O ile wciąż widać, że w najbardziej licznym przedziale wiekowym, czyli między 25 a 45 roku życia, jest też najwięcej klientów, którzy się zdecydowali na lokatę, to ich odsetek jest znacznie niższy w porównaniu do klientów 60+.

Typ zatrudnienia

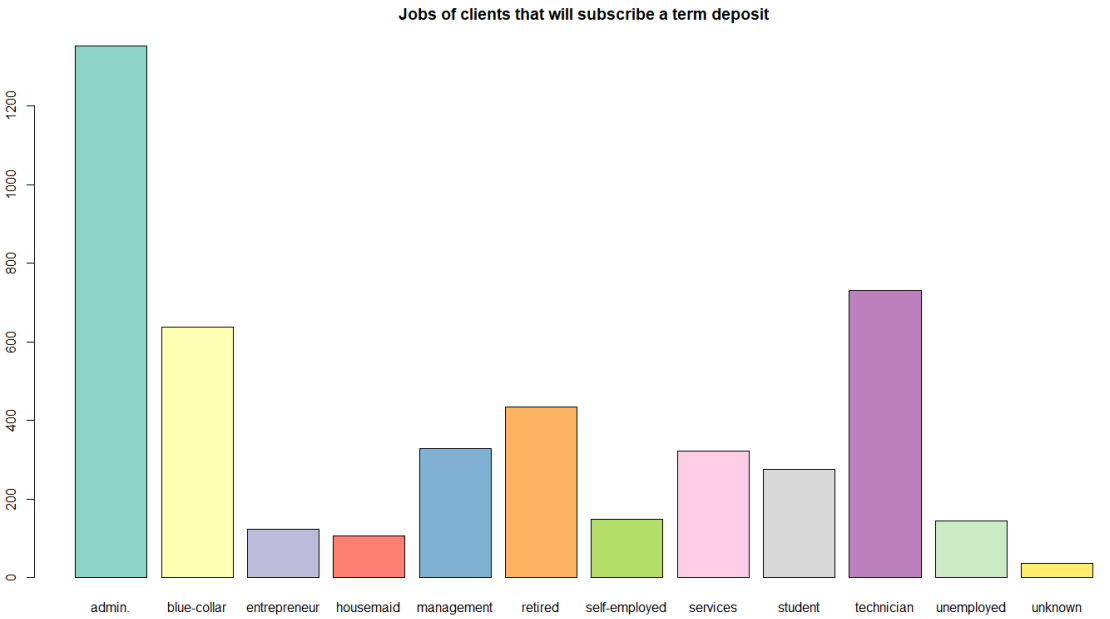
Nie mamy w zbiorze danych informacji o zarobkach klientów, co mogłoby być dobrze skorelowane z tym czy mają wystarczająco oszczędności, żeby zdecydować się na wpłatę pieniędzy na lokatę terminową czy nie. Jednak w zbiorze jest zawarta informacja o tym, jaki zawód wykonują, niesie za sobą pośrednio informacje o tym jakie mogą mieć zarobki i status ekonomiczny – stąd atrybut ten może być istotny w naszym zagadnieniu klasyfikacyjnym:



Jak widać, wśród klientów portugalskiego banku znaczna część osób jest zatrudniona w administracji lub jako tak zwany „blue-collar" (*z ang*. niebieski kołnierzyk) – termin ten określa osoby będące pracownikami w fabrykach, wykonującymi prace fizyczne lub warsztatowe. Termin wziął się od niebieskich ubrań roboczych, które często są noszone przez pracowników zakładów przemysłowych.

W dalszej części struktury zatrudnienia są osoby techniczne oraz pracujące w usługach, dalej mamy kadrę menedżerską, emerytów oraz przedsiębiorców i samozatrudnionych. Część klientów banku stanowią osoby zatrudnione jako pomoc domowa, a na końcu znajdują się osoby bezrobotne oraz studenci.

Sprawdźmy, jak wygląda struktura zatrudnienia wśród klientów, którzy zdecydowali się wykupić lokatę terminową:

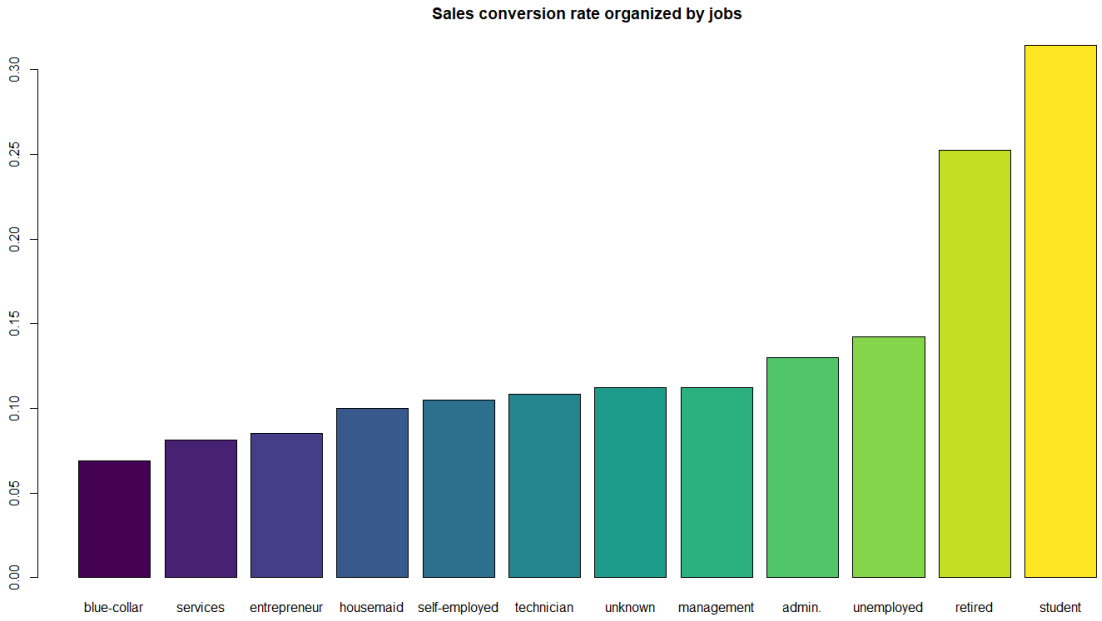


Jak widać dalej dominującą grupą są pracownicy administracji, co może wynikać z tego że jest ich po prostu najwięcej wśród klientów banków. Dalej jednak widać, że osoby techniczne wykupują lokatę

terminową częściej niż „niebieskie kołnierzyki”, a przecież jest ich mniej niż tych drugich. Lokata jest również całkiem popularna wśród emerytów oraz menedżerów i pracowników sektora usługowego. Co ciekawe, studenci częściej zdecydowani na lokatę terminową niż przedsiębiorcy.

W analizie związku między typem zatrudnienia a decyzją o wykupieniu lokaty, istotny wskaźnikiem jest tzw. współczynnik konwersji sprzedaży – czyli jaki procent klientów, do których dotarła kampania marketingowa, zdecydowało się na produkt.

$$współczynnik\ konwersji = \frac{liczba\ klientów\ którzy\ kupili\ usługę\ w\ ramach\ kampanii}{liczba\ wszystkich\ klientów\ w\ kampanii\ marketingowej}$$



Tak jak zauważono wcześniej, mały odsetek „niebieskich kołnierzyków” decyduje się na zakup lokaty bankowej – zaledwie 6,9%. Jest to najmniej spośród wszystkich grup zawodowych. Poniżej 10% konwersji plasują się również klienci z sektora usług ora przedsiębiorcy. Większość grup zawodowych ma współczynnik konwersji w zakresie od 10% do 14,2%, tyle mają osoby zdeklarowane jako bezrobotne. Wśród najliczniejszej grupy zawodowej, pracowników administracyjnych, 13% zdecydowało się na zakup produktu. Kampania marketingowa zdecydowanie najlepiej dotarła do emerytów – 25,2% konwersji oraz do studentów – wśród nich aż 31,4% zdecydowało się na lokatę.

Wśród klientów banku nie ma zbyt wielu studentów, zaledwie 875 na zbiór liczący 41 188 osób. Natomiast wśród tych studentów, aż 275 zdecydowało się wykupić lokatę terminową, co daje tak wysoki współczynnik konwersji.

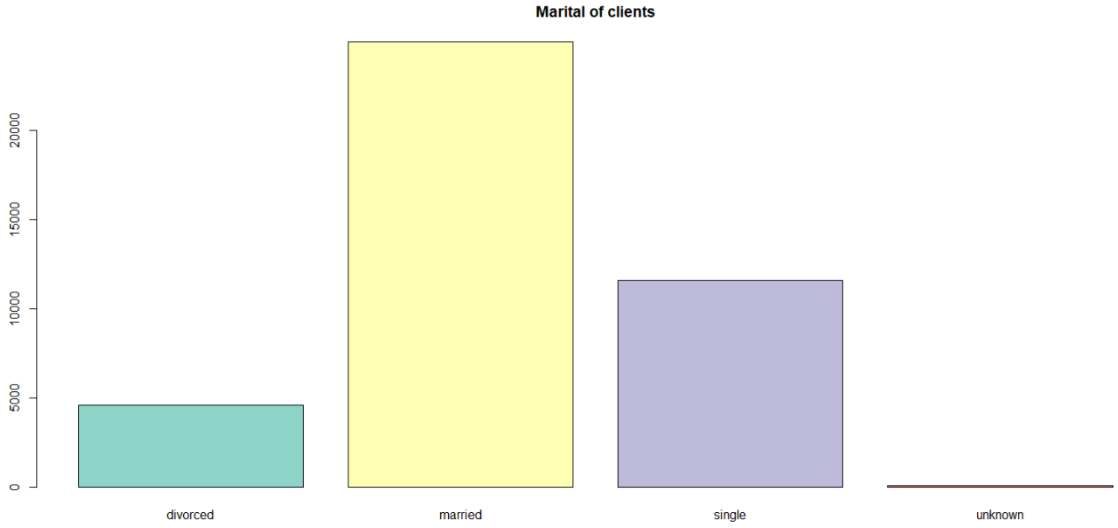
Emerytów jest prawie dwa razy więcej niż studentów, co wciąż jest jednak niskim odsetkiem wśród klientów banku. Jednak wśród nich 434 osoby zdecydowały się na usługę, co wynosi prawie 10% wszystkich którzy ją kupili.

Jak widać niektóre grupy zawodowe klientów są znacznie powiązane z faktem o tym czy wykupią lokatę czy nie, stąd parametr ten będzie istotny przy trenowaniu klasyfikatora.

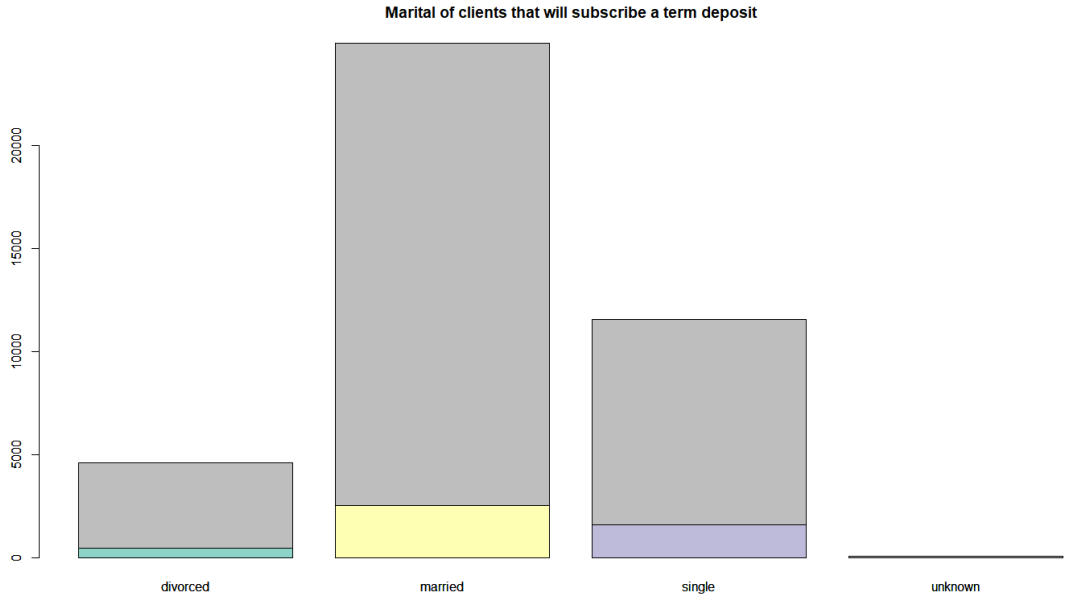
Stan cywilny

W zbiorze znajdują się 3 stany cywilne: w związku małżeńskim, stanu wolnego, oraz rozwiedziony, przy czym rozwiedziony to również osoby owdowiałe. Zaledwie 80 klientów nie podało swojego stanu

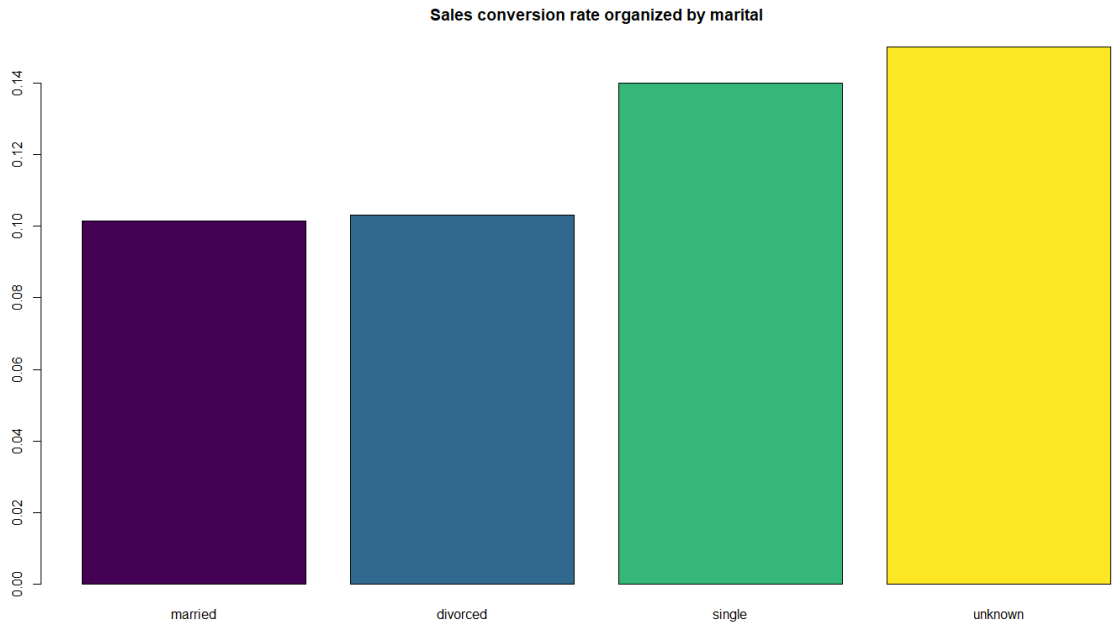
cywilnego.



Jak widać, najwięcej z klientów jest w związku małżeńskim, ponad dwa razy mniej jest stanu wolnego. Mniej niż 1/8 klientów jest po stracie partnera. Sprawdźmy teraz jak wygląda ten układ wśród klientów, którzy kupili lokatę:



Na tle wszystkich klientów widać, że o ile wciąż najwięcej chętnych jest wśród klientów w związku małżeńskim, to ich przewaga nad osobami stanu wolnego nie jest już tak duża jak wśród wszystkich klientów.

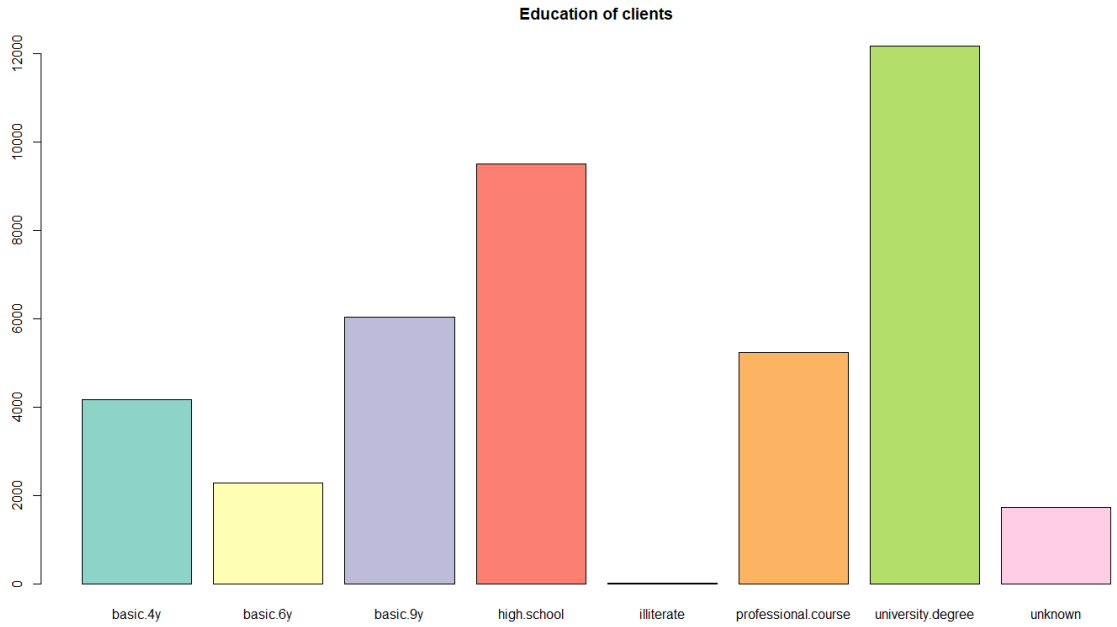


Widać to również wykresie współczynnika konwersji sprzedaży. W przypadku osób stanu wolnego wyniósł prawie o 4 punkty procentowe więcej niż dla pozostałych stanów cywilnych, pomijając przypadek, gdzie nie znamy stanu cywilnego. Z racji, że klientów bez podanego sanu cywilnego jest zaledwie 80, ciężko wnioskować na podstawie tak małej grupy.

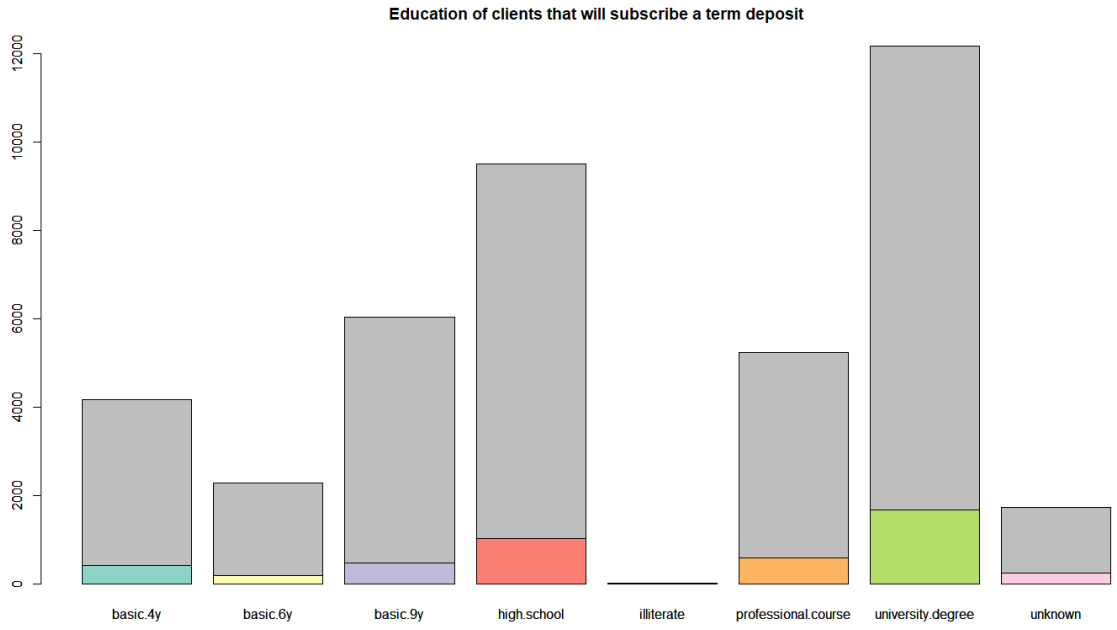
Fakt, że osoby stanu wolnego chętniej decydowały się na zakup lokaty, może świadczyć o tym że bardziej interesuje ich zabezpieczenie finansowe. Być może jest to związane z tym, że nie mając rodziny na utrzymaniu, mają więcej oszczędności.

Wykształcenie

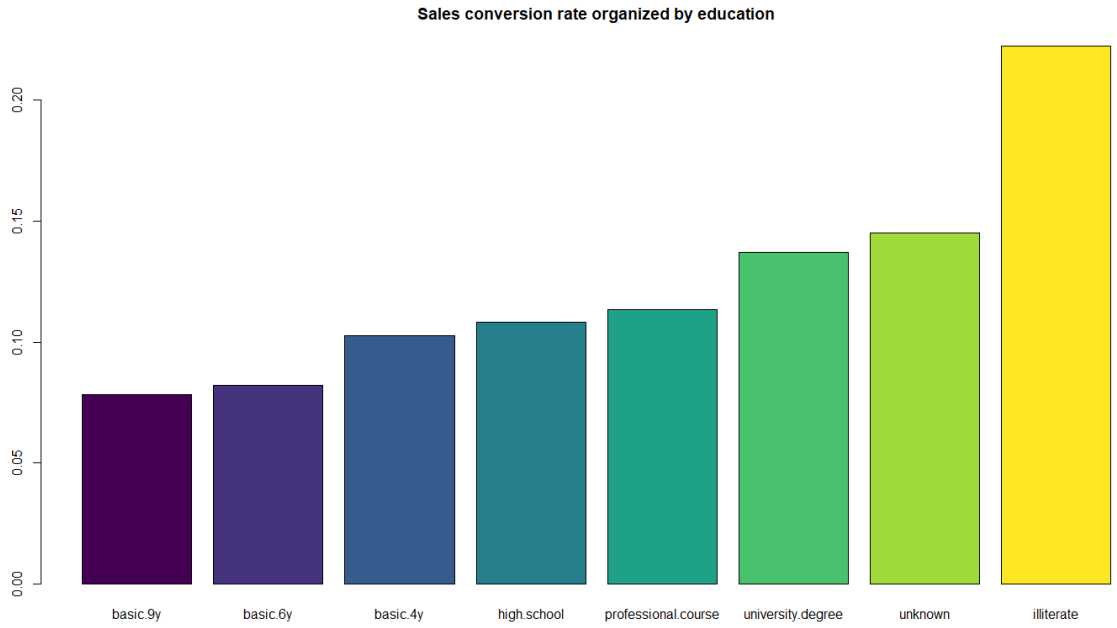
Projektant bazy, do której wprowadzano informacje o klientach, przygotował siedem typów wykształcenia, oraz ósmy typ dla osób, które nie chciały się podzielić tą informacją.



Najwięcej z klientów ma ukończone studia wyższe bądź szkołę średnią. Duża część klientów ma wykształcenie podstawowe które trwało cztery, sześć bądź dziewięć lat – gdyby znieść podział na to, ile trwała ich edukacja, stanowiliby grupę nieco liczniejszą niż osoby o wykształceniu wyższym. Spora część osób jest po kursie zawodowym, zaś 1731 osób nie podało informacji o wykształceniu, co jest dość dużym odsetkiem luk w zbiorze, porównując inne atrybuty. 18 klientów zadeklarowało bycie analfabetami.



Najwięcej klientów to ci posiadający wyższe lub średnie wykształcenie. Część klientów znalazło się" wśród absolwentów kursów zawodowych lub szkół podstawowych.



Największa konwersja sprzedaży występuje wśród analfabetów, natomiast z uwagi na to, że takich klientów jest tylko 18, ciężko wysnuć na tej podstawie jakieś wnioski. Duży współczynnik sprzedaży

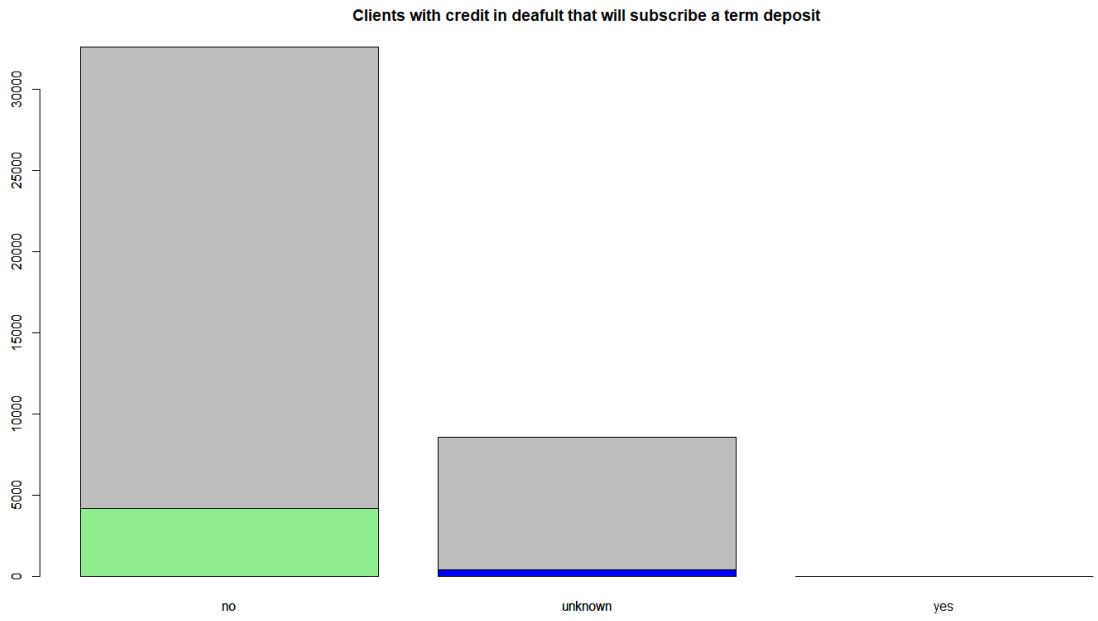
dotyczy również osób, które nie podały informacji o wykształceniu, jednak tu też trudno wnioskować nie mając wiedzy.

Wyraźnym trendem jest jednak to, że osoby z wyższym wykształceniem chętniej wykupują lokatę, niż osoby ze średnim wykształceniem lub po kursie zawodowym, natomiast te z kolei chętniej decydują się na zakup niż klienci z wykształceniem podstawowym. Na tym wykresie dobrze widać, że konwersja sprzedaży rośnie wraz z poziomem edukacji klientów.

Czy klient ma niespłacony kredyt?

Spośród klientów, do których była skierowana kampania marketingowa, zaledwie 3 klientów powiedziało, że ma niespłacony kredyt, żaden z nich nie zdecydował się na lokatę. Aż 20% klientów nie udzieliło informacji o tym czy mają niespłacony kredyt.

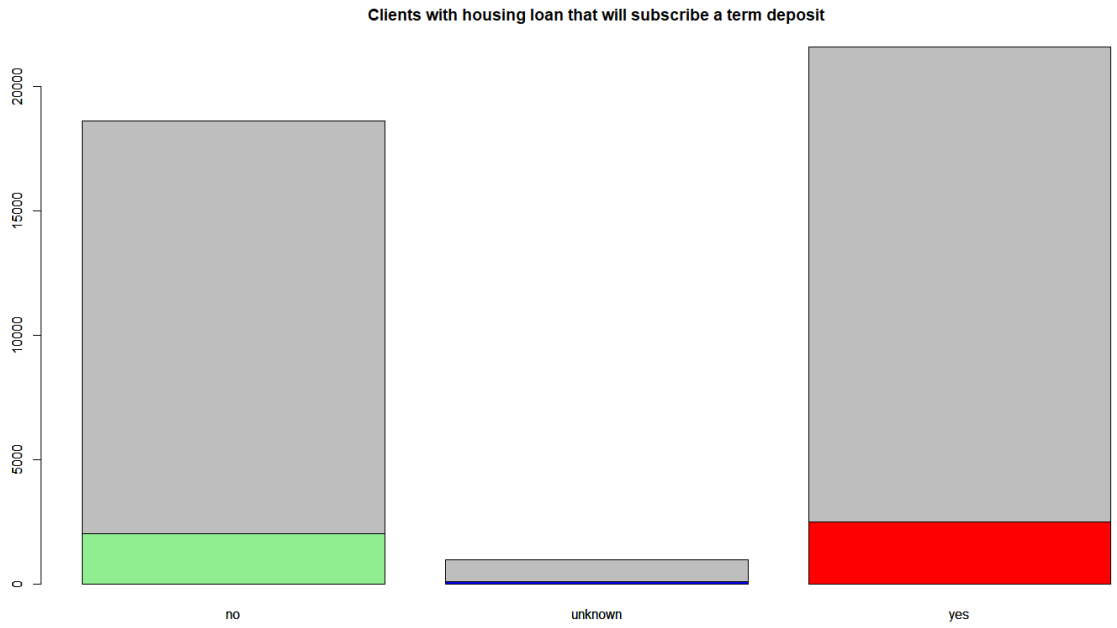
Zdecydowana większość klientów nie ma długu, wśród nich jest wyższa konwersja sprzedaży – 12,9%, niż dla niezdeklarowanych osób – 5,1%.



Oczywiście danych dla klientów z niespłaconym kredytem jest zbyt mało, żeby wykorzystać tą zmienną w analizie – zmienna binarna powinna mieć instancje po stronie „yes”, a 3 to za mało, żeby uznać to za regułę. Stąd ta zmienna nie zostanie wykorzystana w modelu.

Czy klient ma kredyt hipoteczny?

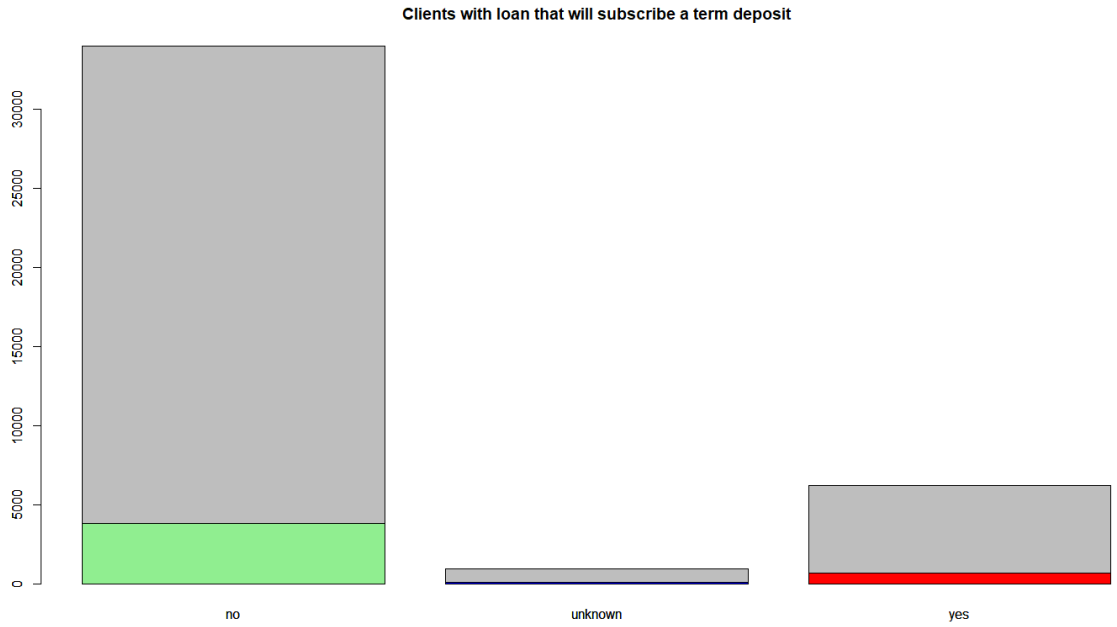
Inna sytuacja jest w przypadku kredytów hipotecznych:



Klientów, którzy mają kredyt hipoteczny jest nieco więcej niż tych, którzy nie mają. Co więcej, osoby mające taki kredyt chętniej zgadzały się na lokatę terminową – 11,6% procent klientów, podczas gdy osoby bez kredytu zgadzały się w 10,9%. Jest to jednak zbliżony do siebie wynik, stąd ciężko rozstrzygnąć o tym, czy ten atrybut ma znaczenie.

Czy klient ma pożyczkę osobistą?

Pożyczka osobista jest mniej popularna wśród klientów banku niż kredyt hipoteczny. Jednak odsetek sprzedaży lokaty terminowej kształtuje się na podobnym poziomie zarówno wśród tych, którzy wzięli pożyczkę osobistą, i tych którzy nie wzięli:

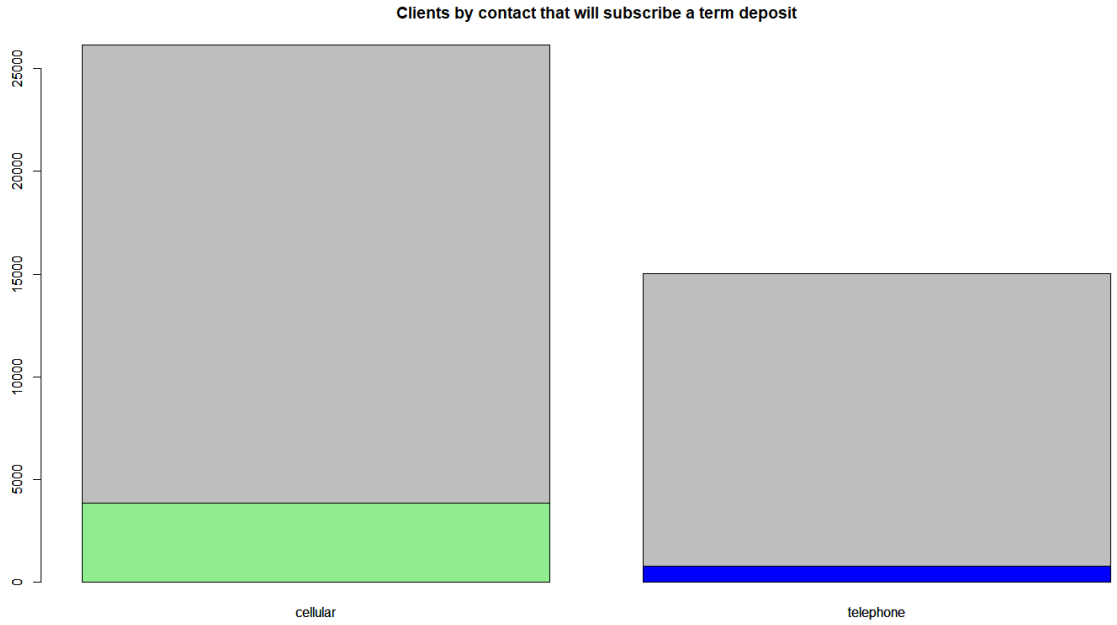


Wśród osób bez pożyczki osobistej odsetek sprzedaży wyniósł 11,3%, zaś wśród klientów mających pożyczkę 10,9% - stąd ciężko mówić o znacznej różnicy.

Informacje związane z bieżącą kampanią marketingową

Rodzaj kontaktu

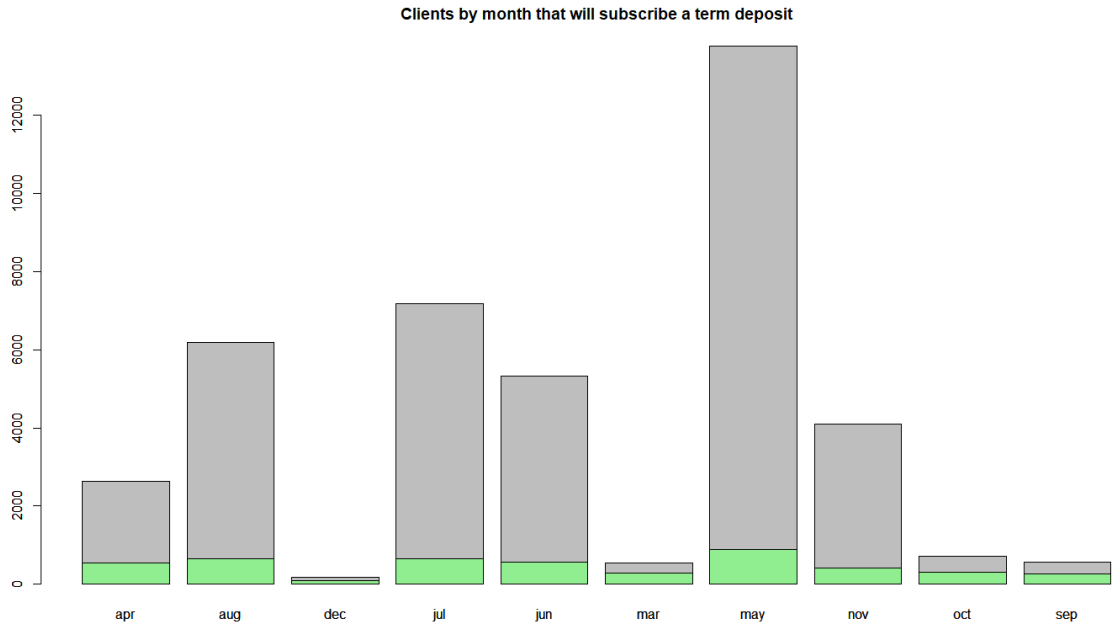
Kampania była prowadzona dwoma kanałami – w telefonii komórkowej oraz stacjonarnej:



Jako że kampania była w okresie Maj 2008 – Listopad 2010, znaczna część kontaktów to te na telefony komórkowe. Co ciekawe, w przypadku tych kontaktów konwersja sprzedaży jest wyższa niż dla telefonów stacjonarnych. W kanale telefonii komórkowej klienci decydowali się w 14,7% procent przypadków, zaś w drugim kanale tylko 5,2%.

Miesiąc ostatniego kontaktu

Kampania była najbardziej zintensyfikowana w okresie maj – sierpień, a więc w okresie wakacyjnym.



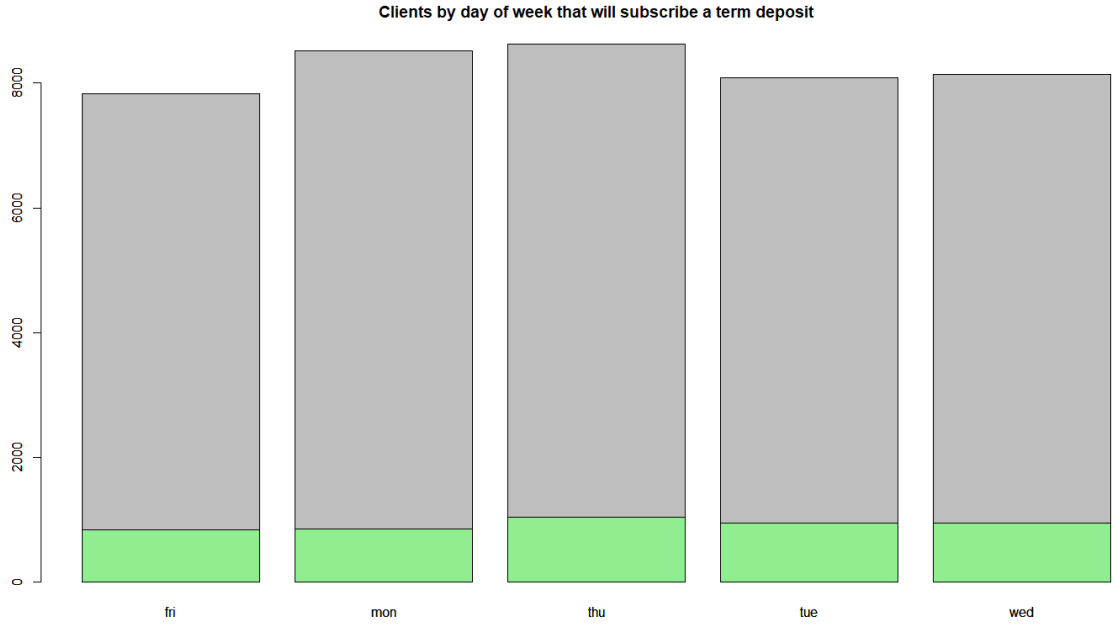
Wyraźnie można zaobserwować, że kampania była zintensyfikowana w okresie letniego półrocza, przy czym dużo kontaktów było również w listopadzie. W poniższej tabeli zestawiono współczynniki konwersji dla sześciu najpopularniejszych miesięcy:

Miesiąc	kwiecień	maj	czerwiec	lipiec	sierpień	listopad
%	20,5	6,4	10,5	9	10,6	10,1

Z istotnych miesięcy, dobrze wypadł kwiecień gdzie konwersja wyniosła 20,5%. Zmienna jest istotna, widać związek między poszczególnymi miesiącami a sprzedażą lokat.

Dzień ostatniego kontaktu

Zmienna ta, określa dzień tygodnia, w jakim nastąpił ostatni kontakt z klientem. Można uznać, że dane są w miarę równomiernie rozłożone.



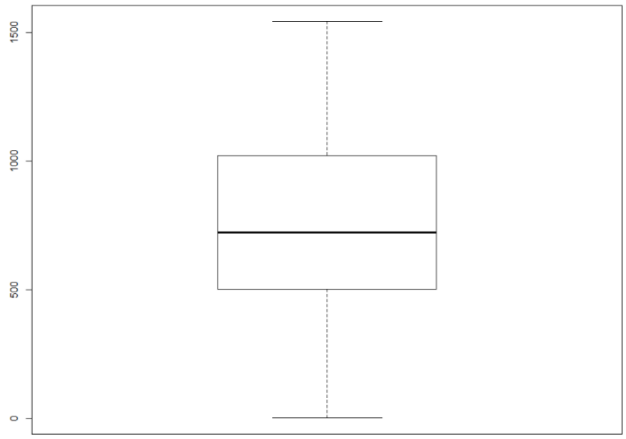
Współczynniki konwersji dla powyższych danych zamieszczono w poniższej tabeli:

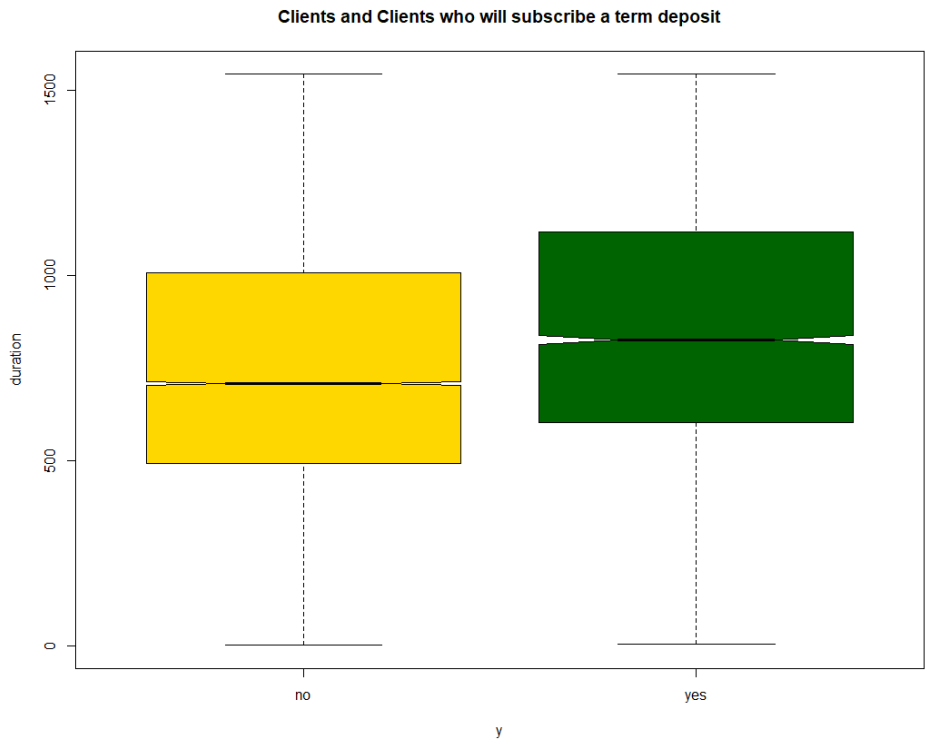
Dzień tygodnia	poniedziałek	wtorek	środa	czwartek	piątek
%	9,9	11,8	11,6	12,1	10,8

Wynika z tego, że większy sukces miały kontakty w środku tygodnia. Jest to ciekawa obserwacja, mówiąca wiele o tym w jakim nastroju są klienci w poszczególne dni tygodnia. Może ona świadczyć też o efektywności telemarketerów, którzy bliżej weekendu gorzej wykonują swoją pracę. Wnioski oparte na tej zmiennej same w sobie są dużą wskazówką dla zarządzania bankiem, dlatego tym bardziej znajdą się w modelu.

Czas ostatniej rozmowy

Rozstęp międzykwartylowy zawiera się w przedziale 502-1022. Wąsy rozciągają się na cały zakres danych.





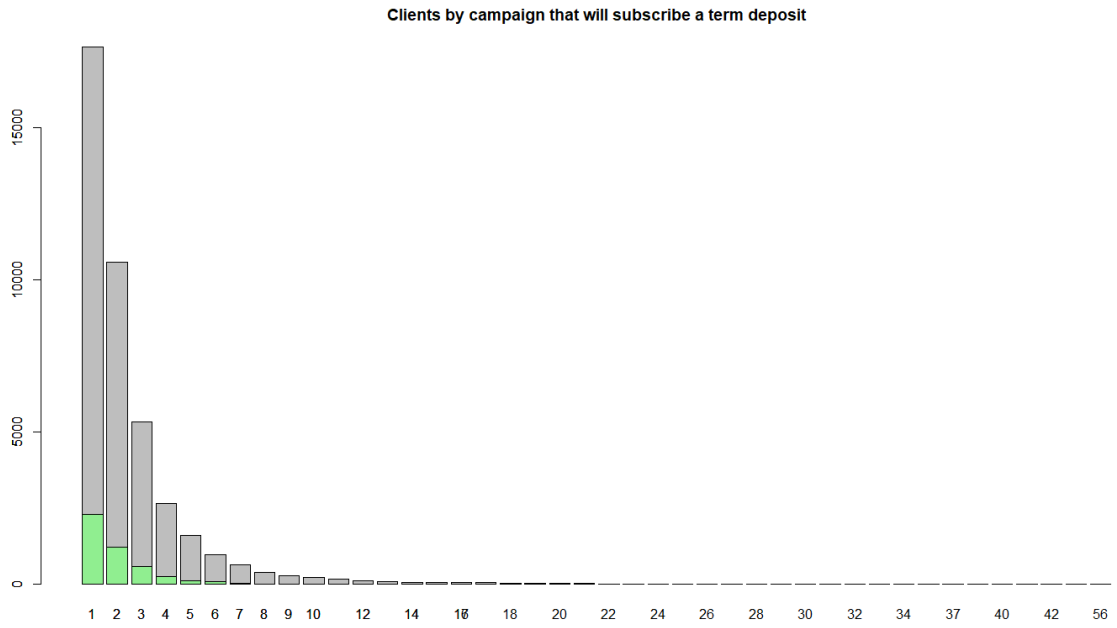
Można stwierdzić, że dla klientów, którzy wykupili lokatę długość rozmowy przesunęła się na dłuższe wartości. Oznacza to, że statystycznie klient, który kupił lokatę rozmawiał dłużej niż ten, który tego nie zrobił.

Z założeń naszego projektu musimy jednak uznać tą zmienną jako tzw. „wyciek danych”, ponieważ nie powinniśmy mieć informacji o trwaniu rozmowy przed jej odbyciem. Ta dana nie powinna zatem zostać użyta w modelu.

Inne informacje o kampanii

Ilość kontaktów wykonanych w bieżącej kampanii

Jak powszechnie wiadomo, telemarketerzy są bardzo wytrwali w swojej pracy i często potrafią dzwonić kilka razy, żeby uzyskać odpowiedź. Nie inaczej było w tej kampanii:

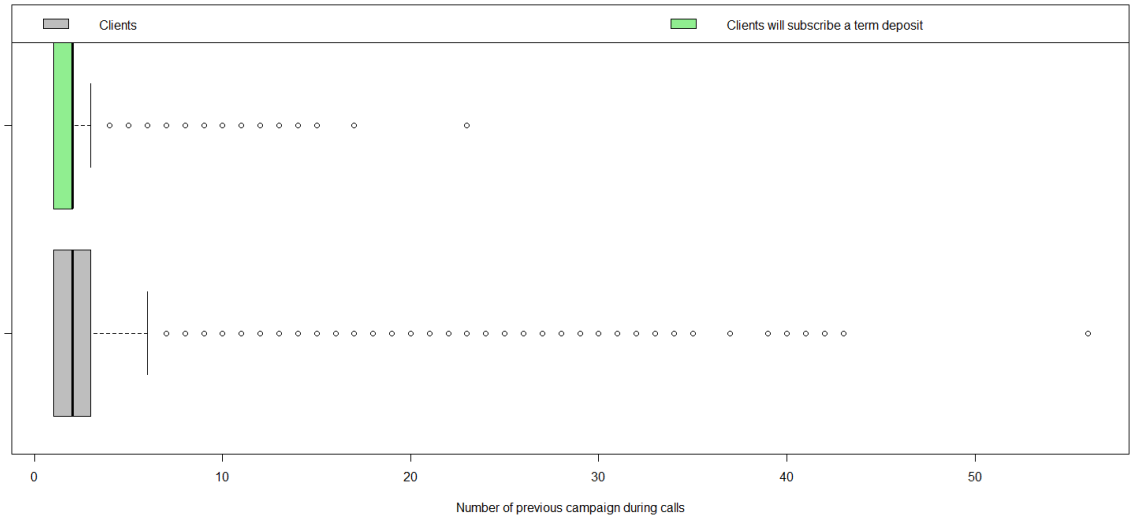


Telemarketerzy do najbardziej lubianych przez nich klientów potrafili dzwonić nawet po kilkadziesiąt razy. Być może jest to związane z faktem, że niektóre kontakty były nieaktualne, stąd wielokrotne próby kontaktu, które kończyły się fiaskiem.

Ilość prób	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	17	23
Zakup	2300	1211	574	249	120	75	38	17	17	12	12	3	4	1	2	4	1

Jak widać w powyższej tabeli, najwięcej sprzedaży to te zawarte po pierwszym lub po drugim kontakcie. Samych kontaktów jest zdecydowanie najwięcej właśnie w przedziale od 1 do 3.

Powyżej 6 kontaktów skupia się już znacznie niższa liczba sprzedaży, ponadto niewiele jest takich klientów, do których dzwoniono tak często. W przypadku wielu kontaktów dość trudno oszacować które kontakty były związane z tym, że do klienta nie można było się dodzwonić, a które z tym, że klient odkładał swoją decyzję w czasie i chciał się dłużej zastanowić nad wykupem lokaty terminowej. Te obserwacje można potraktować jako odstające.



Sprawdźmy jeszcze jak wyglądała konwersja w przypadku klientów, do których dzwoniono od 1 do 6 razy:

Ilość prób	1	2	3	4	5	6
%	13	11,5	10,7	9,4	7,5	7,7

W tabeli widać dość wyraźnie trend, że wraz z ilością kontaktów wykonanych do klienta spada szansa na to, że klient wykupi lokatę. Wygląda na to, że klienci dobrze wiedzą czego chcą, i ciągłe dzwonienie do nich nie jest opłacalne tak bardzo jak przy jednym czy dwóch kontaktach.

Ile dni minęło od ostatniego kontaktu

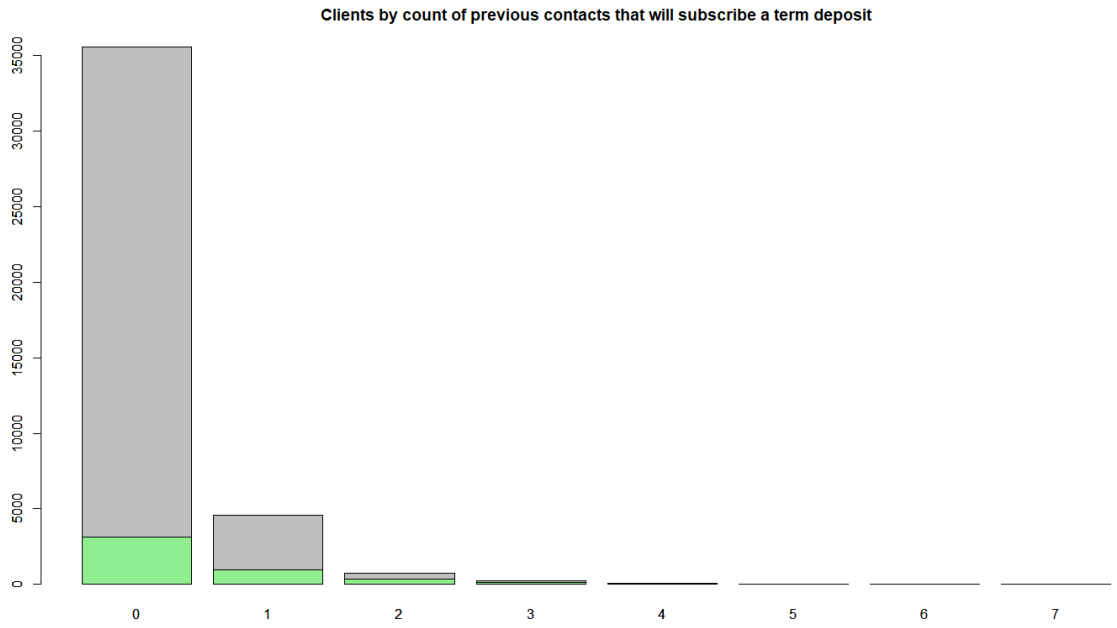
Ilość dni od ostatniego kontaktu określa wszystkie kontakty z klientem, a nie tylko w bieżącej kampanii, stąd nie jest to zmienna bezpośrednio związana z poprzednim atrybutem. Analizując częstotliwość wystąpień poszczególnych wartości w tej zmiennej, zdecydowanie najczęściej występowała liczba 999, oznaczająca brak poprzedniego kontaktu. Występowała ona aż 39 673 razy wśród wszystkich klientów. Inne najczęściej występujące wartości to 3 dni – 439 razy, oraz 6 dni – 412 razy. Widać jednak, że są to liczby o 2 rzędy wielkości mniejsze od braku kontaktu, dlatego ciężko je ze sobą porównywać.

Co ciekawe, rozmowy wykonane po 3 lub 6 dniach od ostatniego kontaktu, cechują się naprawdę dobrą konwersją, odpowiednio 67,9% oraz 70,1%. Wygląda to jak zorganizowana akcja w ramach reklamowania lokaty terminowej do osób, które wcześniej zdecydowały się na jakąś inną usługę.

Nie mniej jednak, ten trend dotyczy małej liczby osób względem całego zbioru, dlatego nie zdecydujemy się na wykorzystanie tej zmiennej.

Ilość kontaktów przed bieżącą kampanią

Podobnie jak w poprzednim przypadku, przeważali klienci, z którymi się wcześniej nie kontaktowano.



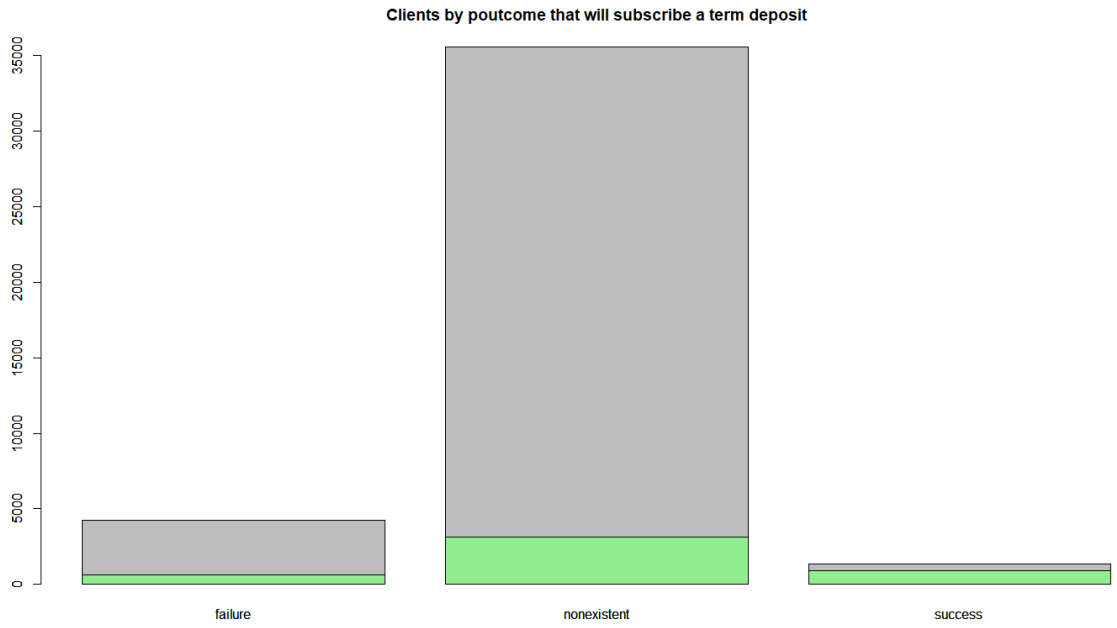
Liczba kontaktów ‘0’ nie pokrywa się jednak z liczbą wystąpień ‘999’ dni od ostatniego kontaktu, co może świadczyć o tym, że te zmienne nie są ze sobą powiązane. Ciężko zatem rozstrzygnąć co dokładnie oznacza ta zmienna, nie została wystarczająco opisana przez dostarczyciela zbioru danych.

Sporą część klientów stanowią ci z którymi kontaktowano się już wcześniej raz – jest ich 11,1% w całym zbiorze. Co ciekawe, o ile wśród klientów, z którymi nie kontaktowano się wcześniej, konwersja sprzedaży jest na poziomie 8,8%, to wśród tych klientów stanowi ona aż 21,2%.

Z racji, że takich klientów jest znaczna ilość w zbiorze danych, należy mieć na uwadze tą zależność.

Wynik poprzedniej kampanii marketingowej

Ostatnią zmienną informującą o kliencie jest to, czy jeśli był z nim kontakt w poprzedniej kampanii marketingowej, to czy zakończył się sukcesem czy nie.



W przypadku tej zmiennej, można znaleźć powiązanie z poprzednią zmienną, o liczbie kontaktów w poprzedniej kampanii. Liczba kontaktów wynoszących 0 oraz liczba wyników poprzedniej kampanii ‘nonexistent’, czyli brak wyniku – jest taka sama – wynosi 35 563. Również analiza tych klientów potwierdziła, że każdy klient, który miał 0 kontaktów w poprzedniej kampanii, również ma status ‘nonexistent’.

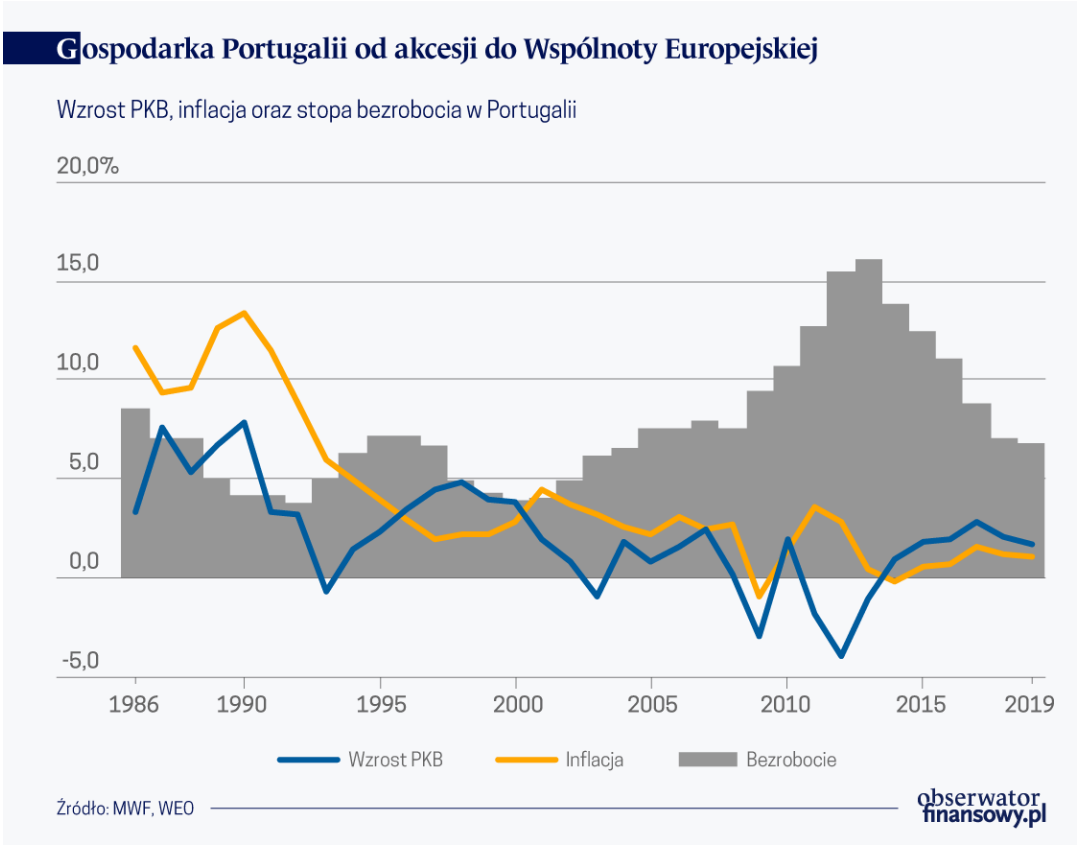
W zbiorze jest zatem 5625 klientów, do których wcześniej kierowano inne kampanie. 4252 z nich nie udało się do nich przekonać, zaś 1373 tak. Współczynnik konwersji wśród klientów, którzy zgodzili się na poprzednią ofertę wynosi aż 65,1%. Ta informacja potwierdza zatem, że opłaca się kierować kampanię do klientów, którzy wcześniej zgodzili się na inną usługę. Klienci, którzy mają usługę z poprzedniej kampanii stanowią 19,3% wszystkich klientów, którzy kupili lokatę terminową.

Jako że ta zmienna i poprzednia, ilość kontaktów w poprzedniej kampanii, są ze sobą powiązane, należy odrzucić jedną z nich. Decydujemy się pozostawić wynik poprzedniej kampanii marketingowej, ze względu na to, że ta zmienna ma tylko 3 kategorie i jest łatwiejsza do analizy, bardziej przejrzysta.

Informacje dotyczące społecznych i ekonomicznych wskaźników, w momencie, gdy kontaktowano się z klientem:

Analizując 5 liczbowych zmiennych odnoszących się do kontekstu społecznego i ekonomicznego, w momencie prowadzenia rozmów, uznaliśmy, że mogą być one kluczowe w działaniu modelu.

Związki kontekstu z decyzją klientów o zakupie lokaty terminowej są mocno widoczne, być może wynika to z działania samego banku lub dużej świadomości klientów o sytuacji ekonomicznej na rynku. Kampania marketingowa była w okresie kryzysu na europejskim rynku gospodarczym, wiadomo też, że w 2010 roku sytuacja była lepsza niż w 2008:

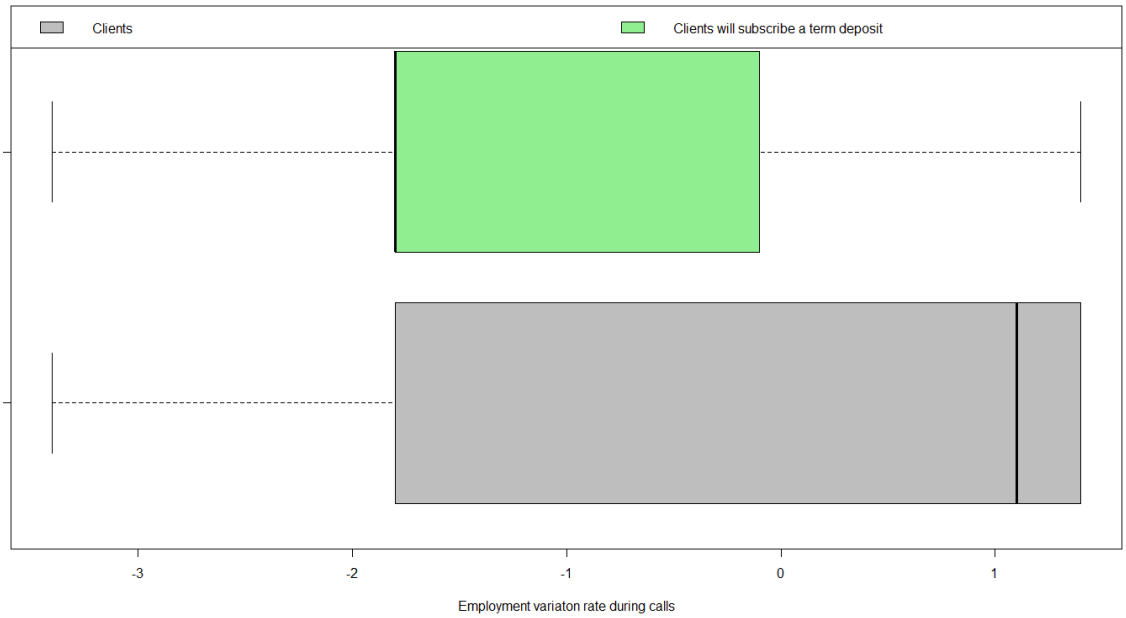


1 Sytuacja ekonomiczna w Portugalii, źródło: https://www.obserwatorfinansowy.pl/forma/rotator/portugalia-rozwaznie-wychodzi-na-prosta/

Widać to po wskaźniku wzrostu PKB na powyższym wykresie. A jak wiadomo, kampania marketingowa trwała w okresie maj 2008 – listopad 2010. Stąd może być wpływ nastrojów społecznych i sytuacji ekonomicznej na zakup lokat.

Wskaźnik zmienności zatrudnienia w ujęciu kwartalnym

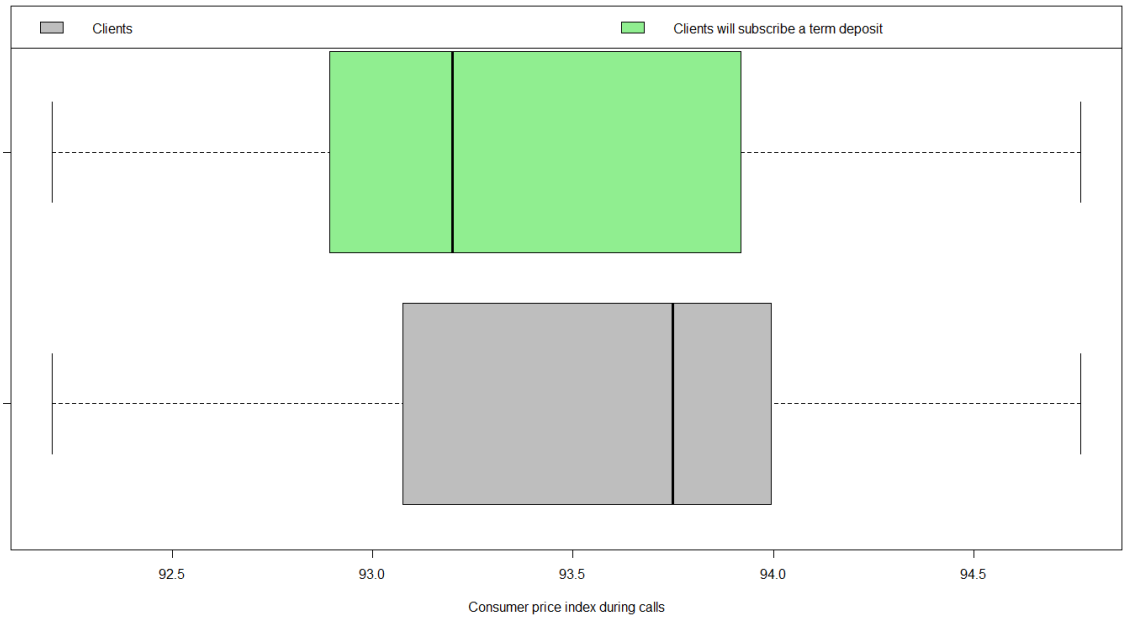
Wskaźnik zmienności zatrudnienia w ujęciu kwartalnym, mówi o tym jak zmienił się wskaźnik zatrudnienia, czyli odsetka osób w wieku produkcyjnym pracujących zawodowo, względem poprzedniego kwartału. Wpływ tego wskaźnika na zakup lokat dobrze widać na wykresie skrzynkowym:



W okresie, w którym ta zmienna była ujemna, zostało zaobserwowane więcej zakupów niż wtedy kiedy była dodatnia. Wygląda zatem na to, że w okresie, kiedy zatrudnienie spadało, więcej osób było chętnych na zabezpieczenie w postaci lokaty terminowej. To cenna uwaga dla naszego modelu.

Indeks kosztów konsumenckich w ujęciu miesięcznym

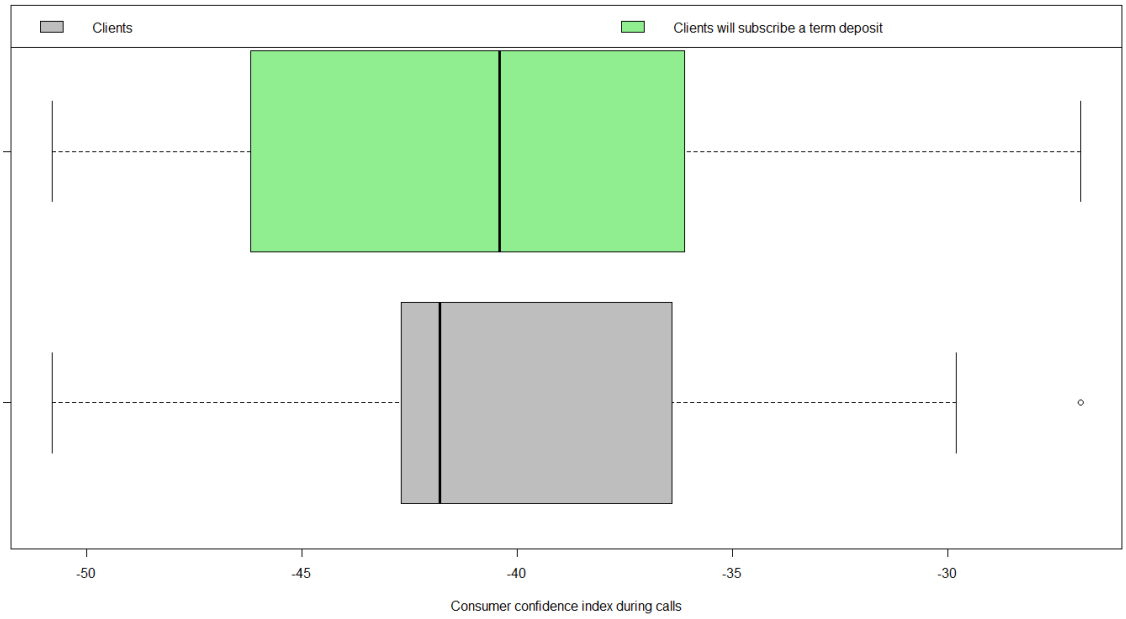
Indeks CPI to indeks zmian cen towarów i usług konsumpcyjnych, jest dobrym wskaźnikiem inflacji i siły nabywczej pieniądza. Jest wyliczany jako iloraz kosztów koszyka zakupów w danym miesiącu względem kosztów w miesiącu bazowym.



Widać zatem, że tu również występuje wyraźne przesunięcie wykresu na korzyść mniejszego CPI. Wygląda zatem na to, że klienci byli chętniejsi na lokatę w okresie, kiedy wartość nabywcza pieniędzy była niższa – w tym okresie bardziej zależało im na zabezpieczeniu oszczędności na lokacie.

Wskaźnik ufności konsumenckiej w ujęciu miesięcznym

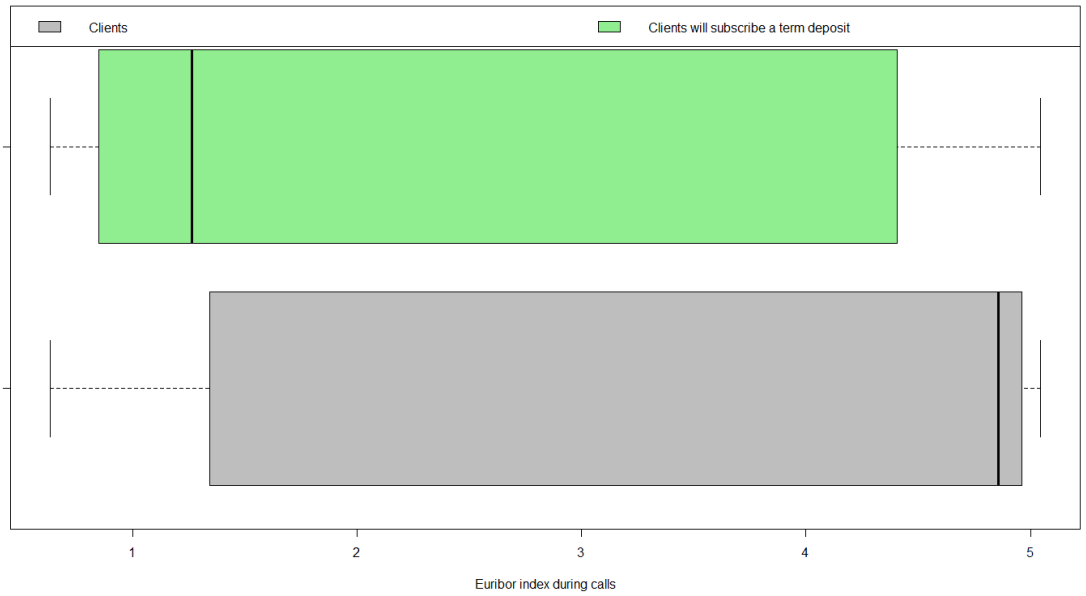
Ta miara określa poziom optymizmu dotyczącego sytuacji gospodarczej, przejawiającej się w wysokości oszczędności i wydatków konsumentów. Banki obserwują comiesięczne zmiany wskaźnika jako jeden z czynników wpływający na podejmowane w danym czasie decyzje. Być może ma to związek z tym, kiedy kampania była zintensyfikowana, a kiedy nie, albo jest to związane z samopoczuciem finansowym klientów.



O ile rozstęp międzykwartylowy klientów, którzy powiedzieli „tak” telemarketerowi jest szerszy niż dla wszystkich klientów, to widać, że mediana oraz prawy „wąs” wykresu przesunęły się w kierunku wyższego zaufania konsumenckiego. Jest to logiczny objaw – klienci są pewniejsi i chętnie decydują się na usługę.

Wskaźnik EURIBOR z 3 miesięcy w ujęciu dziennym

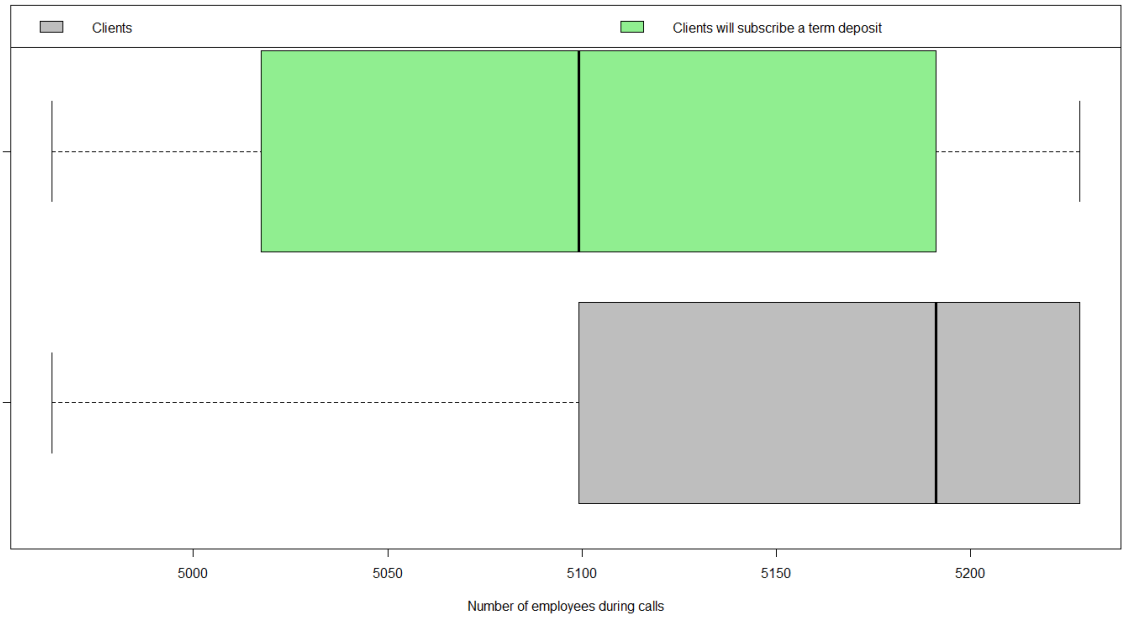
Wskaźnik EURIBOR to międzybankowa stopa procentowa, stosowana do udzielania pożyczek między bankami na rynku międzybankowym Unii Europejskiej. Analitycy finansowi stosują ją jako punkt odniesienia przy ustalaniu stopy procentowej innych pożyczek.



W momencie, kiedy transakcje między bankami były oprocentowane niższą stopą, więcej klientów decydowało się na wykup lokaty terminowej – być może w tym okresie jej oprocentowanie było bardziej korzystne, stąd ten związek.

Liczba pracowników w ujęciu kwartalnym

Ta zmienna dotyczy liczby osób zatrudnionych w rozważanym przez nas banku, a zatem dotyczy działań strategicznych banku, a nie decyzji klientów. Liczba ta mogła wpłynąć na budżet banku, a co za tym idzie na to, że bank dysponował większymi środkami, przez co mógł bardziej aktywizować kampanię marketingową:



Gdy bank zatrudniał mniej osób, klienci częściej decydowali się na lokatę terminową, widać to po tym jak przesunięty jest wykres zdecydowanych na usługę klientów względem wszystkich. Wynika to być może z bardziej intensywnych działań banku w tym okresie, a może z bardziej korzystnego oprocentowania lokaty terminowej.

Wybór zmiennych istotnych

W poprzednim paragrafie przeanalizowano 20 zmiennych zawartych w modelu. Podsumowując przemyślenia spisane powyżej, wybrane zmienne istotne to:

- wiek
- typ zatrudnienia
- stan cywilny
- wykształcenie
- rodzaj kontaktu
- miesiąc ostatniego kontaktu
- dzień kontaktu
- ilość kontaktów wykonanych w bieżącej kampanii
- wynik poprzedniej kampanii
- wskaźnik zmienności zatrudnienia w ujęciu kwartalnym
- indeks kosztów konsumenckich w ujęciu miesięcznym
- wskaźnik ufności konsumenckiej w ujęciu miesięcznym
- wskaźnik EURIBOR z 3 miesięcy w ujęciu dziennym

- liczba pracowników banku w ujęciu kwartalnym

Uzupełnienie danych dla wybranych zmiennych istotnych

Analizując strukturę eksplorowanych zmiennych, widać, że w niektórych zmienne kategorycznych część rekordów jest nieokreślonych i oznaczonych jako ‘unknown’. Z kolei wśród zmiennych liczbowych, które wybrano, tylko jedna ma wiele obserwacji odstających.

W kolumnie ‘typ zatrudnienia’ 330 rekordów jest oznaczona jako nieznane, zatem tyle osób nie chciało podać swojego zatrudnienia. Niestety, z racji faktu, że mamy aż 11 kategorii tego atrybutu, ciężko wybrać typy, którymi można byłoby uzupełnić braki w danych. Dominantą są pracownicy administracji, ale czy właśnie oni by nie podali informacji o swoim zatrudnieniu? Na szczęście dziur w danych jest o 2 rzędy wielkości mniej niż wszystkich rekordów, stąd zdecydowano się usunąć rekordy bez tej informacji.

Inną sytuację mamy w zmiennej ze stanem cywilnym. Mimo że braków jest zaledwie 80, to wyraźną dominację w zbiorze mają osoby w związku małżeńskim, a typów stanu cywilnego jest zaledwie 3. Stąd statusy ‘unknown’ zamieniliśmy na ‘married’.

Z racji trudności w wyborze jak zastąpić brakujące informacje o poziomie wykształcenia klientów, i tego, że braków jest mniej o 1 rząd wielkości, zdecydowaliśmy się nie usuwać braków w danych dla tej kolumny.

Jedyną zmienną liczbową, którą wykorzystamy i która ma wyraźnie dużo obserwacji odstających jest informacja o ilości poprzednich kontaktów w ramach bieżącej kampanii. Wykres skrzynkowy wyraźnie wskazuje na to, że nietypowe są obserwacje powyżej 6 kontaktów, stąd decyzja o usunięciu rekordów z taką wartością. Zastąpienie tej zmiennej inną wartością byłoby problematyczne, ciężko ingerować w tą liczbę.

W wyniku powyższych operacji zbiór z 41 188 instancji zmalał do 38 479. Utraciliśmy nieco obserwacji, ale wciąż dysponujemy dużym zbiorem. Wyczyszczony zbiór pozbawiony zmiennych uznanych za nieistotne został zapisany w pliku *bank-additional-cleaned.csv*.

Podział na zbiory uczące i testowe

W poniższej tabeli zamieszczono podział na dane testowe i uczące:

	%	Liczba obserwacji
Treningowy	80	30 783
Testowy	20	7696

Oczywiście, wszystkie obserwacje zostały wymieszane przed podziałem.

Pierwsze drzewo klasyfikacyjne

Opis drzewa

Dla pierwszego drzewa parametry podziału ustawiono na wysokie wartości z racji wykorzystania dużego zbioru.

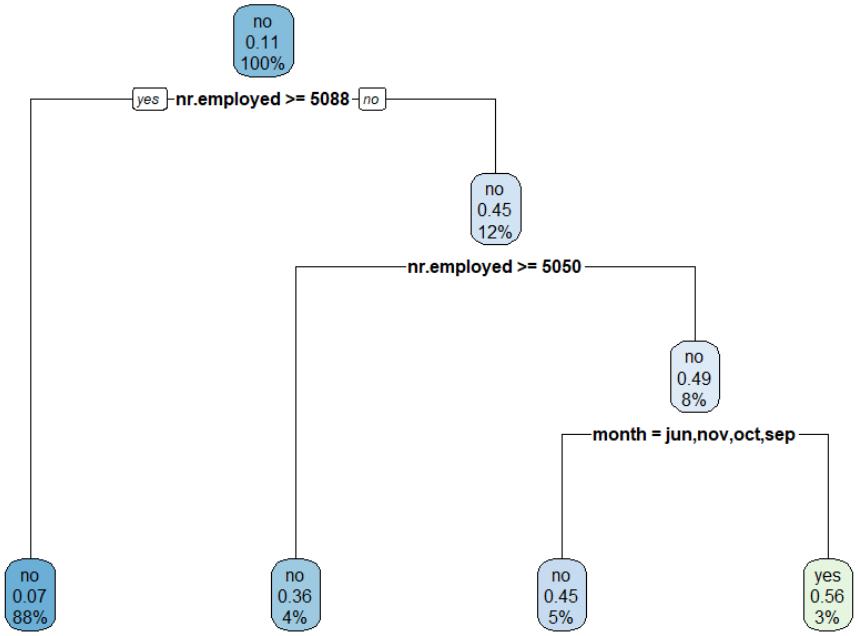
Parametr	Wartość
minsplit	1000
minbucket	1000

xval	0

Oczywiście parametr xval dotyczący cross-walidacji ustawiono na 0 a reszta parametrów miała domyślne wartości.

Analiza drzewa

Otrzymane drzewo prezentuje się następująco:



Reguły prezentują się następująco:

```
1) root 32950 3727 no (0.88688923 0.11311077)
  2) nr.employed>=5087.65 28979 1947 no (0.93281342 0.06718658) *
  3) nr.employed< 5087.65 3971 1780 no (0.55175019 0.44824981)
    6) nr.employed>=5049.85 1341  485 no (0.63832960 0.36167040) *
    7) nr.employed< 5049.85 2630 1295 no (0.50760456 0.49239544)
     14) month=jun,nov,oct,sep 1544  692 no (0.55181347 0.44818653) *
     15) month=apr,aug,dec,jul,mar,may 1086  483 yes (0.44475138 0.55524862) *
>
```

Poniżej zamieszczono tabelę z wynikami predykcji:

Przewidywanie	no	yes
no	7268	763
yes	130	150

Otrzymano skuteczność 89%. Nie można uznać tego, za miarodajny wynik, gdyż, gdy zmienna przewidywana wynosi “yes” skuteczność wynosi trochę ponad 16% a na dobrym określeniu tej wartości najbardziej w tym przypadku zależy. Wynik jest zawyżony przez dysproporcje pomiędzy ilością poszczególnych wartości zmiennej przewidywanej. Niestety niektóre zmienne są praktycznie nieużywane przez model o czym świadczy obliczona istotność zmiennych, którą pokazano poniżej:

```
Variable importance
   nr.employed      euribor3m   emp.var.rate  cons.conf.idx cons.price.idx          month
            27             23             16             14             12              8
```

Drugie drzewo - poprawa jakości drzewa przez zmianę jego parametrów

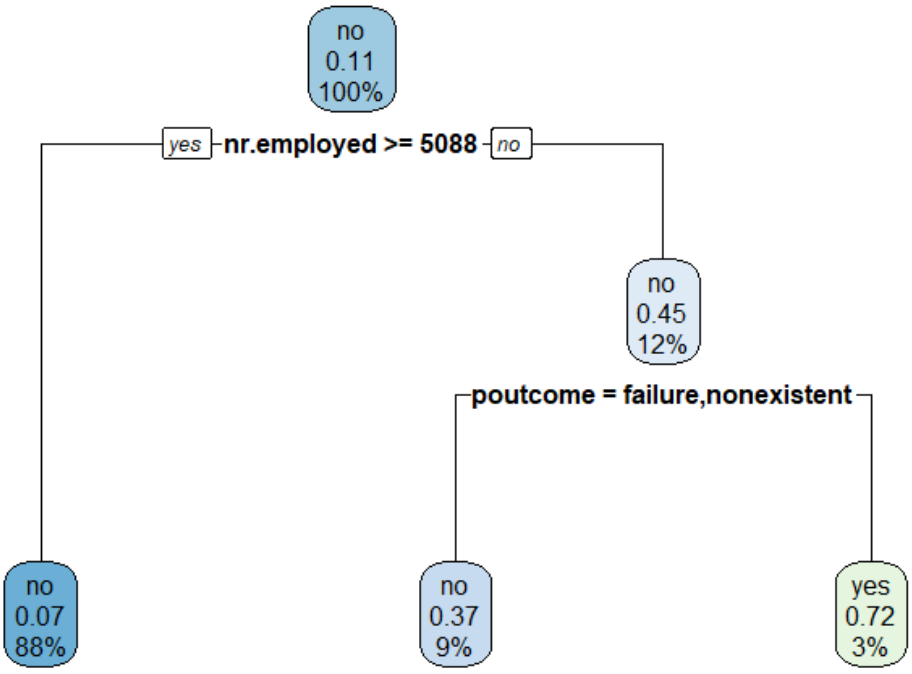
Opis drzewa

Parametry drugiego drzewa:

Parametr	Wartość
minsplit	500
minbucket	500
xval	0

Analiza drzewa

Otrzymane drzewo prezentuje się następująco:



Klasyfikacja ograniczyła się tylko do rozpatrzenia wartości dwóch zmiennych.

Reguły prezentują się następująco:

```
1) root 32950 3727 no (0.88688923 0.11311077)
  2) nr.employed>=5087.65 28979 1947 no (0.93281342 0.06718658) *
  3) nr.employed< 5087.65 3971 1780 no (0.55175019 0.44824981)
    6) poutcome=failure,nonexistent 3061 1126 no (0.63214636 0.36785364) *
    7) poutcome=success 910  256 yes (0.28131868 0.71868132) *
```

Poniżej zamieszczono tabelę z wynikami predykcji:

Przewidywanie	no	yes
no	7195	763

yes	57	150

Uzyskano lepsze wyniki dla wartości “no” zmiennej przewidywanej. Niestety wyniki przeciwnej wartości pozostały bez zmian.

```
variable importance
   nr.employed       euribor3m    emp.var.rate  cons.conf.idx cons.price.idx           month        poutcome
            26              22              15              14              12               7               4
```

Udało się zwiększyć istotność zmiennej “poutcome”.

Trzecie drzewo - poprawa jakości drzewa przez zmianę jego parametrów

Opis drzewa

Postanowiono zastosować następujące wartości parametrów:

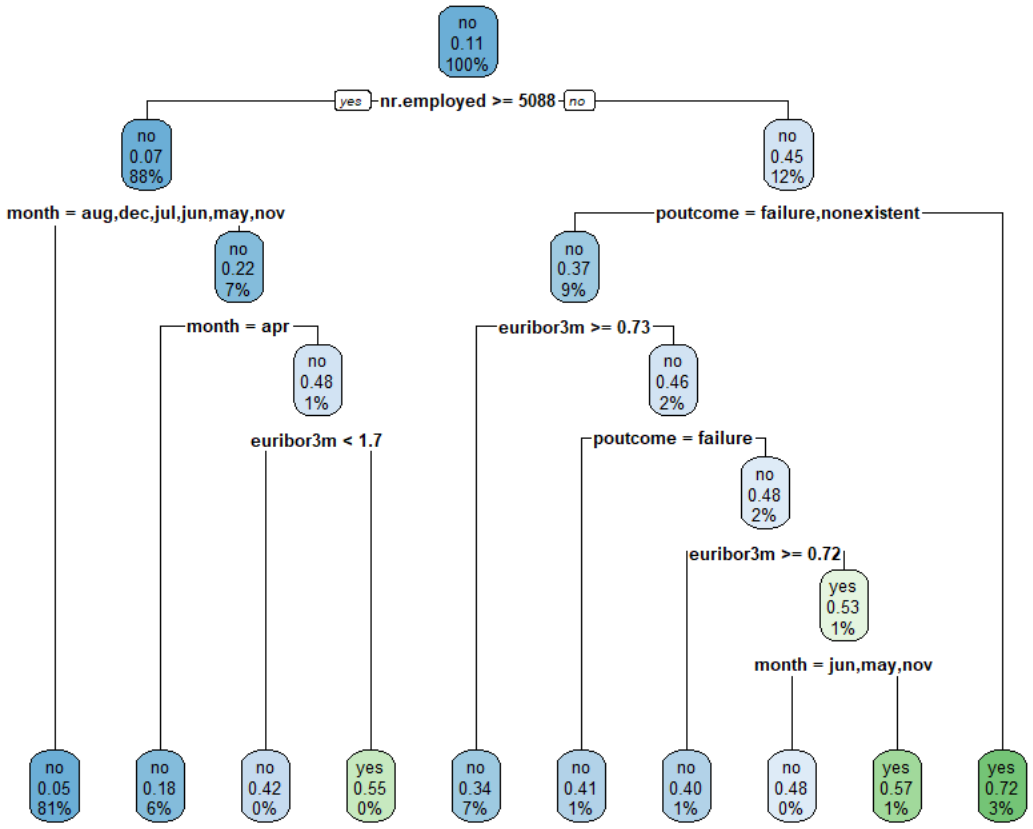
Minbucket =100

Minsplit =100

Cp = 0.001

Analiza drzewa

Otrzymano następujące drzewo:



Reguły prezentują się następująco:

```
1) root 32950 3727 no (0.88688923 0.11311077)
  2) nr.employed>=5087.65 28979 1947 no (0.93281342 0.06718658)
    4) month=aug,dec,jul,jun,may,nov 26734 1462 no (0.94531308 0.05468692) *
    5) month=apr,mar,oct 2245  485 no (0.78396437 0.21603563)
     10) month=apr 1975  356 no (0.81974684 0.18025316) *
     11) month=mar,oct 270  129 no (0.52222222 0.47777778)
       22) euribor3m< 1.7145 152   64 no (0.57894737 0.42105263) *
       23) euribor3m>=1.7145 118   53 yes (0.44915254 0.55084746) *
  3) nr.employed< 5087.65 3971 1780 no (0.55175019 0.44824981)
    6) poutcome=failure,nonexistent 3061 1126 no (0.63214636 0.36785364)
     12) euribor3m>=0.7305 2243  753 no (0.66428890 0.33571110) *
     13) euribor3m< 0.7305 818  373 no (0.54400978 0.45599022)
       26) poutcome=failure 302  123 no (0.59271523 0.40728477) *
       27) poutcome=nonexistent 516  250 no (0.51550388 0.48449612)
         54) euribor3m>=0.7155 178   71 no (0.60112360 0.39887640) *
         55) euribor3m< 0.7155 338  159 yes (0.47041420 0.52958580)
          110) month=jun,may,nov 142   68 no (0.52112676 0.47887324) *
          111) month=apr,dec,mar 196   85 yes (0.43367347 0.56632653) *
    7) poutcome=success 910  256 yes (0.28131868 0.71868132) *
```

Poniżej zamieszczono tabelę z wynikami predykcji:

Przewidywanie	no	yes
no	7230	716
yes	95	197

Sumaryczne wyniki są lepsze niż w poprzednich modelach. Pomimo wzrostu błędnych wartości przewidywań w przypadku “no” znacznie więcej wzrosła wartośc przeciwna co czyni model bardziej skutecznym kosztem większej złożoności.

Czwarte drzewo utworzone przy pomocy kroswalidacji

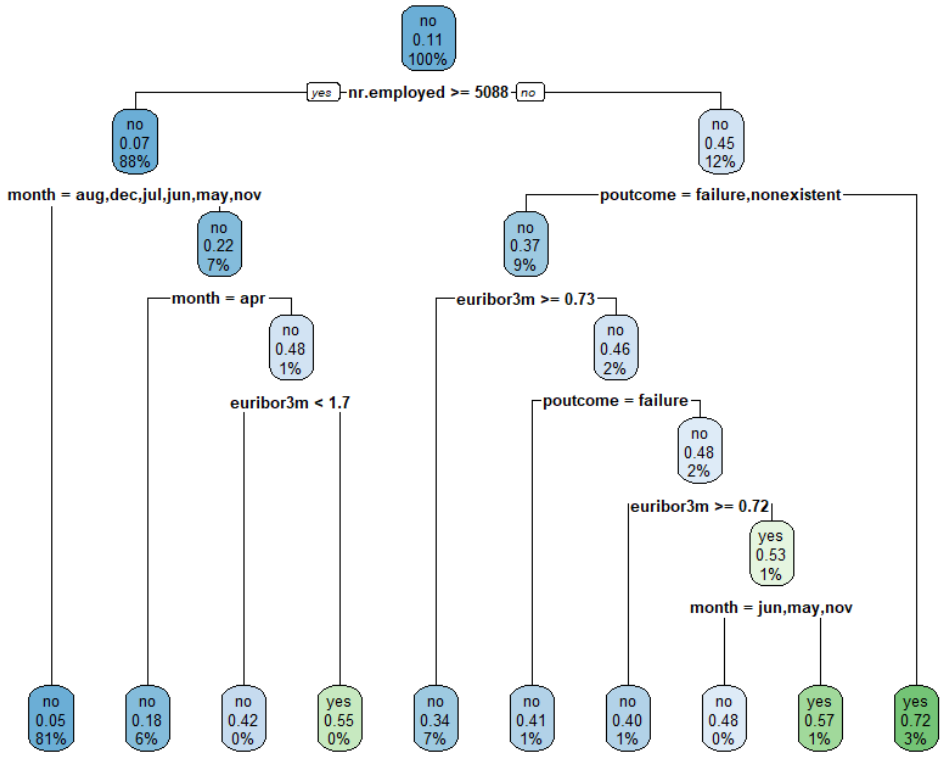
Opis drzewa

Parametry drzewa:

Parametr	Wartość
minsplit	200
minbucket	100
xval	20
cp	0,001

Analiza drzewa

Otrzymano następujące drzewo:



Reguły prezentują się następująco:

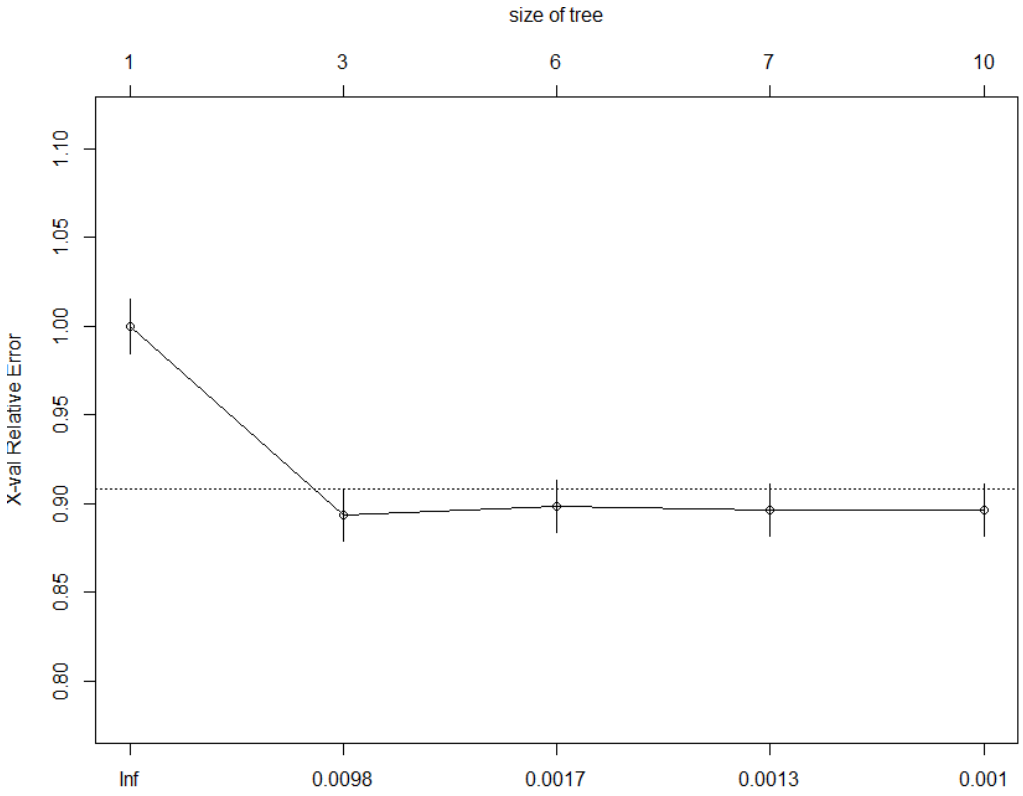
```
1) root 32950 3727 no (0.88688923 0.11311077)
  2) nr.employed>=5087.65 28979 1947 no (0.93281342 0.06718658)
    4) month=aug,dec,jul,jun,may,nov 26734 1462 no (0.94531308 0.05468692) *
    5) month=apr,mar,oct 2245  485 no (0.78396437 0.21603563)
     10) month=apr 1975  356 no (0.81974684 0.18025316) *
     11) month=mar,oct 270  129 no (0.52222222 0.47777778)
       22) euribor3m< 1.7145 152   64 no (0.57894737 0.42105263) *
       23) euribor3m>=1.7145 118   53 yes (0.44915254 0.55084746) *
  3) nr.employed< 5087.65 3971 1780 no (0.55175019 0.44824981)
    6) poutcome=failure,nonexistent 3061 1126 no (0.63214636 0.36785364)
     12) euribor3m>=0.7305 2243  753 no (0.66428890 0.33571110) *
     13) euribor3m< 0.7305 818  373 no (0.54400978 0.45599022)
       26) poutcome=failure 302  123 no (0.59271523 0.40728477) *
       27) poutcome=nonexistent 516  250 no (0.51550388 0.48449612)
         54) euribor3m>=0.7155 178   71 no (0.60112360 0.39887640) *
         55) euribor3m< 0.7155 338  159 yes (0.47041420 0.52958580)
          110) month=jun,may,nov 142   68 no (0.52112676 0.47887324) *
          111) month=apr,dec,mar 196   85 yes (0.43367347 0.56632653) *
    7) poutcome=success 910  256 yes (0.28131868 0.71868132) *
```

Poniżej zamieszczono tabelę z wynikami predykcji:

Przewidywanie	no	yes
no	7230	716
yes	95	197

Uzyskane wyniki są takie same jak w poprzednim modelu. Ustawienie parametrów takich samych jak w poprzednich modelach i tylko zmienianie liczby cross-walidacji nie wpływało na zmianę wyników. Zmieniał się tylko czas budowy modelu.

Wykres xval error:



Jak widać wraz ze wzrostem rozmiarów drzewa błąd pozostaje praktycznie niezmienny.

Podsumowanie

Wnioski

Problemem z jakim przyszło nam się zmierzyć w obliczu zbioru danych, który wybraliśmy, było przewidzenie wyniku kampanii marketingowej, na podstawie danych o kliencie, do którego jest adresowana, a także na postawie informacji o kontekście ekonomicznym i społecznym, w którym były prowadzone rozmowy.

Podczas eksploracji danych przeanalizowano wszystkie 20 atrybutów wejściowych jakie były zawarte w zbiorze. Po sprawdzeniu jak bardzo związane są ze zmienną przewidywaną i jaką mają strukturę, zdecydowano się nie skorzystać z 5 zmiennych, których związek z zakupem lokaty terminowej uznano za zbyt mały. Dodatkowo spotkano się z tzw. *wyciekiem danych*, czyli zawarciem w zbiorze danych wejściowych informacji o danej wyjściowej. Taka sytuacja była przy zmiennej opisującej czas trwania rozmowy, która w założeniu zagadnienia klasyfikacyjnego się jeszcze nie odbyła. Dlatego zdecydowano się nie skorzystać z tej zmiennej. Do modelu wykorzystano zatem 14 zmiennych.

Informacje o kliencie, takie jak wiek, miejsce zatrudnienia, wykształcenie oraz stan cywilny, uznaliśmy za cenne informacje. Pozwalają one na dobre określenie kim jest odbiorca kampanii i dopasowanie go do określonego segmentu rynku. Analiza współczynników konwersji sprzedaży pozwoliła na wyłapanie prawidłowości wiążących klienta z jego decyzją o zakupie lokaty terminowej.

Sposób prowadzenia kampanii, określony przez takie parametry jak miesiąc, w którym wykonano połączenie, a także dzień tygodnia również miały związek z decyzją klienta. W danych odkryto prawidłowość, że sprzedaż lepiej idzie w środku tygodnia, zaś poniedziałek i piątek są gorzej produktywnymi dniami. Z kolei spojrzenie na miesiące kontaktów pokazało, że kampania była intensywniejsza w miesiącach kwiecień – sierpień oraz w listopadzie.

Inne parametry w zbiorze, takie jak ilość kontaktów w bieżącej kampanii marketingowej oraz wynik poprzedniej kampanii przeprowadzonej na klientach, także wykazują korelację tym czy klient powie „tak”, czy „nie”. Wykryto trend, że wraz z ilością kontaktów spada szansa na to, że klient zgodzi się na oferowaną mu propozycję. Z kolei klienci, którzy w poprzedniej kampanii byli chętni na proponowaną usługę, aż w 65% zdecydowali się na lokatę terminową. Stanowią 19% klientów z lokatą terminową.

Sytuacja ekonomiczna i społeczna w czasie rozmowy miała znaczny wpływ na jej wynik. Zdecydowano się wybrać wszystkie 5 wskaźników opisujących kontekst rozmowy. Analiza wykresów skrzynkowych jasno potwierdziła ich znaczenie, a najwyraźniejszy wpływ ma parametr opisujący liczbę pracowników aktualnie pracujących w banku.

Interpretacja modeli

Uruchomienie podstawowego modelu drzewa decyzyjnego pozwoliło na szybkie zweryfikowanie wybranych przez nas zmiennych istotnych. Algorytmy zaimplementowane w środowisku R, wyliczające ważność zmiennych wprowadzonych do modelu, odrzuciły połowę z wybranych przez nas zmiennych, pozostawiając ich jedynie 7. Właśnie, dlatego możliwe byłoby stworzenie modeli uwzględniających tylko te zmienne.

Mimo, uwzględnienia wielu zmiennych dla budowy modelu jedynie wartościowe były zmienne ekonomiczne. Jest to zastanawiające, gdyż eksploracja pozwoliła zauważyć związki między poszczególnym wartościami obserwacji innych zmiennych a wartościami przewidywanymi. Możliwe, że wynika to z tego, iż dla zmiennych ekonomicznych istniało wiele typów obserwacji o wartości, które niemal w całości potrafiły odpowiadać wartości “yes” w zmiennej przewidywanej. Sprawiało to, że zmienne ekonomiczne znacznie wyróżniały się na tle pozostałych.

Uzyskana skuteczność modeli obraca się wokół 90%. Jednakże dla wartości parametru przewidywanego “yes” jest ona znacznie mniejsza niż dla przeciwnej wartości. Ma to swoje wady zalety. W ten sposób na pewno poznamy metody jakie statystycznie NIE PROWADZĄ do uzyskania danej wartości oczekiwanej. Jednakże bardziej wartościowe byłoby uzyskanie informacji o tym, jakie czynniki w dużym stopniu mogłyby prowadzić do przekonania klienta do stworzenia lokaty. Istotne jest to, że model tworzony metodą krosswalidacji praktycznie nie zmieniał innych modelów o tych samych parametrach. Zastanawiające jest również to, że w każdym modelu na podstawie jednej zmiennej(nr.employed) model tworzył liść który na starcie klasyfikował 80-90% danych. Wydaje, się logiczne, że liczba zatrudnionych ludzi w banku ma tak duży wpływ na podjęcie lokaty/

Naszym zdaniem na uzyskane wyniki wpłynęły ogólnie mała ilość wartości “yes” przewidywanej. Jeżeli w zbiorze takich wartości jest mało to jeszcze trudniej określić jej związki z pozostałymi wartościami obserwacji innych zmiennych. Możliwe jest, że również niemożliwe jest utworzenie takiego klasyfikatora o wysokiej skuteczności dla zmiennej “yes”, gdyż po prostu brakuje większych trendów w związkach między zmiennymi, a wypatrzenie samych niewielkich powiązań nie wystarcza do zbudowania modelu o wysokiej skuteczności.