2 LUTEGO 2020

KAMPANIA MARKETINGOWA BANKU

ANALIZA DANYCH KAMPANII MAKETINGOWEJ I OKREŚLENIE PRAWDOPODOBIEŃSTWA ZAŁOŻENIA DEPOZYTU PRZEZ KLIENTA

SŁAWOMIR LISOWSKI
POLITECHNIKA GDAŃSKA
Wydział Fizyki Technicznej i Matematyki Stosowanej
Studia podyplomowe
Kierunek: Inżynieria Danych – Data Science

1. Wstęp.

Opracowanie ma na celu stworzenie modelu predykcyjnego na podstawie bazy danych zawierającej dane osób, które brały udział w kampanii marketingowej jednego z portugalskich banków. Model ma jak najtrafniej przewidzieć, czy dany klient założy depozyt w danym banku, czy też nie. Dzięki temu bank w przyszłości będzie mógł skierować kampanie marketingowe do odpowiednich klientów.

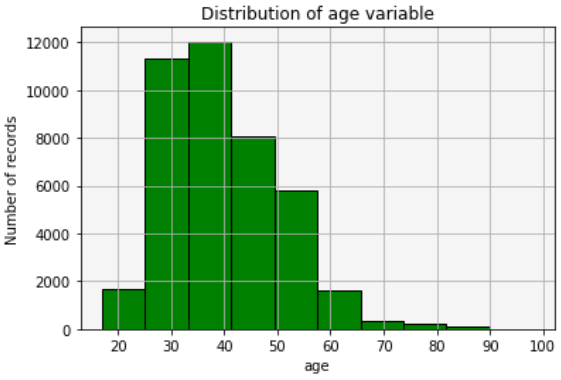
2. Opis danych.

Zbiór danych zawiera 41188 rekordów, które odpowiadają osobom do których była skierowana kampania marketingowa. Każdy rekord ma 21 atrybutów:

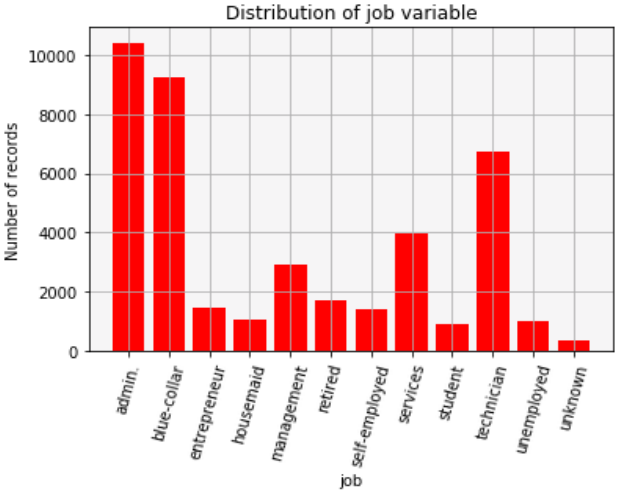
lp	Nazwa	Nazwa polska	Cechy	Typ danych
1	age	wiek	wiek potencjalnego klienta	numeric
2	job	zawód	typ pracy: admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown	categorical
3	maritial	stan cywilny	"divorced","married","single","unknown"; note: "divorced" means divorced or widowed	categorical
4	education	wykształcenie	"basic.4y","basic.6y","basic.9y","high.school" ,"illiterate","professional.course", "university.degree","unknown"	categorical
5	default	opóźnienie w spłacie kredytu?	"no","yes","unknown"	categorical
6	housing	kredyt mieszkaniowy?	"no","yes","unknown"	categorical
7	loan	pożyczka gotówkowa?	"no","yes","unknown"	categorical
8	contact	typ komunikacji	"cellular","telephone"	catrgorical
9	month	miesiąc ostatniego kontaktu	"jan", "feb", "mar", ..., "nov", "dec"	categorical
10	day_of_week	dzień tygodnia ostatniego kontaktu	"mon","tue","wed","thu","fri"	catrgorical
11	duration	długość ostatniego kontaktu w sekundach		numeric
12	campaign	liczba kontaktów z klientem wykonanych podczas obecnej kampani		numeric
13	pdays	dni od ostatniego kontaktu z poprzedniej kampani	999 oznacza, że nie było kontaktu w poprzedniej kampanii marketingowej	numeric
14	previous	liczba kontaktów przed kampanią		numeric
15	poutcome	wynik poprzedniej kampanii	'failure','nonexistent','success'	categorical
16	emp.var.rate	wskaźnik zmiany zatrudnienia	wyliczany kwartalnie	numeric
17	cons.price.idx	indeks cen konsumpcyjnych	wyliczany miesięcznie - relacja cen reprezentatywnego zestawu dóbr konsumpcyjnych w kolejnych latach badania do ceny tego koszyka dóbr w roku bazowym	numeric
18	cons.conf.idx	indeks zaufania konsumentów	wyliczany miesięcznie - na wskaźnik składają się dwa subindeksy: ocena obecnej kondycji gospodarstwa domowego oraz wskaźnik oczekiwań kondycji gospodarstwa w przyszłości.	numeric
19	euribor3m	stawka euribor 3 miesięczna		numeric
20	nr.employed	liczba pracowników		numeric
21	y	zmienna celu	Informacja czy klient założył depozyt ‘yes’, or ‘no’	binary

Są to dane dotyczące osób, do których była skierowana kampania marketingowa jednego z portugalskich banków. Dane pochodzą ze strony https://archive.ics.uci.edu/ml/datasets/Bank+Marketing# . W zbiorze występują dane numeryczne i kategoryczne. Zmienną celu jest zmienna y i przyjmuje dwie wartości: ‘yes’ – jeżli osoba założyła depozyt w banku, ‘no’ – jeżeli osoba nie założuła takiego depozytu.

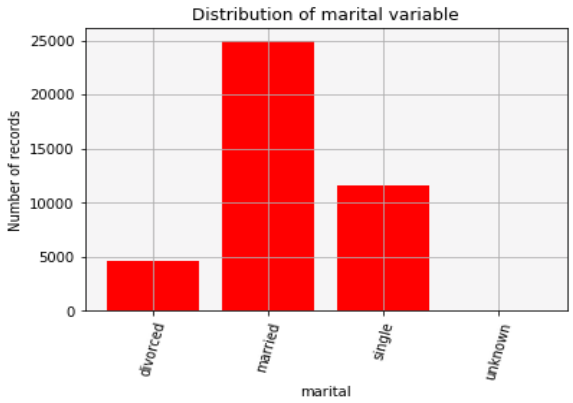
3. Rozkłady wybranych zmiennych losowych.



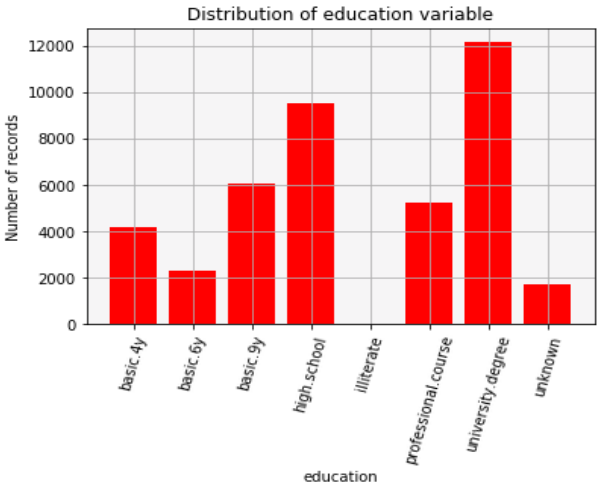
Rysunek 1. Zmienna losowa wiek ma rozkład normalny. Większość osób do, których była skierowana kampania jest między 30 a 40 rokiem życia



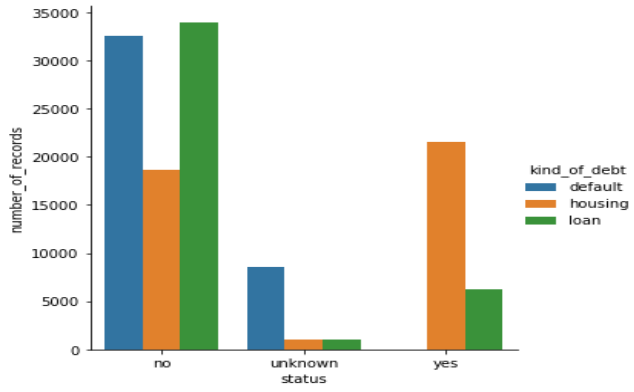
Rysunek 2. Większość osob do których była kierowana kampania tto pracownicy administracyjni oraz fizyczni



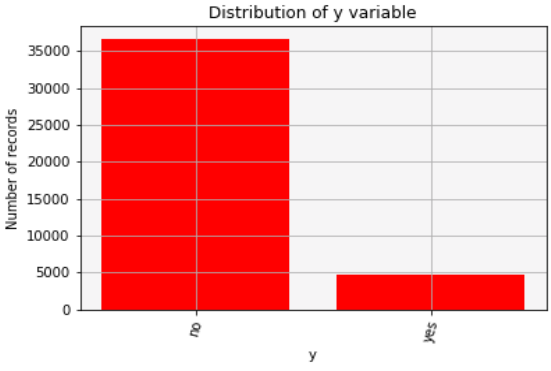
Rysunek 3. Jeżeli chodzi o stan cywilny to największy odsetek osób stanowią osoby w związku małżeńskim



Rysunek 4. Wykształcenie osób do których kierowana była kampania było najczęściej na poziomie ukończenia liceum



Rysunek 5

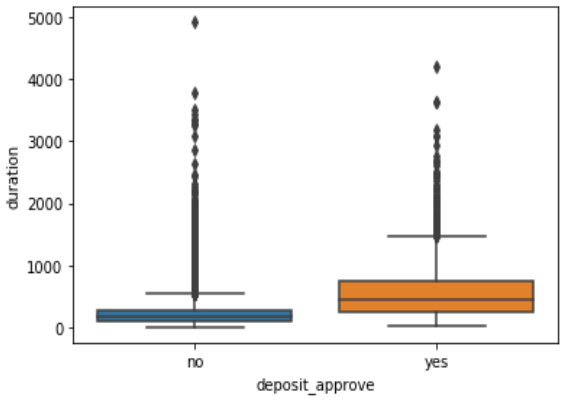


Rysunek 6

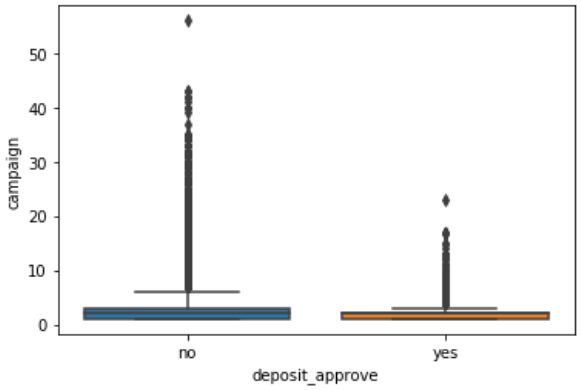
Pozostałe rozkłady wszystkich zmiennych znajdują się w plikach pdf dołaczonych do projektu.

4. Rozkład zmiennych opisowych w stosunku do zmiennej celu.

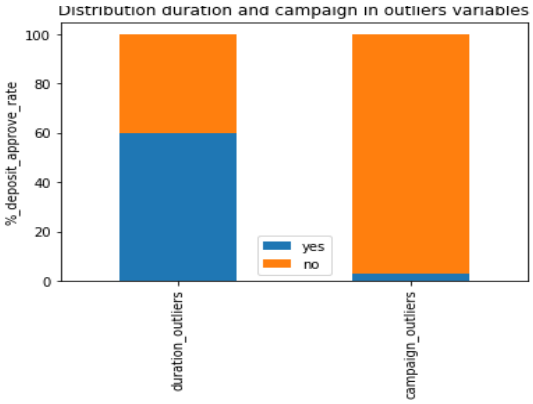
Analizę rozpoczynamy od dwóch zmiennych, które zawierają wartości znacznie odstające od średniej. Są to zmienne duration – czas ostatniej rozmowy z klientem w sekundach oraz campaign – liczba kontaktów z klientem podczas ostatniej kampanii.



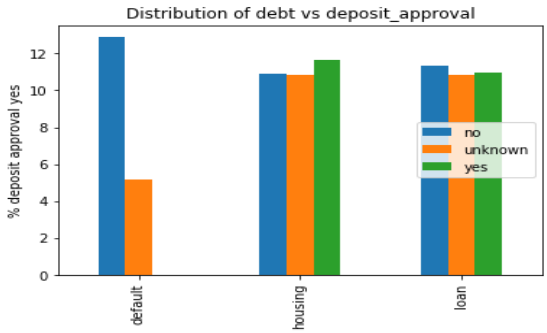
Rysunek 7 Na wykresie typu boxplot widać wartości znacznie odbiegające od średniej w przypadku osób, które założyły depozyt lub nie w zależności od długości ostatniego kontaktu telefonicznego. Możemy jednak zauważyć zależność im dłuższa rozmowa tym chętniej zakładany depozyt



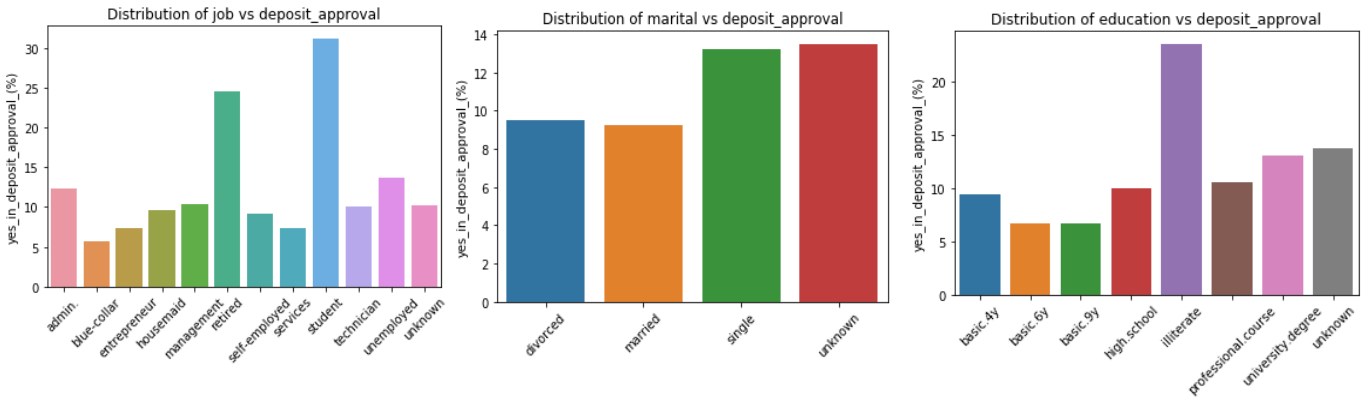
Rysunek 8 Wykres przedstawia zależność między liczbą kontaktów w ostatniej kampanii a osobami zakładającymi depozyty.



*Rysunek 8 Wykres przedstawia rozkład wartości oddalonych (dalej niż 3*odch. Std od średniej)w zmiennych duration i campaign było ich odpowiednio 861 i 869. Rekordy te zostaną wyrzucone ze zbioru*



Rysunek 9 Wykres przedstawia zależność między statusem zadłużeń a zmienną celu. W zmiennych housing i loan proporcje są równe i nie wnoszą żadnych informacji dlatego usuwamy te kolumny. Kolumna default, zawiera rekordy z wartościami no lub unknown dlatego tę kolumnę również wyrzucamy



Rysunek 10,11,12 Wykresy pokazują rozkłady poszczególnych zmiennych względem zmiennej celu. Wykresy pokazują, że poszczególne grupy względem zmiennej celu rozkładają się inaczej niż ogólne liczebności poszczególnych grup. W przypadku zmiennych 'job' i 'education' widać przewagi poszczególnych grup, natomiast w przypadku zmiennej marital, udział osób zakładajacyh depozyty w poszczególnych grupach jest podobny z nieznaczną przewagą wartości 'single' i 'unknown'.

Zmienną pdays zmieniamy na binarną gdzie wartośc 1 przyjmują rekordy w których był kontakt w poprzedniej kampanii i 0 jeżeli takiego kontaktu nie było

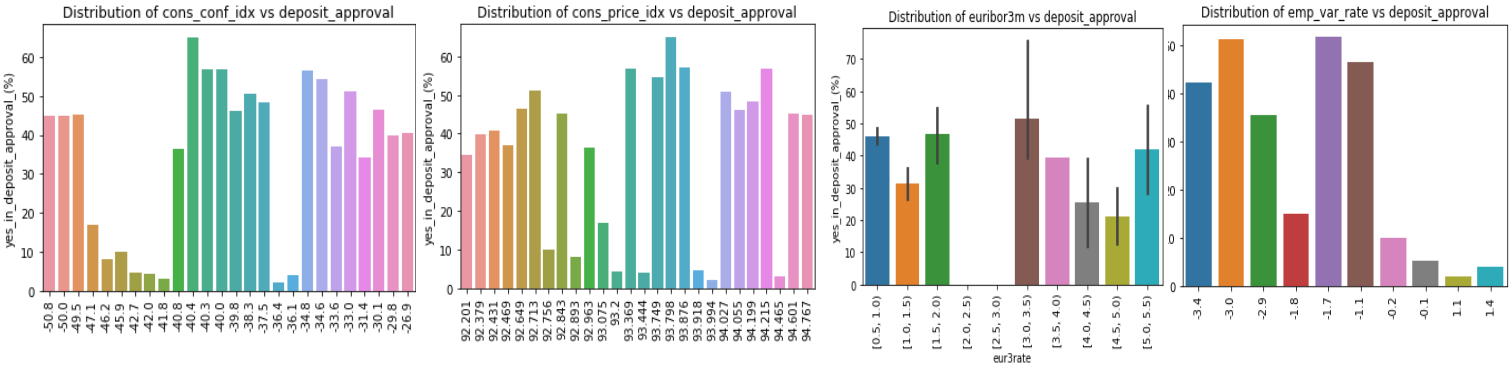
```
In [98]: bankData['pdays']=np.where(bankData['pdays']==999,0,1)

In [99]: bankData['pdays'].value_counts()
Out[99]:
0 36193
1 1375
Name: pdays, dtype: int64

In [100]: bankData.groupby(['pdays', 'deposit_approve']).size()
Out[100]:
pdays deposit_approve
0 no 33209
  yes 2984
1 no 503
  yes 872
dtype: int64
```

	yes_in_deposit_approval_(%)
fri	10.042849
mon	9.113060
thu	11.177474
tue	10.861183
wed	10.815308

W przypadku zmiennej day of week nie widać korelacji odnośnie dnia kontaktu a liczbą zakładanych depozytow dlatego odrzucamy tą zmienną

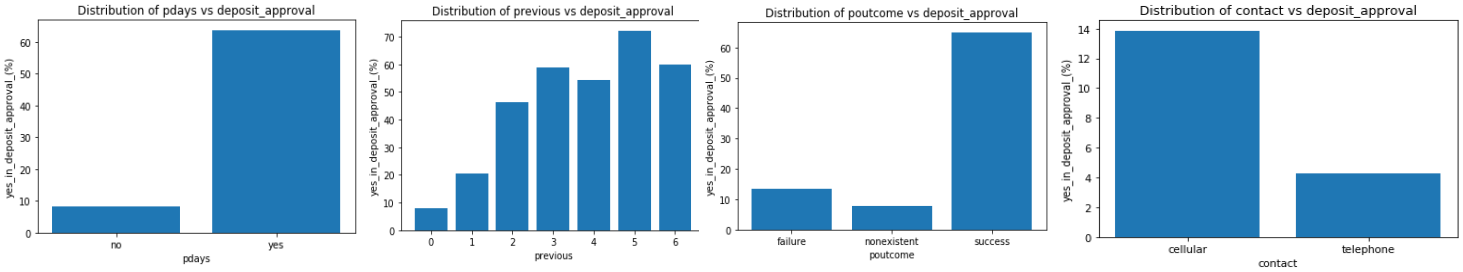
```
bankData_yes['nr_employed'].value_counts()
Out[20]:
5099.1    979
5228.1    619
5076.2    581
5017.5    444
4991.6    392
5008.7    360
4963.6    289
5195.8    194
5191.0    161
5023.5     85
5176.3      1
```

Wartości te oznaczają liczbę pracowników w ujęciu kwartalnym, nie widać tutaj jasnej zależności między tą zmienną a zmienną celu, dlatego tą zmienną odrzucamy. Nie wiemy również co dokładnie ta liczba oznacza

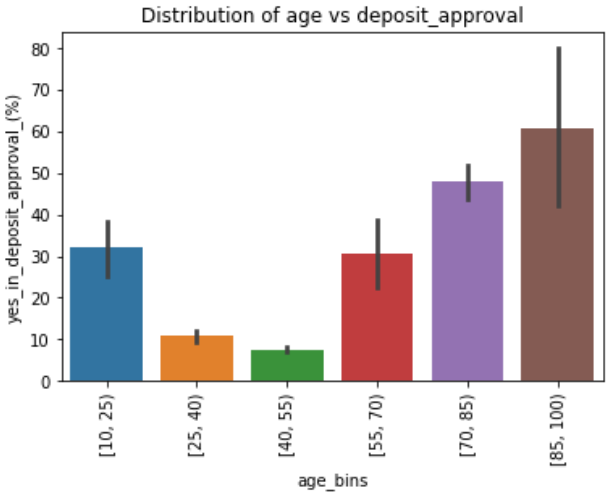


Rysunek 13,14,15,16 Wykresy przedstawiają zależność między zmiennymi opisującymi dane z gospodarki oraz zmienną celu.Na wykresach nie widać jednak jasnej zaleśności między poziomami tych indeksów a % zakładanych depozytów. Możemy jedynie zauważyć, że jeżeli indeksy osiągały wartość mediany to procent zakładanych depozytów był największy. Jedynie w przypadku zmiennej emp_var_rate opisującej stopę rotacji w zatrudnieniu pracowników w przypadku wartości poniżej -1, udział zakładanych depozytów był znacznie mniejszy.

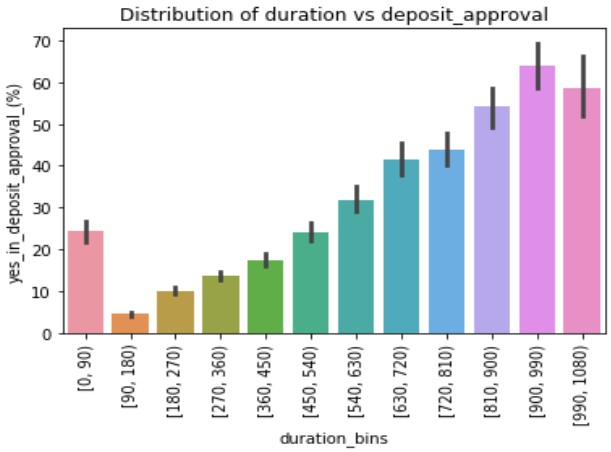
Sprawdzamy rozkłady pozostałych zmiennych względem zmiennej celu.



Rysunek 17,18,19,20 W przypadku zmiennych opisujących charakter kontaktu dostajemy wiele przydatnych informacji. Jeżeli w poprzedniej kampanii marketingowej był kontakt z klientem to ponad 60% tych klientów założyło depozyty w obecnej kampanii. Im większa liczba kontaktów przd kampanią tym chętniej były zakładane depozyty. Jeżeli w poprzedniej kampanii osoba założyła depozyt to w tej kampanii również to zrobiła. Osoby posiadające telefony komórkowe również chętniej zakładały depozyty



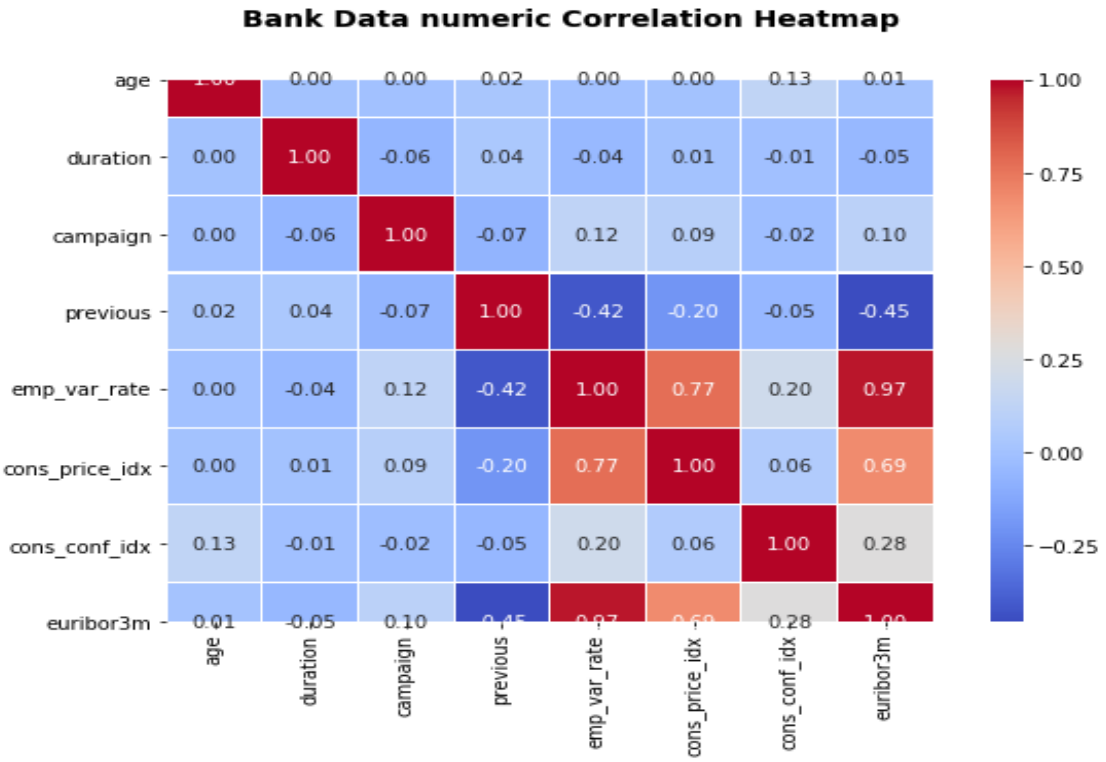
Rysunek 21. Zależność wieku do chęci założenia depozytu ma pozytywną korelację. Im więcej lat ma dana osoba tym chętniej zaklada depozyt.



Rysunek 22. Wykres pokazuje jasno – im dłużej trwał ostatni kontakt telefoniczny tym chętniej depozyty były zakładane.

5. Przekształcanie danych do zastosowania w algorytmach uczenia maszynowego.

Sprawdzamy skorlowanie zmiennych numerycznych



Rysunek 23. Na heatmapie widzimy korelację zmiennych numerycznych. Na podstawie tego możemy odrzucić zmienne silnie skorelowane. Z pary zmiennych cons_price_idx oraz euribor_3m odrzucamy con_price_idx, emp_var_rate oraz euribor_3m odrzucamy euribor_3m.

Kolejne kroki:

1. Usuwamy kolumnę marital ponieważ wykres korelacji ze zmienną celu nie pokazuje jasnej zależności.

2. Nazwę zmiennej ‘pdays’ zmieniamy na ‘prev_contact’ a wartości zamieniamy na 1 w przypadku ‘yes’ i 0 dla ‘no’ .

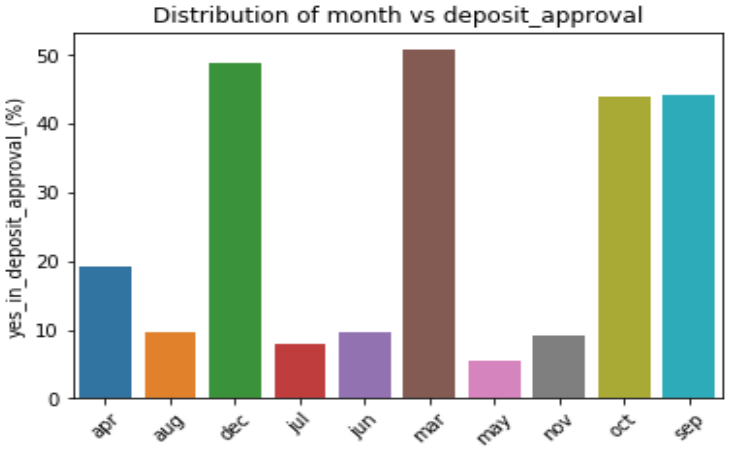
3. Wartości dla zmiennej ‘contact’ również zmieniamy na 1 dla wartości cellular i 0 dla wartości telephone.

4. Zmienną celu zamieniamy na 1 w przypadku wartości yes i 0 dla no.

5. Zmienne kategoryczne 'month', 'poutcome', 'job', 'education' kodujemy binarnie za pomocą pd.get_dummies

Stworzona ramka danych zawiera aż 40 kategorii dlatego na podstawie wykresów korelacji zmiennych kategorycznych ze zmienną celu stworzymy ramkę danych z pierwotną liczbą kategorii. Zmienne kategoryczne zakodujemy według wykresów korelacji zmiennych opisowych ze zmienną celu:

Patrząc na wykres korelacji zmiennej month względem zmiennej celu, jako 1 oznaczymy misiące apr, dec, mar, oct oraz sep, a 0 pozostałe miesiące



W przypadku zmiennej poutcome jako 1 oznaczymy rekordy z wartością ‘succes’, jako 0 pozostałe rekordy. Na podstawie wykresu zminnej ‘job’ w stosunku do zmennej deposit approval jako 1 oznaczymy wartości ‘admin’, ‘retired’, ‘student’, oraz ‘unemployed’, jako 0 pozostałe kategorie. Zmienną ‘education’ zakodujemy jako 1 dla wykształcenia ‘high school’ i wyższego, oraz 0 dla wykształcenia poniżej ‘high school’.

6. Zastosowanie danych w modelach.

6.1 Drzewo decyzyjne.

W pierwszej kolejności zastosujemy model drzewa decyzyjnego dla danych gdzie stosowaliśmy „OneHotEncoding”, czyli danych z 40 kategoriami.

Dzielimy model na model uczący i model testowy w stosunku 80% do 20%. Po podstawieniu do modelu sprawdzamy metryki modelu

```
pd.Series(fit_classifier(my_tree, X_ucz_tree, X_test_tree, y_ucz_tree, y_test_tree))
Out[85]:
ACC     0.890912
P       0.455764
R       0.450331
F1      0.453031
```

Podstawimy do modelu dane, w których wartości kategoryczne kodowaliśmy jako 0 i 1 według korelacji zmiennych objaśniających i zmiennej celu.

```
pd.Series(fit_classifier(my_tree_bin, X_ucz_tree_bin, X_test_tree_bin, y_ucz_tree_bin, y_test_tree_bin))
Out[96]:
ACC     0.898352
P       0.506702
R       0.487742
F1      0.497041
```

W przypadku drzewa decyzyjnego lepsze metryki ma model, w którym używalismu ramki danych, w której kodowaliśmy wartości na podstawie wykresów korelacji poszczrgólnych zmiennych ze zmienną celu.

6.3 **Model regresji liniowej.**

Do modelu regresji liniowej używamy ramki danych, w której kodowaliśmy wartości na podstawie wykresów korelacji poszczrgólnych zmiennych ze zmienną celu – *'bankData_binary_cat.csv'.* Aby zastosować model regresji liniowej standaryzujemy dane, które były numeryczne. Dane kategoryczne, którym nadaliśmy wartości 0 i 1 pozostawiamy.

Wyliczamy współczynniki regresji

```
pd.Series(reglinear_std_all.coef_, index=X_std.columns.to_list()).round(4).sort_values(ascending=False)
Out[57]:
duration          0.3508
prev_contact      0.1733
month             0.1005
cons_conf_idx     0.0969
poutcome          0.0763
job               0.0395
contact           0.0357
education         0.0213
age               0.0050
campaign          0.0005
previous         -0.0244
emp_var_rate     -0.1995
```

Wartości współczynników regresji pokazują, które zmienne w najlepszy sposób objaśniają model. Z wyliczeń wynika, że 4 pierwsze zmienne czyli 'prev_contact' 'mont', 'cons_conf_idx' objaśniają w ponad 70% zmienną celu

Podstawowe metryki modelu

	r_score_u	r_score_t	MSE_u	MSE_t	MAE_u	MAE_t
Reg. liniowa	0.31316	0.322058	0.063753	0.060539	0.153222	0.14963