Politechnika Gdańska

Wydział Fizyki Technicznej i Matematyki Stosowanej

Studia podyplomowe

Kierunek – Inżynieria Danych – Data Science

Metody statystyczne i analityczne Big Data

Sławomir Lisowski

Określenie wystąpienia świadczenia ubezpieczeniowego z danej polisy ubezpieczenia komunikacyjnego – wybór modelu predykcyjnego

1. Wstęp i cele

Praca ma na celu wybór najbardziej odpowiedniego modelu, który przewidzi, czy z danej polisy ubezpieczenia komuni kacyjnego wystąpi roszczenie czy też nie. Na podstawie danych trzeba znaleźć te, od których w największym stopni zależy prawd opodobieństwo roszczenia z danej polisy i będą wykorzystywane w fazie wyboru modelu.

2. Opis danych i zrozumienie danych

Zbiór danych, który będzie służył do badań zawiera 10296 obserwacji. Pojedynczy rekord zawiera informacje na tema t klienta, który zawarł umowę ubezpieczeniową odpowiedzialności cywilnej z firmą ubezpieczeniową. Każdy rekord w tym wypad ku polisa jest opisany poprzez zestaw 33 zmiennych:

1. ID – numer ID (integer)
2. KIDSDRIV – liczba dzieci użytkujących pojazd (integer)
3. PLCYDATE – data zawarcia aktualnej polisy (Date)
4. TRAVTIME – dystans pokonywany w drodze do pracy (integer)
5. CAR_USE – sposób użytkowania samochodu: „Private”, „Commercial” (factor)
6. POLICYNO – numer polisy (character)
7. BLUEBOOK – wartość samochodu (integer)
8. INITDATE – data pierwszej polisy (Date)
9. RETAINED – ilość lat jako klient (integer)
10. NPOLICY – liczba polis (integer)
11. CAR_TYPE – typ samochodu: „Panel Truck”, „Pickup”, „Sedan”, „Sports Car”, „SUV”, „Van” (factor)
12. RED_CAR – informacja czy kolor samochodu to czerwony: „Yes”, „No”(factor)
13. OLDCLAIM – suma roszczeń z polis w poprzednich latach (integer)
14. CLM_FREQ – liczba roszczeń z polis w ostatnich pięciu latach (integer)
15. REVOKED – prawo jazdy było uzyskane w ostatnich 7 latach: „Yes”, „No”(factor)
16. MVR_PTS – liczba wykroczeń drogowych (integer)
17. CLM_AMT – suma roszczeń z obecnej polisy (integer)
18. CLM_DATE – data roszczenia z obecnej polisy (integer)
19. CLM_FLAG – czy roszczenie wystąpiło: „Yes”, „No”(factor)
20. AGE – wiek kierowcy (integer)
21. AGE*GENDER – wiek kierowcy łącznie z informacją o płci (integer)
22. HOMEKIDS – liczba dzieci w gospodarstwie domowym (integer)
23. YOJ – liczba lat zatrudnienia w obecnej pracy (integer)
24. INCOME – roczny przychód (integer) GENDER – płeć kierowcy: „F”, „M” (factor)
25. MARRIED – czy kierowca jest w związku małżeńskim: „Yes”, „No” (factor)
26. PARENT1 – czy kierowca samotnie wychowuje dziecko: „Yes”, „No”(factor)
27. JOBCLASS – wykonywany zawód: : ”Unknown”, ”Blue Collar”, ”Clerical”, ”Doctor”,”Home Maker”, ”Lawyer”, ”Manager ”, ”Professional”, ”Student”(factor)
28. MAX_EDUC – poziom edukacji (factor)
29. HOME_VAL – wartość ubezpieczonego domu (integer)
30. SAMEHOME – lata zamieszkania pod obecnym adresem (integer)
31. DENSITY – miejsce zamieszkania: „Highly Rural”, „Highly Urban”, „Rural”, „Urban” (factor)
32. YEARQTR – factor

Zmienna celu jest oznaczona kolumną CLM_FLAG i ma wartości „Yes” lub „No”, które określają czy świadczenie z danej polisy wy stąpiło czy też nie.

2.1 Charakterystyka zmiennych.

Zmienne w tym zbiorze danych mają róży charakter. Występują tu zmienne o charakterze:

- binarnym,
- zmienne ciągłe,
- zmienne dyskretne,
- zmienne jakościowe

Tabela przedstawia ilość unikalnych wartości poszczególnych zmiennych oraz ich rozkład

.

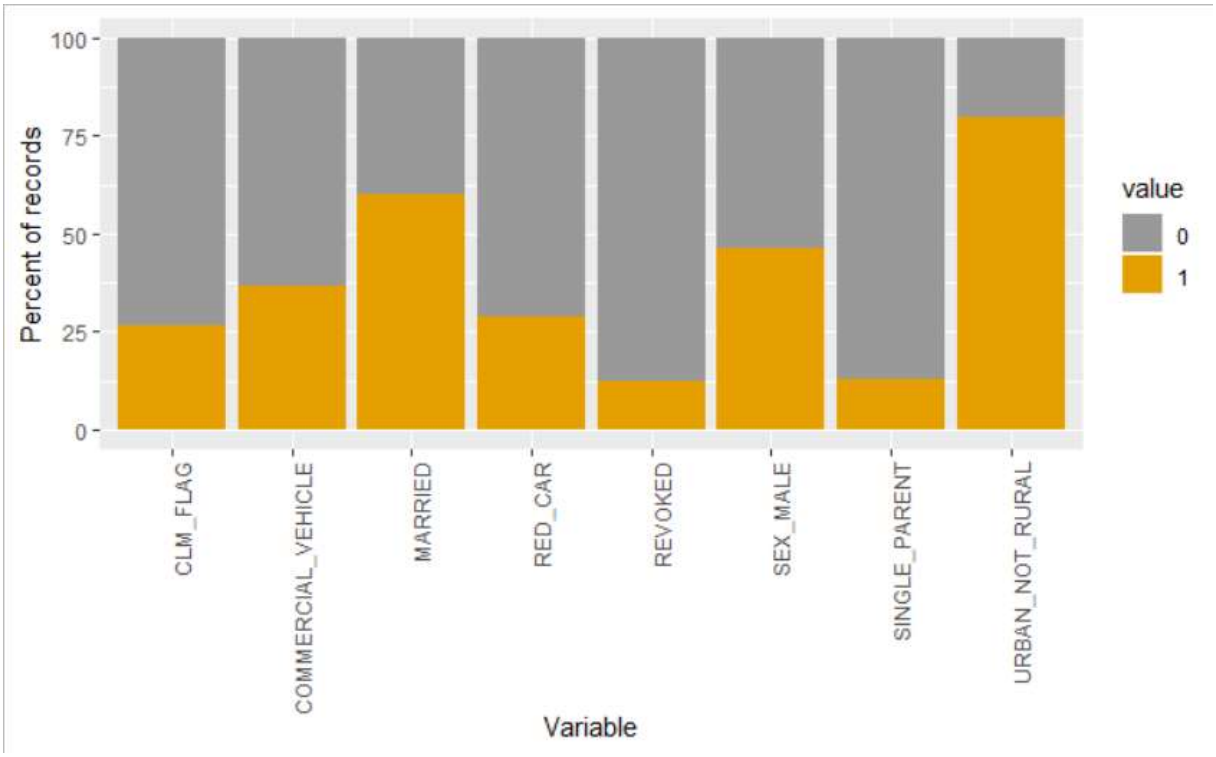
key	uniques	Rodzaj zminnej
CAR_USE	2	binarna
CLM_FLAG	2	binarna
GENDER	2	binarna
MARRIED	2	binarna
PARENT1	2	binarna
RED_CAR	2	binarna
REVOKED	2	binarna
AGE	4	jakościowa
DENSITY	4	jakościowa
KIDSDRIV	5	dyskretna
MAX_EDUC	5	jakościowa
CAR_TYPE	6	jakościowa
CLM_FREQ	6	dyskretna
HOMEKIDS	6	dyskretna
AGE.GENDER	8	jakościowa
NPOLICY	8	dyskretna
JOBCLASS	9	jakościowa
MVR_PTS	14	dyskretna
YOJ	21	ciągła
RETAINED	23	ciągła
SAMEHOME	30	ciągła
AGE_YEARS	62	ciągła
TRAVTIME	100	ciągła
CLM_AMT	2341	ciągła
BLUEBOOK	2983	ciągła
OLDCLAIM	3542	ciągła
HOME_VAL	6335	ciągła
INCOME	8150	ciągła

Wartości NA, które występowały w kolumnach YOJ, INCOME, HOME_VAL, SAME_HOME zostały zastąpione wartościami średnimi z tych kolumn. Do zbioru danych została również kolumna mówiąca o dokładnym wieku AGE_YEARS. Powstała z różnicy między kolumnami PLCYDATE, BIRTH. Daty były napisane w języku francuskim. Po konwersji zostały zamienione na typ Date.

2.2 Rozkłady zmiennych losowych.

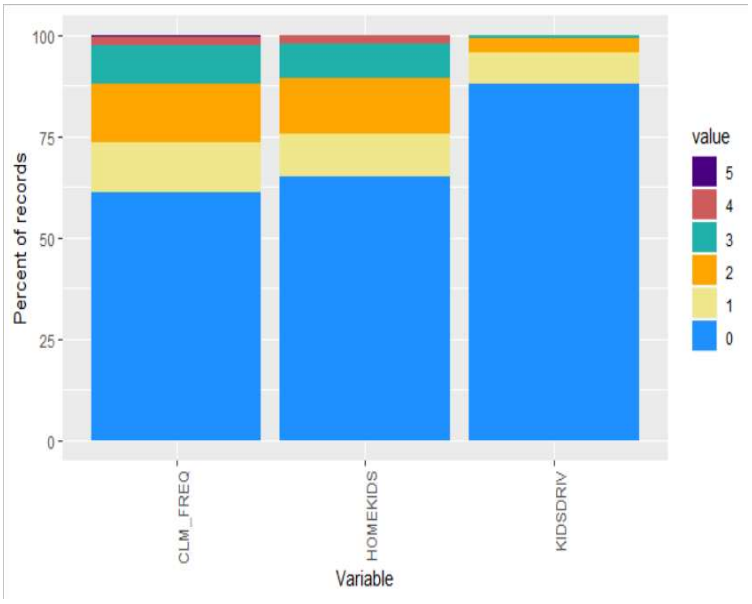
Z tego względu, że mamy do czynienia z różnego rodzaju zmiennymi losowymi ich rozkłady również będą kształtować się w inny sposób. Zmienna w kolumnie DENSITY została zamieniona na wartość binarną , 1 – w przypadku wartości „Highly Urban" i „Urban", 0 – w przypadku wartości „Highly Rural", „Rural".

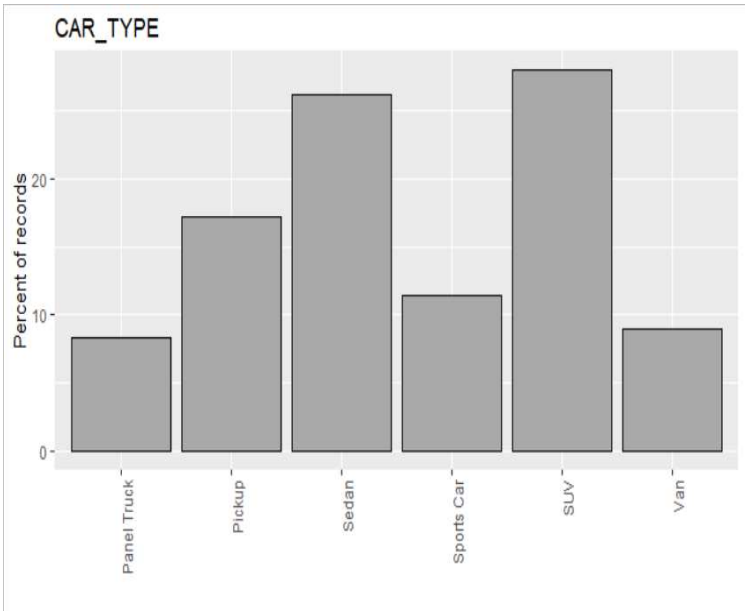
Rozkład zmiennych binarnych



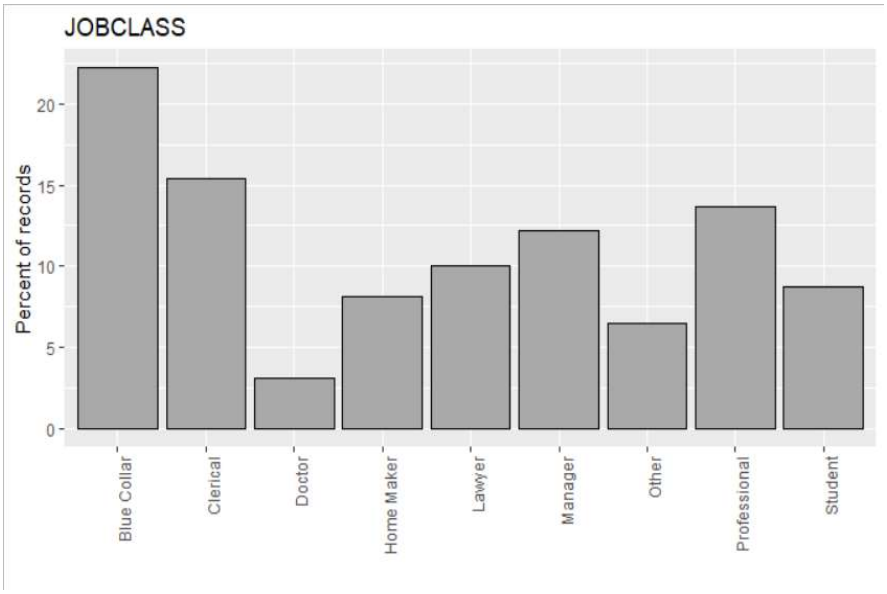
Z rozkładu możemy wywnioskować, że znacząca część właścicieli polis zamieszkiwała tereny zurbanizowane lub wysoko zurbanizowane. W przypadku wszystkich polis z ok 25% polis wystąpiły roszczenia. W ok 10% osoby posiadające polisy samotnie wychowywały dziecko, również ok 10% właścicieli polis

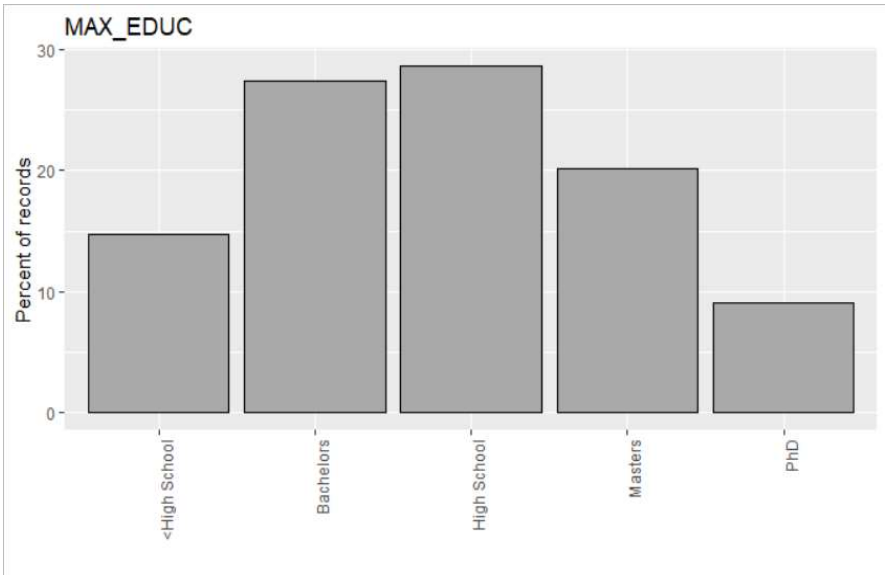
Rozkład zmiennych dyskretnych i faktorów





Biorąc pod uwagę rozkład zmiennych dyskretnych CLM_FREQ, HOMEKIDS oraz KIDSDRIV, możemy je zamienić na zmienne binarne. Po tej zmianie nowe nazwy zmiennych to: Past_claim (1 jeżeli w przeszłości wystąpiło roszczenie), Driving_kids (1 jeżeli samochód prowadzili nastolatkowie), Kids (1 jeżeli właściciel polisy posiadał dzieci)

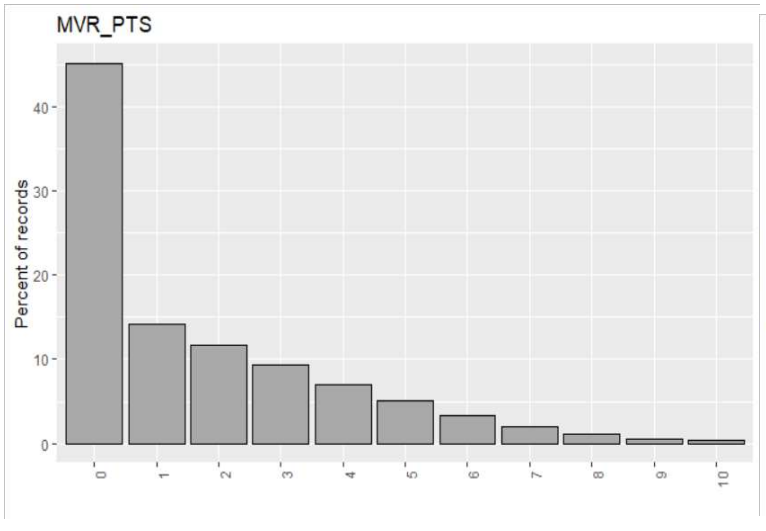


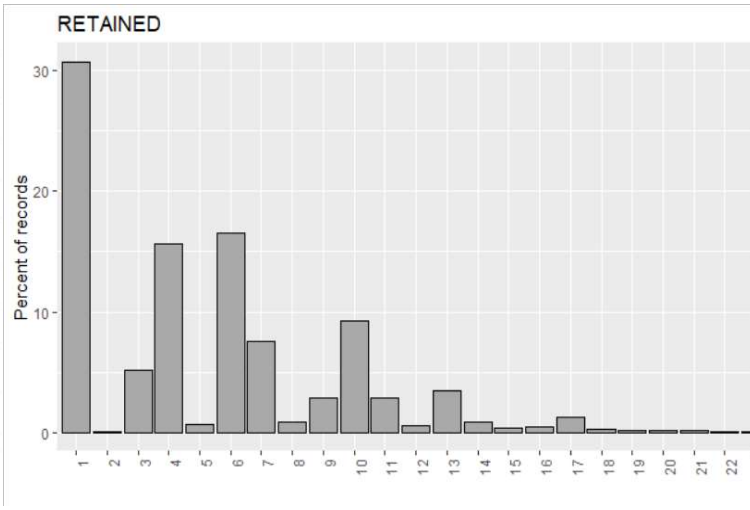


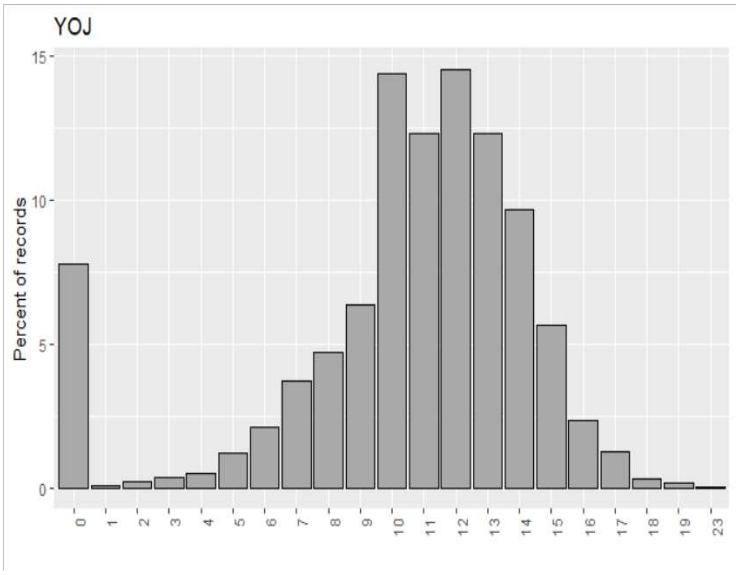
Przedstawione rozkłady faktorów, pokazują jak kształtowały się zmienne jakościowe w badanym zbiorze danych. Dane są pokazane w wartościach procentowych. Rozkład zmiennej CAR_TYPE opisuje jaki procent ogólnej liczby ubezpieczonych pojazdów stanowiły poszczególne typy samochodów. Możemy zauważyć znaczną przewagę samochodów typu Sedan oraz SUV, które w sumie mają 50% udział w rozkładzie tej zmiennej. W zmiennej JOBCLASS, mówiącej o statusie zawodowym właścicieli ubezpieczonych samochodów, ponad 20% stanowią pracownicy fizyczni oraz administracyjni niższego szczebla.

Zmienna jakościowa MAX_EDUC opisuje stopień edukacji jaki został zdobyty prze właścicieli polis. Tutaj ok 30% to osoby z ukończoną szkołą średnią

Rozkład zmiennych dyskretnych

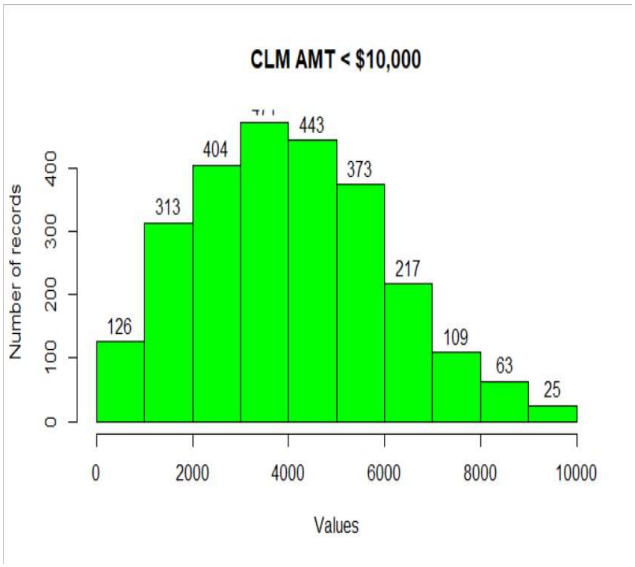


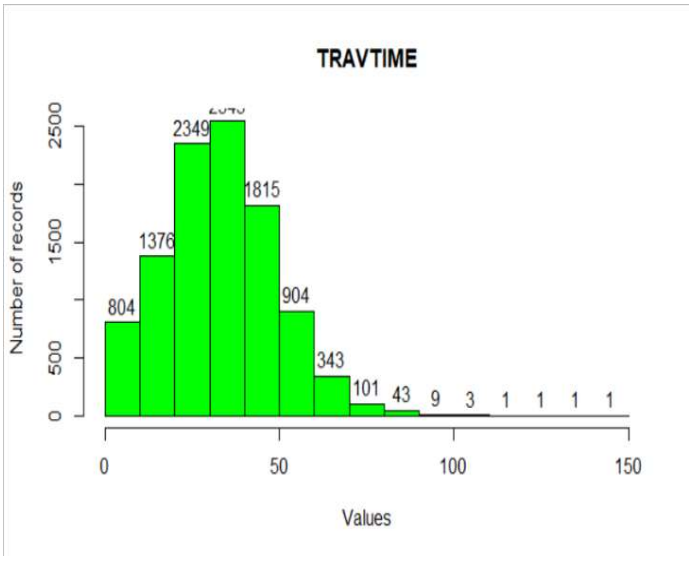


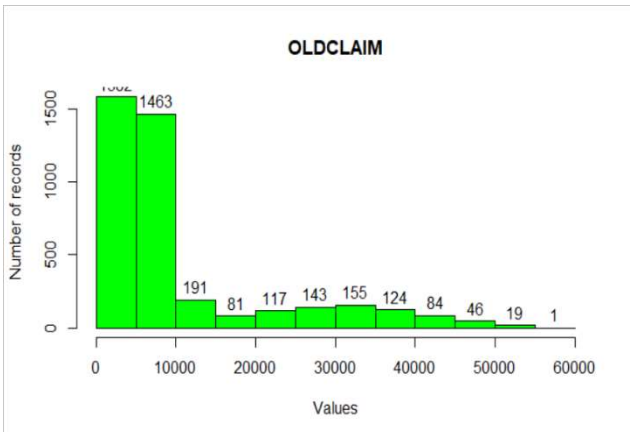


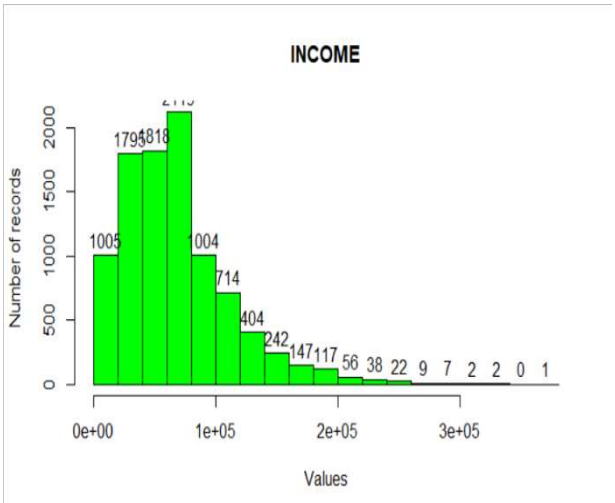
Rozkład zmiennej MVR_PTS, która mówi o liczbie punktów za wykroczenia drogowe. W ponad 40% obserwacji właściciele polis nie mieli punktów za takie wykroczenia. MVR_PTS ograniczamy do max 10 punktów ponieważ udział osób, które miały więcej punktów jest bliski 0. Rozklad zmiennej RETAINED pokazuje, że większość klientów to osoby, które były stałymi klientami maksymalnie 6 lat. W przypadku zmiennej YOJ obserwujemy, że jest spory odsetek osób zaczynającyhch pracę jednak przeważają osoby z min. 10 letnim stażem

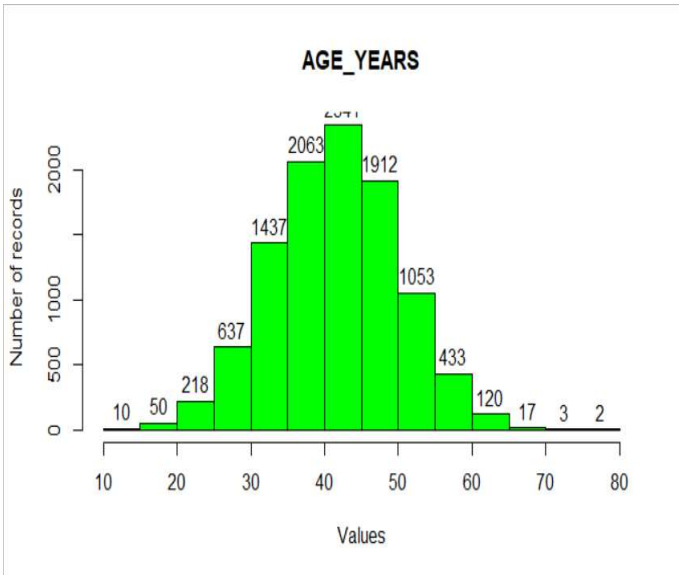
Rozkład zmiennych ciągłych



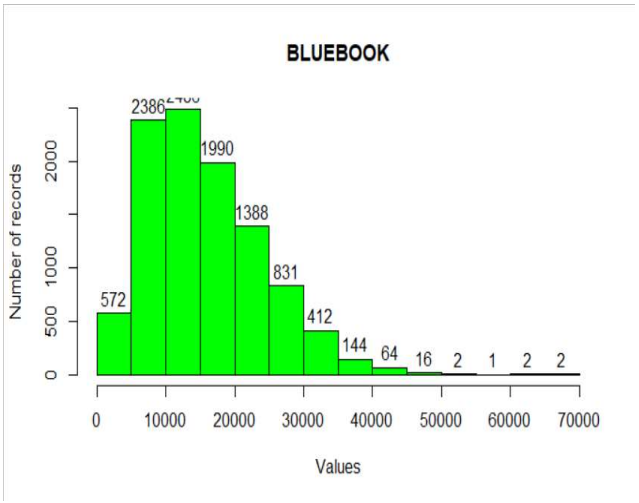


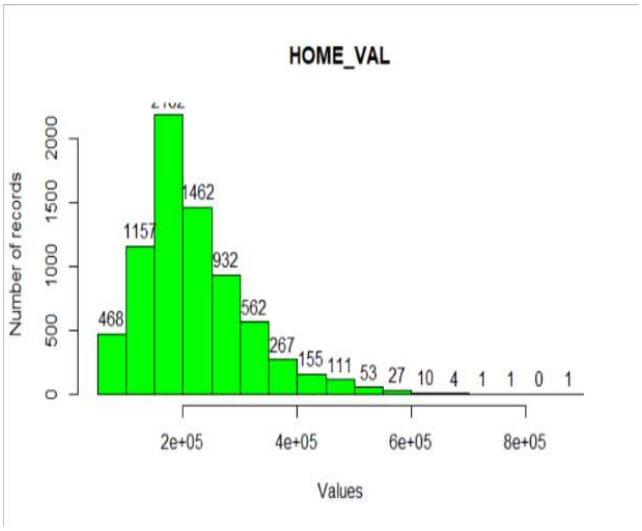






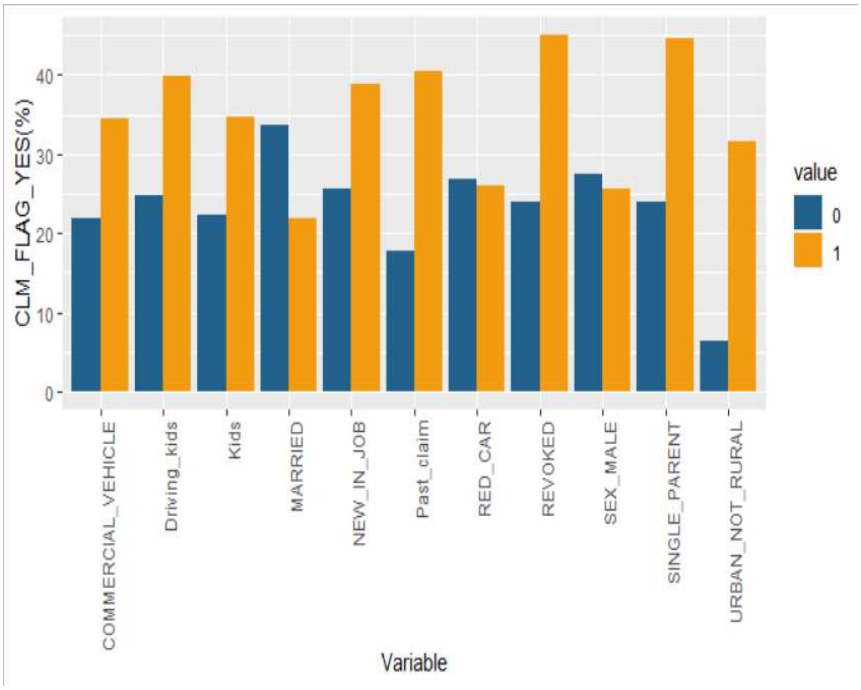
Większość zmiennych ciągłych ma rozkład normalny. Rozkłady zmiennych OLDCLAIM, INCOME, HOME_VAL, BLUEBOOK są rozkładami o asymetrii prawostronnej. Dzieje się tak dlatego, że są to wartości wyrażone w dolarach. Aby pozbyć się tej asymetrii możemy dokonać transformacji logarytmicznej tych zmiennych. Rozkład zmiennej TRAVTIME ma również asymetrię prawostronną. Zmienna OLDCLAIM pokazuje jak kształtowała się kwota roszczeń z polis w poprzednich latach





3. Korelacja zmiennych losowych ze zmienną celu CLM_FLAG.

3.1 Korelacja zmiennych binarnych ze zmienną celu.



Zmienne RED_CAR oraz SEX_MALE mają podobny % udział w powodowanych roszczeniach dlatego zmienne te odrzucamy . Z wykresu wynika że osoby które jeżdżą samochodami firmowymi częściej powodują wypadki. Roszczenia występowały również częściej jeżeli:

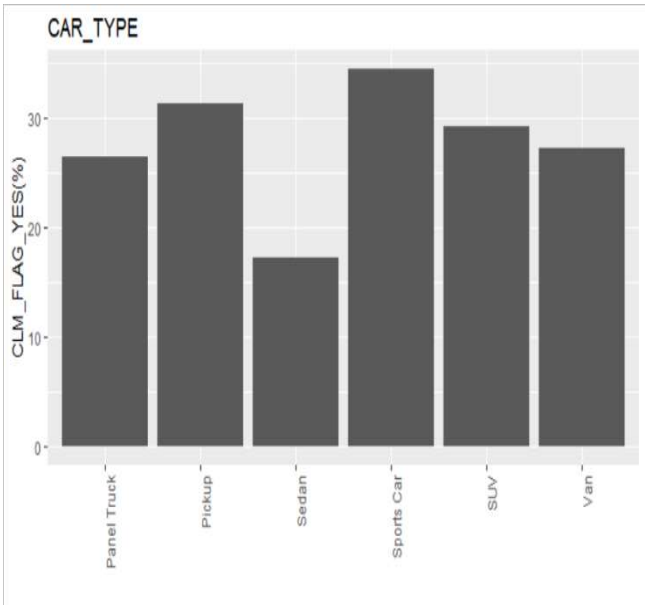
- samochód prowadzili nastolatkowie,

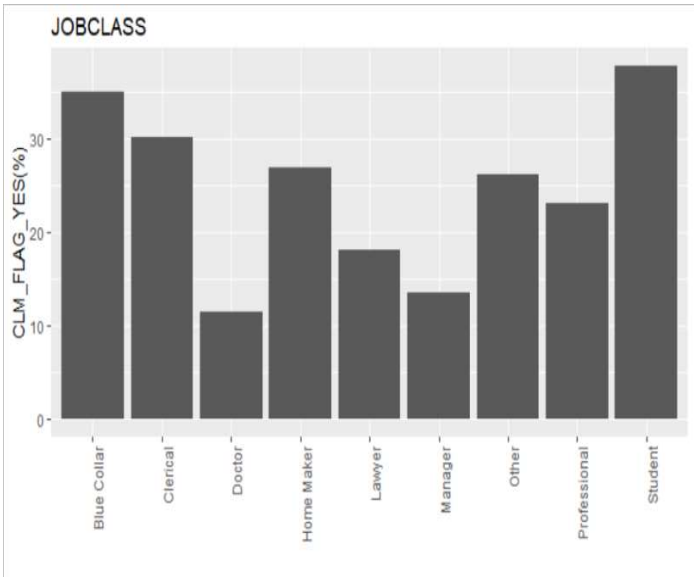
- osoby miały roszczenia z polisy w poprzednich latach,

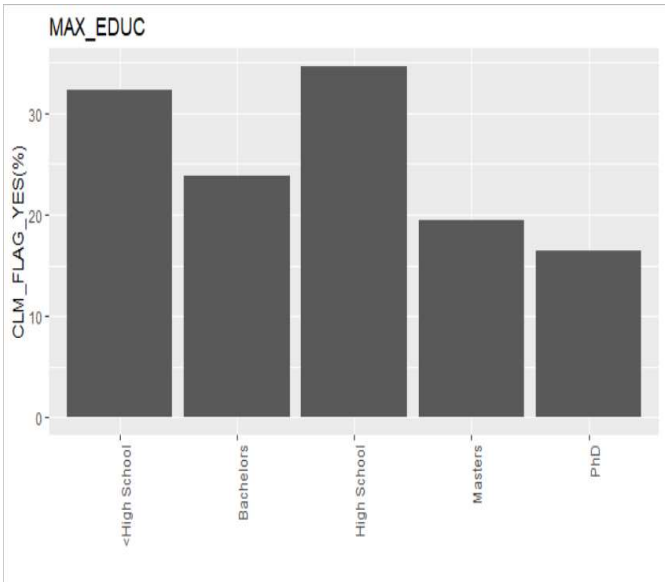
- osoby uzyskały prawo jazdy w ostatnich 7 latach,

- osoby używające pojazdów w terenach zurbanizowanych

3.2 Korelacja zmiennych jakościowych ze zmienną celu.







Na wykresach widzimy jakie grupy w podziale na typ ubezpieczonego samochodu, wykonywany zawód oraz poziom wykształcenia najczęściej uzyskiwały roszczenia z polisy. Możemy z nich wyciągnąć następujące wnioski

- osoby, które uzyskały przynajmniej licencjat, rzadziej uzyskiwały roszczenia,

- użytkownicy samochodów typu sedan rzadziej powodują roszczenia

- osoby uczące się oraz pracownicy fizyczni i administracyjni niższego szczebla częściej powodują roszczenia

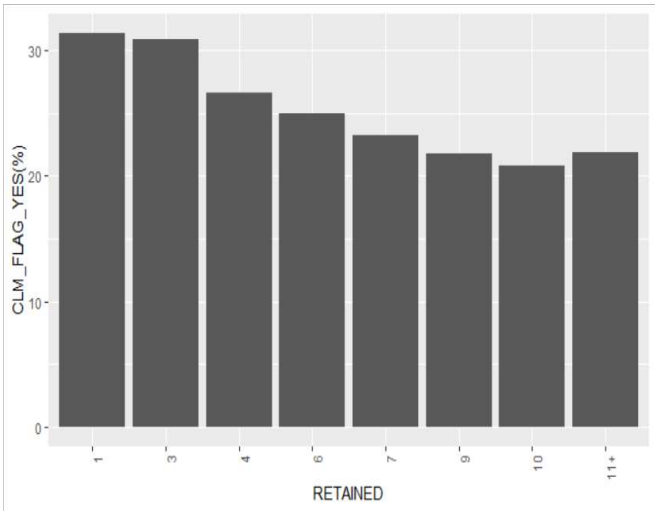
Biorąc pod uwagę te wnioski, z tych zmiennych tworzymy zmienne typu binarnego biorąc pod uwagę:

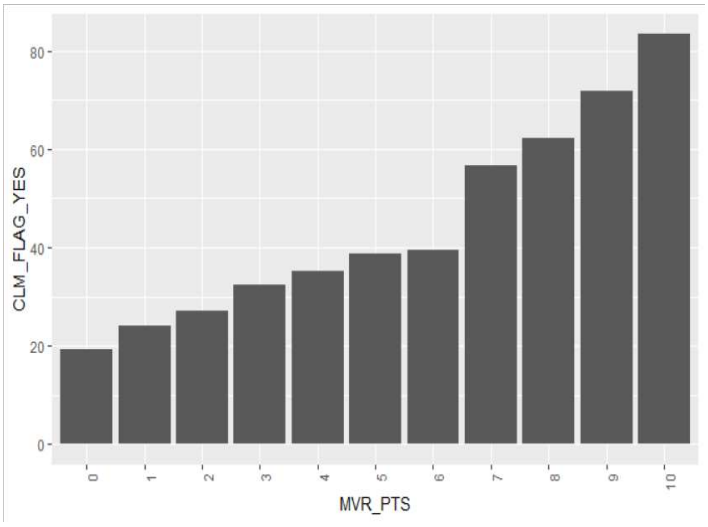
- czy osoba miała co najmniej licencjat(1) czy też nie(0), nowa zmienna College_MAX_EDUC

- czy osoba kierowała samochodem typu sedan(1) czy też nie(0) nowa zmienna Sedan

- czy osoba nie należała do grupy uczących się lub pracowników fizycznych lub administracyjnych niższego szczebla(1), czy należała do taj grupy(0), nowa zmienna Blue_collar

3.3 Korelacja zmiennych numerycznych ze zmienną celu.



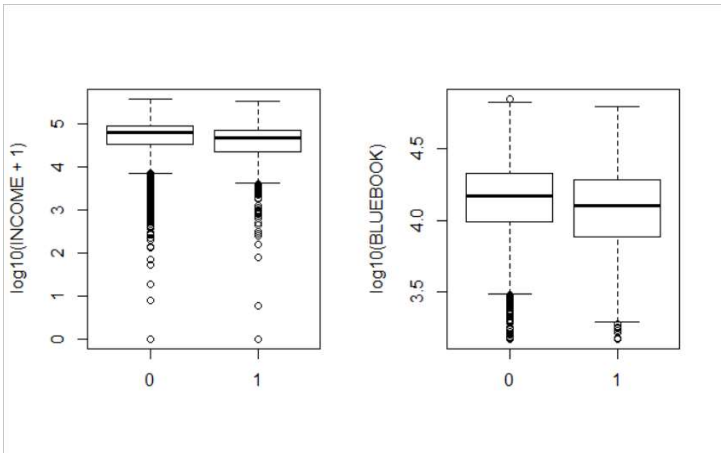


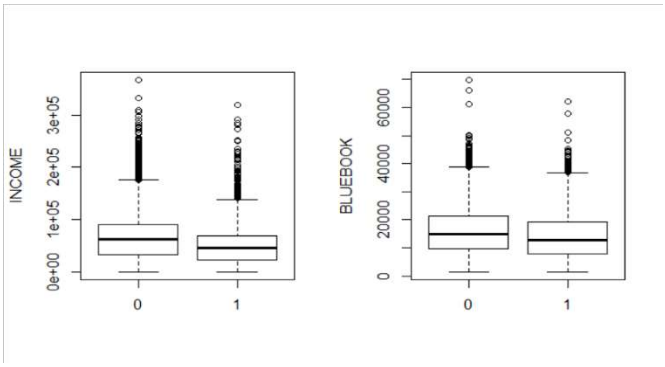
Z wykresu wynika, że im dłużej klient korzystał z usług ubezpieczeń komunikacyjnych tym mniejsza jest jego skłonność do powodowania roszczeń. Zależność ta jednak nie jest bardzo silna. W większości przypadków niezależnie ile lat klient miał polisy i tak roszczenia były powodowane w przedziale od 20 do 30%. W związku z tym rezygnujemy z ujmowania tej zmiennej w modelu.

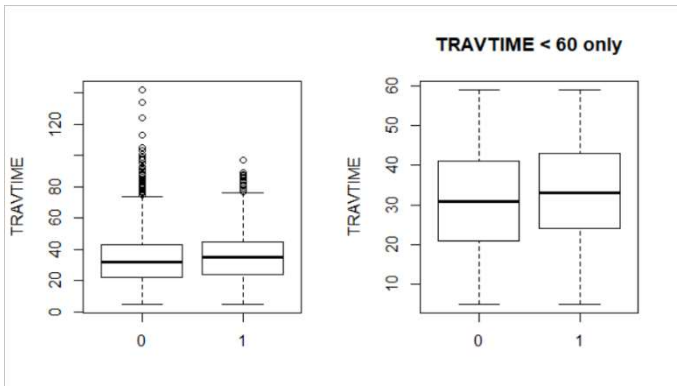
W przypadku zmiennych, MVR_PTS widzimy silną korelację. Jeżeli chodzi o zmienną MVR_PTS to jest ona silnie dodatnio skorelowana ze zmienną celu. Sprawdzamy również korelację zmiennej AGE_YEARS do CLM_FLAG i wynika z niej:

```
[1] "Procent roszczeń dla grupy Młodzi (<25): 56.04"
[1] "Procent roszczeń dla grupy Pomiędzy (25-64): 26"
[1] "Procent roszczeń dla grupy Seniorzy (65+): 29.41"
```

Z tych proporcji wynika, że najczęściej roszczenia powodują osoby młode, dlatego też zmienną AGE_YEARS zamieniamy na binarną 1 – dla osób poniżej 25 lat, 0 – dla osób powyżej tego wieku i tworzymy nową zmienną Young_age.



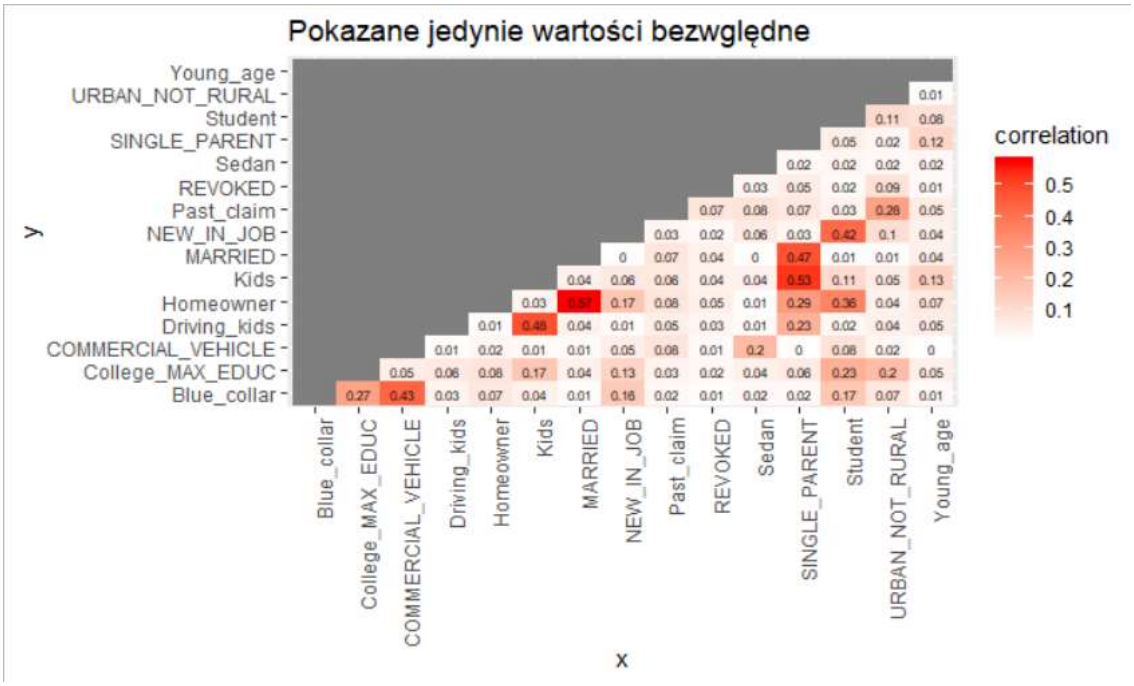




Z wykresów typu boxplot wynika, że osoby podróżujące częściej, powodują roszczenia natomiast osoby z większymi zarobkami i droższymi samochodami robią to rzadziej.

4. Korelacje między zmiennymi.

4.1 Korelacja między zmiennymi binarnymi.



W większości przypadków korelacje między zmiennymi binarnymi są słabe, ale kilka się wyróżnia:

- korelacja pomiędzy statusem zatrudnienia - Blue_collar a użytkowaniem samochodu – Commercial_Vehicle

- korelacja pomiędzy zmiennymi HOMEOWNER a MARRIED

- korelacja pomiędzy NEW_IN_JOB oraz STUDENT

- pomiędzy Driving_kids oraz kids

- pomiędzy SINGLE_PARENT a MARRIED

W związku z tym nie będziemy brać pod uwagę zmiennych Blue Collar, oraz Student w dalszych rozważaniach

Natomiast ze zmiennych HOMEOWNER, MARRIED, Driving_kids oraz kids utworzymy jedną zmienną abstrakcyjną

- „marriage_home_and_kids_score”:

1. 8 punktów –właściciel domu w związku małżeńskim
2. 7 punktów – właścicielem domu, nie w związku małżeńskim bez dzieci
3. 6 punktów – w związku małżeńskim nie właściciel domu
4. 5 punktów – nie w związku, bez dzieci, nie jest właścicielem domu
5. 4 punktów – właściciel domu, samotnie wychowujący dziecko
6. 0 punktów – samotnie wychowujący dziecko, bez własności domu

Korelacja między zmiennymi numerycznymi

```
Korelacja INCOME i BLUEBOOK: 0.42
Korelacja INCOME i TRAVTIME: -0.05
Korelacja TRAVTIME i BLUEBOOK: -0.02
```

Widzimy korelację między zarobkami a wartością samochodu dlatego nie będziemy brać pod uwagę jednej z tych wartości. Rezygnujemy z wartości samochodu

Korelacja między zarobkami (INCOME) a zmiennymi binarnymi

	Variable	Correlation
8	NEW_IN_JOB	-0.36
13	Student	-0.34
6	Kids	-0.15
15	Young_age	-0.09
9	Past_claim	-0.06
12	SINGLE_PARENT	-0.06
4	Driving_kids	-0.04
7	MARRIED	-0.04
1	Blue_collar	-0.03
10	REVOKED	-0.02
11	Sedan	0.04
3	COMMERCIAL_VEHICLE	0.09
5	Homeowner	0.11
14	URBAN_NOT_RURAL	0.19
2	College_MAX_EDUC	0.49

Korelacja między czasem podróżowania (TRAVTIME) a zmiennymi binarnymi

	Variable	Correlation
14	URBAN_NOT_RURAL	-0.17
2	College_MAX_EDUC	-0.04
8	NEW_IN_JOB	-0.02
5	Homeowner	-0.01
11	Sedan	-0.01
12	SINGLE_PARENT	-0.01
6	Kids	0.00
9	Past_claim	0.00
10	REVOKED	0.00
15	Young_age	0.00
3	COMMERCIAL_VEHICLE	0.01
4	Driving_kids	0.01
7	MARRIED	0.01
13	Student	0.02
1	Blue_collar	0.03

Korelacja zmiennych binarnych z sumą roszczenia (CLM_AMT)

	Variable	Correlation
5	Homeowner	-0.08
7	MARRIED	-0.08
11	Sedan	-0.07
2	College_MAX_EDUC	-0.06
8	NEW_IN_JOB	0.03
13	Student	0.03
15	Young_age	0.05
1	Blue_collar	0.06
4	Driving_kids	0.07
6	Kids	0.07
10	REVOKED	0.07
3	COMMERCIAL_VEHICLE	0.10
12	SINGLE_PARENT	0.10
14	URBAN_NOT_RURAL	0.12
9	Past_claim	0.14

Ostatecznie po wszystkich transformacjach do użycia w modelach wybraliśmy zmienne:

1. Driving_kids
2. TRAVTIME
3. Past_claim
4. REVOKED
5. MVR_PTS
6. NEW_IN_JOB
7. Young_age
8. College_MAX_EDUC
9. marriage_kids_and_home_score – nowo utworzona zmienna abstrakcyjna z połączenia MARRIED, SINGLE_PARENT, HOMEOWNER, KIDS
10. Sedan
11. URBAN_NOT_RURAL

5. Wybrane modele.

Do dalszych obliczeń wybraliśmy model „Regresji Poissona”,„Regresji dwumianowej” oraz „Regresji ujemnej dwumianowy” będziemy oceniać na podstawie kryterium AIC. Zbiór podzieliliśmy na zbiór treningowy, który zawiera 6999 rekordów. Zbiór walidacyjny zawiera pozostałe 3296 rekordów.

5.1 Regresja Poissona

```
Call:
glm(formula = CLM_FLAG ~ Driving_kids + TRAVTIME + Past_claim +
    REVOKED + MVR_PTS + NEW_IN_JOB + Young_age + College_MAX_EDUC +
    marriage_home_and_kids_score + Sedan + URBAN_NOT_RURAL, family = "poisson",
    data = model_data_train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.13262  -0.68098  -0.49078   0.09423   2.41199

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -2.592467   0.133135 -19.472  < 2e-16 ***
Driving_kids                0.361676   0.062830   5.756 8.59e-09 ***
TRAVTIME                    0.010282   0.001462   7.034 2.00e-12 ***
Past_claim                  0.334813   0.052867   6.333 2.40e-10 ***
REVOKED                     0.399631   0.057400   6.962 3.35e-12 ***
MVR_PTS                     0.052017   0.010302   5.049 4.44e-07 ***
NEW_IN_JOB                  0.268116   0.073351   3.655 0.000257 ***
Young_age                   0.383232   0.116394   3.293 0.000993 ***
College_MAX_EDUC           -0.547526   0.047637 -11.494  < 2e-16 ***
marriage_home_and_kids_score -0.079911   0.008457  -9.449  < 2e-16 ***
Sedan                      -0.468970   0.062168  -7.544 4.57e-14 ***
URBAN_NOT_RURAL             1.532659   0.107448  14.264  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4934.9  on 6998  degrees of freedom
Residual deviance: 3877.3  on 6987  degrees of freedom
AIC: 7637.3
```

```
                             Overall
Driving_kids                5.756405
TRAVTIME                    7.034431
Past_claim                  6.333054
REVOKED                     6.962208
MVR_PTS                     5.049016
NEW_IN_JOB                  3.655244
Young_age                   3.292556
College_MAX_EDUC           11.493814
marriage_home_and_kids_score  9.449330
Sedan                       7.543592
URBAN_NOT_RURAL            14.264174
```

Z modelu wynika, że wszystkie przyjęte predyktory są znaczące. Potwierdza również korelacje, które zostały zauważone wcześniej. Za pomocą funkcji varImp() z pakietu caret sprawdzamy również jak znaczące są wybrane predyktory.

5.2 Model dwumianowy

```
Call:
glm(formula = CLM_FLAG ~ Driving_kids + TRAVTIME + Past_claim +
    REVOKED + MVR_PTS + NEW_IN_JOB + Young_age + College_MAX_EDUC +
    marriage_home_and_kids_score + Sedan + URBAN_NOT_RURAL, family = binomial,
    data = model_data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4627  -0.7416  -0.4590   0.6691   3.1240

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -2.544512   0.161774 -15.729  < 2e-16 ***
Driving_kids                  0.691841   0.091047   7.599 2.99e-14 ***
TRAVTIME                      0.018031   0.001972   9.146  < 2e-16 ***
Past_claim                    0.487743   0.068312   7.140 9.34e-13 ***
REVOKED                       0.781000   0.084025   9.295  < 2e-16 ***
MVR_PTS                       0.112487   0.014989   7.504 6.17e-14 ***
NEW_IN_JOB                    0.550913   0.109431   5.034 4.80e-07 ***
Young_age                     0.831417   0.195120   4.261 2.03e-05 ***
College_MAX_EDUC             -0.922041   0.062987 -14.639  < 2e-16 ***
marriage_home_and_kids_score -0.162349   0.012810 -12.674  < 2e-16 ***
Sedan                        -0.737778   0.077158  -9.562  < 2e-16 ***
URBAN_NOT_RURAL               2.185165   0.122327  17.863  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8120.9  on 6998  degrees of freedom
Residual deviance: 6521.4  on 6987  degrees of freedom
AIC: 6545.4

Number of Fisher Scoring iterations: 5
```

```
                              Overall
Driving_kids                 7.598739
TRAVTIME                     9.145643
Past_claim                   7.139887
REVOKED                      9.294808
MVR_PTS                      7.504479
NEW_IN_JOB                   5.034326
Young_age                    4.261044
College_MAX_EDUC            14.638569
marriage_home_and_kids_score 12.673925
Sedan                        9.561867
URBAN_NOT_RURAL             17.863249
```

Biorąc pod uwagę kryterium AIC model ten wydaje się lepszy. Ponownie za pomocą funkcji varImp() z pakietu caret sprawdzamy, które zmienne są najbardziej zanczące. Potwierdza się, że najbardziej znaczącymi zmiennymi dla naszego modelu są zmienne URBAN_NOT_RURAL, College_MAX_EDUC oraz zmienna stworzona z kombinacji MARRIED, SINGLE_PARENT, KIDS i HOMEOWNER czyli marriage_home_and_kids_score

5.3 Model ujemny dwumianowy.

```
Call:
glm.nb(formula = CLM_FLAG ~ Driving_kids + TRAVTIME + Past_claim +
    REVOKED + MVR_PTS + NEW_IN_JOB + Young_age + College_MAX_EDUC +
    marriage_home_and_kids_score + Sedan + URBAN_NOT_RURAL, data = model_data_train,
    init.theta = 4615.872777, link = log)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.13241  -0.68097  -0.49077   0.09419   2.41196

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -2.592472   0.133139 -19.472  < 2e-16 ***
Driving_kids                  0.361686   0.062834   5.756 8.60e-09 ***
TRAVTIME                      0.010282   0.001462   7.034 2.00e-12 ***
Past_claim                    0.334822   0.052870   6.333 2.41e-10 ***
REVOKED                       0.399640   0.057404   6.962 3.36e-12 ***
MVR_PTS                       0.052019   0.010303   5.049 4.44e-07 ***
NEW_IN_JOB                    0.268134   0.073356   3.655 0.000257 ***
Young_age                     0.383239   0.116403   3.292 0.000994 ***
College_MAX_EDUC             -0.547535   0.047639 -11.494  < 2e-16 ***
marriage_home_and_kids_score -0.079914   0.008457  -9.449  < 2e-16 ***
Sedan                        -0.468975   0.062170  -7.543 4.58e-14 ***
URBAN_NOT_RURAL               1.532665   0.107450  14.264  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(4615.873) family taken to be 1)

    Null deviance: 4934.6  on 6998  degrees of freedom
Residual deviance: 3877.0  on 6987  degrees of freedom
AIC: 7639.5

Number of Fisher Scoring iterations: 1
```

```
                              Overall
Driving_kids                 5.756200
TRAVTIME                     7.034247
Past_claim                   6.332945
REVOKED                      6.961908
MVR_PTS                      5.048878
NEW_IN_JOB                   3.655220
Young_age                    3.292343
College_MAX_EDUC            11.493503
marriage_home_and_kids_score 9.449029
Sedan                        7.543411
URBAN_NOT_RURAL             14.264010
```

Regresja ujemna dwumianowa wydaje się być najgorszym z tych trzech modeli biorąc pod uwagę kryterium AIC. Regresja ta ma bardzo podobną charakterystykę rozkładu parametrów znaczących do Regresji Poissona

5.4 Porównanie modeli.

	AIC	BIC	loglik
model_1	7637.286	7719.528	-3806.643
model_2	6545.436	6627.679	-3260.718
model_3	7639.451	7728.547	-3806.726

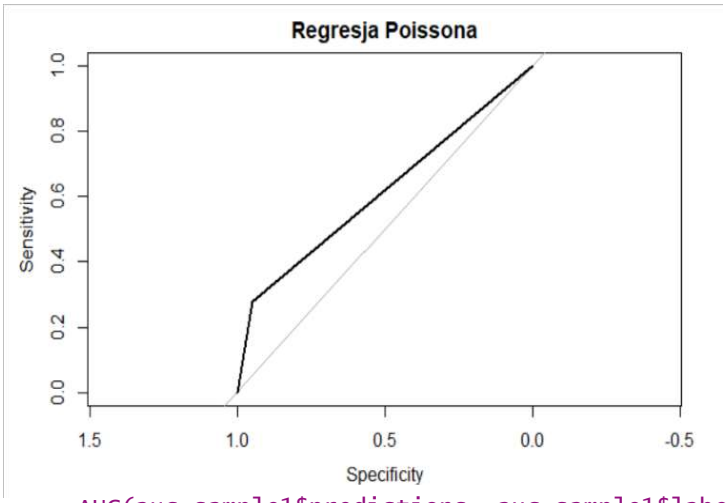
Biorąc pod uwagę kryterium AIC najlepszym modelem jest model dwumianowy. We wszystkich wypadkach wybrane predyktory były znaczące. Korzystając z biblioteki caret w pakiecie RStudio sprawdziliśmy, że najbardziej znaczącymi parametrami były URBAN_NOT_RURAL, College_MAX_EDUC oraz zmienna marriage_home_and_kids_score. Wybór ten potwierdza również kryterium BIC oraz logarytm wiarygodności.

6. Ewaluacja modeli.

Modele będą ewaluowane za pomocą funkcji confusionMatrix() z pakietu caret na zbiorze testowym 3297 obserwacji. Poza tym wygenerujemy krzywą ROC oraz obliczymy AUC (Area Under Curve). Pole powierzchni pod krzywą ROC. Im większa jest ta wartość tym lepiej

6.1 Regresja Poissona.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2304  628
         1  121  244

               Accuracy : 0.7728
                 95% CI : (0.7581, 0.787)
    No Information Rate : 0.7355
    P-Value [Acc > NIR] : 4.5e-07

                  Kappa : 0.2825

 Mcnemar's Test P-Value : < 2e-16

            Sensitivity : 0.27982
            Specificity : 0.95010
         Pos Pred Value : 0.66849
         Neg Pred Value : 0.78581
             Prevalence : 0.26448
         Detection Rate : 0.07401
   Detection Prevalence : 0.11071
      Balanced Accuracy : 0.61496

       'Positive' Class : 1
```

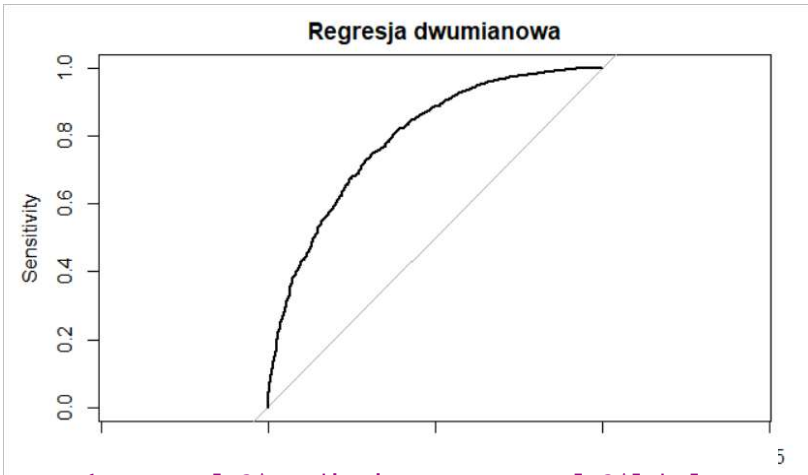



```
AUC(auc_sample1$predictions, auc_sample1$labels,
label.ordering = NULL)
[1] 0.7905431
```

Po ewaluacji modelu możemy spojrzeć na macierz pomyłek oraz podstawowe statystyki dla tego modelu. Wygenerowaliśmy również krzywą ROC oraz za pomocą funkcji AUC z biblioteki cvAUC w pakiecie R obliczyliśmy pole pod powierzchnią krzywej. Skuteczność modelu(Accuracy) wynosi 0,77. Jest to jedna z najważniejszych miar skuteczności modelu mówi o liczbie prawidłowych predykcji w stosunku do całego zbioru danych

6.2 Regresja dwumianowa

```
          Reference
Prediction    0    1
         0 2265  568
         1  160  304

               Accuracy : 0.7792
                 95% CI : (0.7646, 0.7933)
    No Information Rate : 0.7355
    P-Value [Acc > NIR] : 3.91e-09

                  Kappa : 0.3325

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.34862
            Specificity : 0.93402
         Pos Pred Value : 0.65517
         Neg Pred Value : 0.79951
             Prevalence : 0.26448
         Detection Rate : 0.09221
   Detection Prevalence : 0.14073
      Balanced Accuracy : 0.64132

       'Positive' Class : 1
```

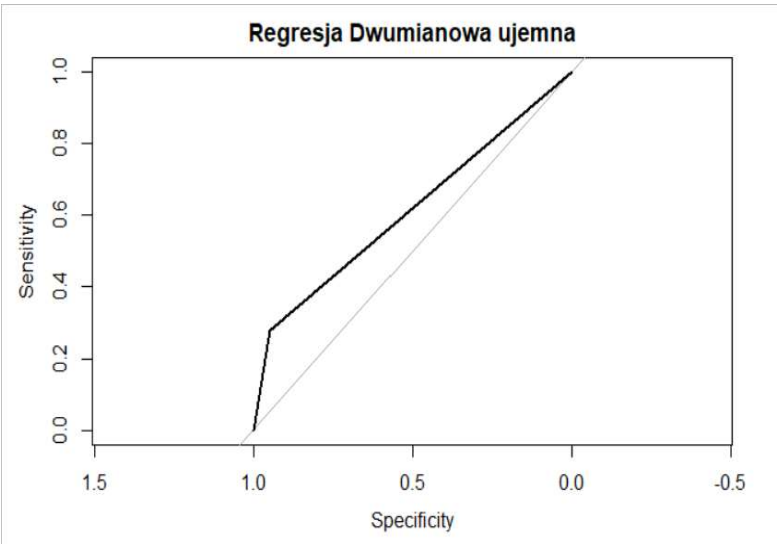



```
AUC(auc_sample2$predictions, auc_sample2$labels,
label.ordering = NULL)
[1] 0.7916034
```

Ewaluacja modelu Regresji dwumianowej, potwierdza nasze wcześniejsze stwierdzenia dotyczące najlepszej jakości tego modelu. Wartość Accuracy oraz AUC to potwierdzają.

6.3 Regresja ujemna dwumianowa

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2304  628
         1  121  244

               Accuracy : 0.7728
                 95% CI : (0.7581, 0.787)
    No Information Rate : 0.7355
    P-Value [Acc > NIR] : 4.5e-07

                  Kappa : 0.2825

 Mcnemar's Test P-Value : < 2e-16

            Sensitivity : 0.27982
            Specificity : 0.95010
         Pos Pred Value : 0.66849
         Neg Pred Value : 0.78581
             Prevalence : 0.26448
         Detection Rate : 0.07401
   Detection Prevalence : 0.11071
      Balanced Accuracy : 0.61496

       'Positive' Class : 1
```




```
AUC(auc_sample3$predictions, auc_sample3$labels,
label.ordering = NULL)
[1] 0.6149598
```

Ostatni model wypada najgorzej ze wszystkich trzech co również zostało potwierdzone w wyliczeniach z użyciem funkcji confusionMatrix(), oraz roc() i AUC().