LabStat2 - Sprawozdanie 3

. . . czyli jak znaleźć recydywistów..?

Makowski Michał

31 stycznia 2017r.

Spis treści

1 Wstęp

1.1 Problem

W wielu systemach wymiaru sprawiedliwości na całym świecie, więźniowie którzy nie stanowią zagrożenia dla społeczeństwa (albo przynajmniej nie zdawali się go stwarzać do momentu wypuszczenia), są wysyłani na tzw. zwolnienie warunkowe. Będąc na takim zwolnieniu nadal traktowani są jako odbywający karę, jeśli naruszą ustalone warunki, to zostaną przywróceni do odbywania kary w więzieniu.

W USA sądy tzw. *parole boards* (parole - zwolnienie warunkowe) decydują, którzy więźniowie są dobrymi kandydatami do zwolnienia warunkowego. Mają one za zadanie oceniac, czy więzień dopuści się kolejnego wykroczenia będąc na zwolnieniu. Problematyczne jest to, że takie sądy postępują subiektywnie i ich decyzja może być obarczona fatalna w skutkach. Postaramy się pomóc *parole boards* poprze z zbudowanie modelu, który w obiektywny, matematyczny sposób pomoże podjąc decyzję o zwolnieniu. Model miałby pomogać podjąc decyzję, a nie zastępować ocenę komisji.

1.2 Dane

Do tego zadania będziemy wykorzystywać dane z *United States 2004 National Corrections Reporting Program*, pochodzące prawdopodobnie ze strony icpsr.umich.edu. My posługujemy się wersją przygotowaną specjalnie na nasze potrzeby. NCRP jest to ogólnokrajowy spis wieźniów My badamy tylko tych których dotyczyło zwolnienie warunkowe. Zbiór ograniczony jest do osobników, którzy spędzili więzieniu nie więcej niż 6 miesięcy, a ich całkowity wyrok nie przekracza 18-tu miesięcy pozbawienia wolności. Nałożone jest też dodatkowe ograniczenie, które zakłada, że więzień musiał albo odbyć poprawnie zwolnienie, albo musiał złamać jego zasady i powrócił spowrotem do zakładu karnego.

1.3 Cel

Tak jak wcześniej wspomniano, głównym celem będzie stworzenie modelu regresji logistycznej, który ma przewidywać prawdopodobieństwo naruszenia zwolnienia warunkowego.

1.4 Droga do celu

Aby poprawnie zbudować model podzielimy nasz zbiór na podzbiory: treningowy i walidacyjny, w proporcjach 0.7, 0.3 odpowiednio. Pierwszy będzię służył do ekploracji danych i budowy kilku modeli, drugi do sprawdzenia modeli i wyboru tego ostatecznego. Łącznie do dyspozycji mamy 675, jednakże po podziale zbiór treningowy bedzie zawierał 472 rekordów co może nie jest potężną wielkością, ale wystarcza eksploracja miała sens.

2 Podstawowa analiza i przygotowanie danych

Poniżej prezentujemy opis, informacje jakie posiadamy:

```
1. male: 1 if the parolee is male, 0 if female
2. race: 1 if the parolee is white, 2 otherwise
3. age: the parolee's age in years at release from prison
4. state: a code for the parolee's state. 2 is Kentucky, 3 is Louisiana, 4 is Virginia, and 1 is any other st
5. time.served: the number of months the parolee served in prison (limited by the inclusion criteria to not e
6. max.sentence: the maximum sentence length for all charges, in months (limited by the inclusion criteria to
8. crime: a code for the parolee's main crime leading to incarceration. 2 is larceny, 3 is drug-related crime
crime, and 1 is any other crime.
9. violator: 1 if the parolee violated the parole, and 0 if the parolee completed the parole without violatio
```

Tak jak wcześniej wspomnieliśmy, do dyspozycji mamy 472 obserwacji 9 zmiennych. Zmienną objaśnianą jest oczywiście ostatnia kolumna, która określa, czy nastąpiło naruszenie zwolnienia wrunkowego czy też nie.

2.1 Przygotowanie i eksploracja danych

Na początku sprawdźmy, czy nasze dane są kompletne:

```
identical(parole, parole[complete.cases(parole),])
```

```
[1] TRUE
```

Wiedzac to możemy przejść do *sfaktoryzowania* danych. Znakomita większość informacji jakie posiadamy to dane kategoryczne, informacja czy dane zdarzenia miało miejsce czy nie. Zmienimy także wartości zmiennych sfaktoryzowanych, ma to jedynie na celu zwiększenie czytelności i ułatwienie prac (np. w ggplot2).

Sporna pozostaje kwestia długości wyroku, być może warto zbudować dwa modele, jeden opierający się na (w pewien sposób) skategoryzowanych danych, drugi na traktujący długość odsiadki jako zmienną jakościową.

Przyjrzyjmy się danym:

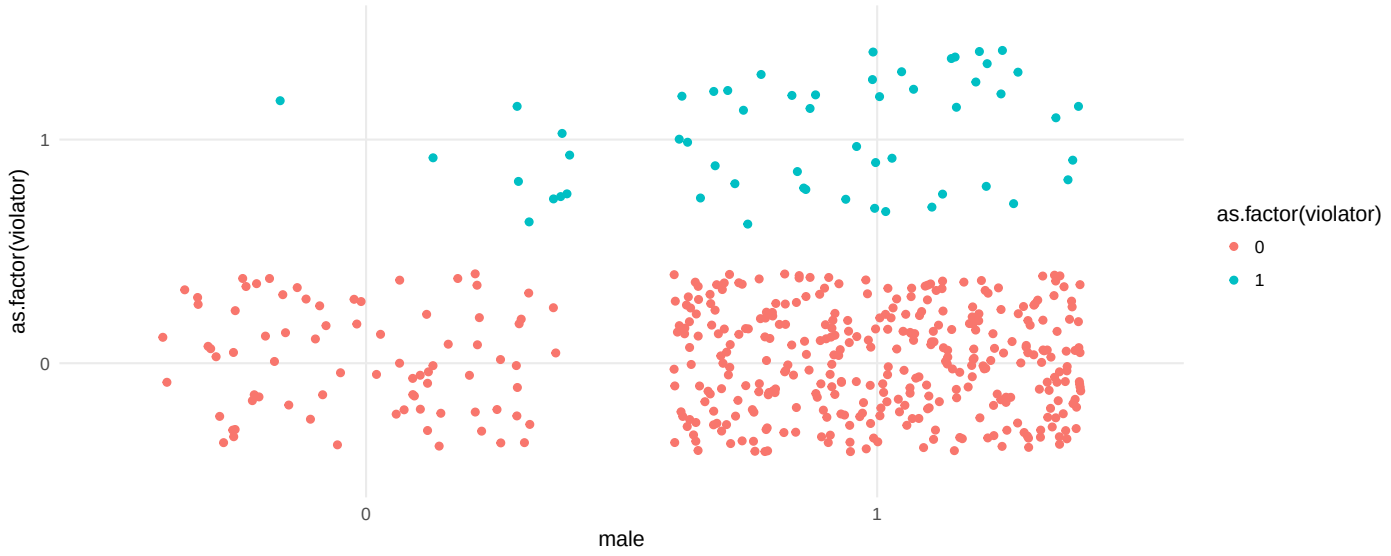
```
 male       race          age              state      time.served     max.sentence
 0:130   Other:286   Min.   :18.40   Kentucky :120   Min.   :0.000   Min.   : 1.00
 1:545   White:389   1st Qu.:25.35   Louisiana: 82   1st Qu.:3.250   1st Qu.:12.00
                     Median :33.70   Other    :143   Median :4.400   Median :12.00
                     Mean   :34.51   Virginia :330   Mean   :4.198   Mean   :13.06
                     3rd Qu.:42.55                   3rd Qu.:5.200   3rd Qu.:15.00
                     Max.   :67.00                   Max.   :6.000   Max.   :18.00
 multiple.offenses             crime        violator
 0:313             Driving-reletd:101   Min.   :0.0000
 1:362             Drug-reletd   :153   1st Qu.:0.0000
                   Larceny       :106   Median :0.0000
                   Other         :315   Mean   :0.1156
                                        3rd Qu.:0.0000
                                        Max.   :1.0000
```

Liczba więźniów łamiących zasadę zwolnienia warunkowego jest stosunkowo mała, aby w odpowiedni sposób podzielić dane na dwa zbiory posłużymy się funkcją *sample.split* z pakietu *caTools*, która zapewnia podział próby wg. odpowiednich proporcji.

W kolejnych podrozdziałach przyjrzymy się każdej ze zmiennych, postaramy się znaleźć zależnośći miedzy nią, a złamaniem zwolnienia warunkowego.

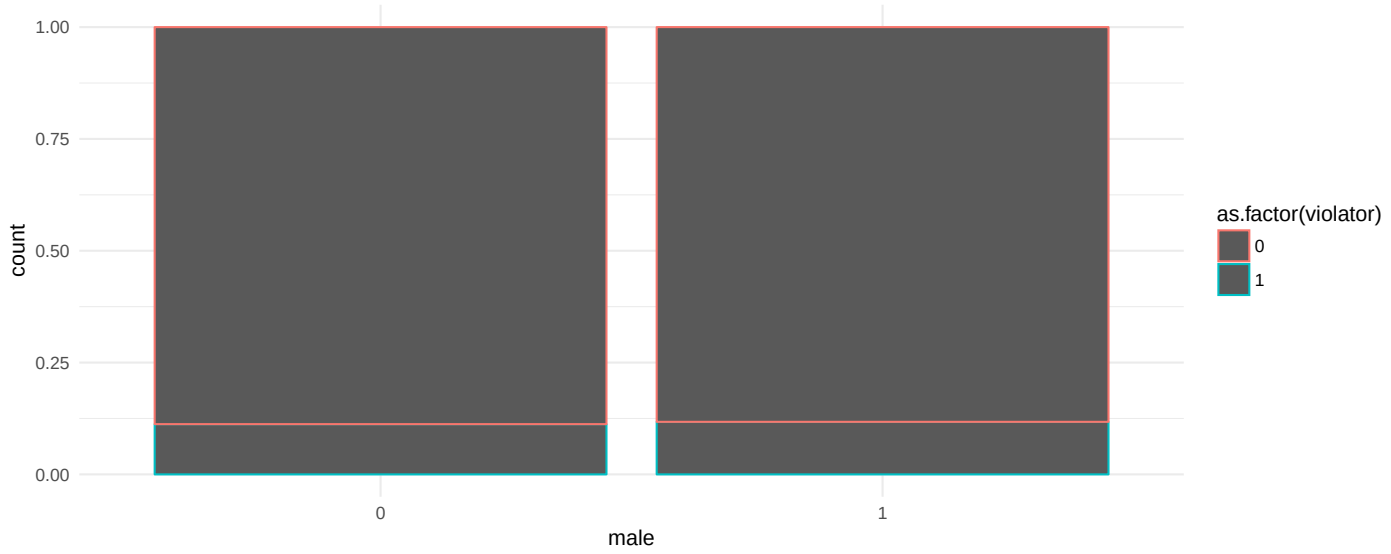
2.2 Płeć

Przyjrzyjmy się rozkładowi recydywistów ze względu na płeć:



Powyższy wykres nie mówi nic, poza tym, że wśród więźniów na zwolnieniu obserwujemy więcej mężczyzn. Jest to intuicyjne, gdyż wśród więźniów ogólnie przeważają mężczyźni -> LINK.

Znormalizujmy liczbę więźniów każdej z płci i przyjrzyjmy się danym po raz kolejny:



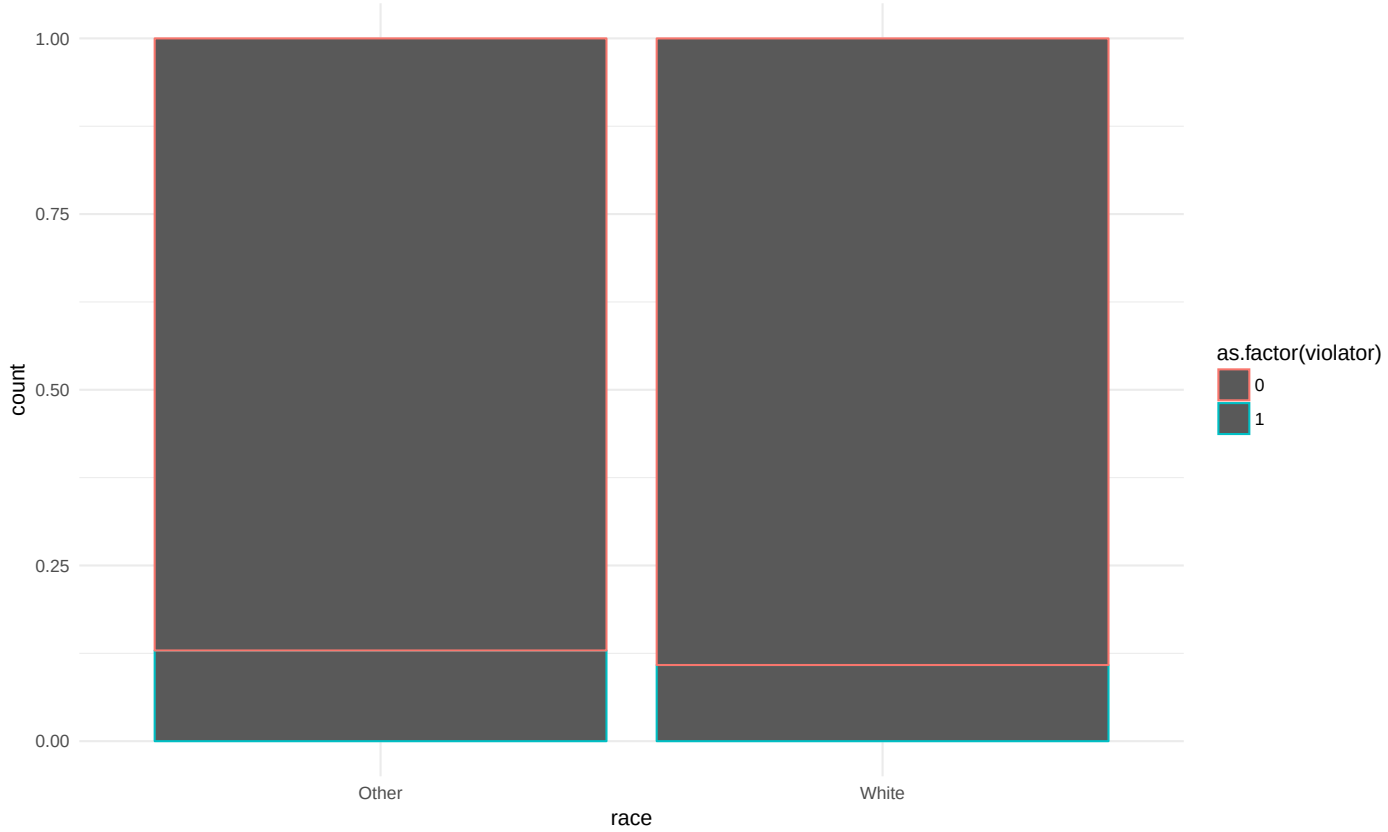
Tym razem widać, a raczej nie widać zależności pomiędzy płcią, a recydywą. Sprawdźmy jeszcze jak wygląda udział procentowy recydywistów w każdej z płci:

```
sum(paroleTrain$male=="1" & paroleTrain$violator==1)/sum(paroleTrain$male=="1")
[1] 0.1174935
sum(paroleTrain$male=="0" & paroleTrain$violator==1)/sum(paroleTrain$male=="0")
[1] 0.1123596
```

Różnica jest praktycznie żadna, więc płeć nie powinna w żaden sposób wpływać na decyzję komisji (co niekoniecznie może mięc miejsce w rzeczywistości).

2.3 Rasa

Przedstawiamy rozkładowi recydywistów ze względu na kolor skóry. Tym razem pominiemy pierwszy z wykresów używanych przy analizie płci, skupimy się tylko na znormalizowanych wartościach:



Zależność istnieje, choć jest bardzo mała, to jednak widoczna, więźniowie biali rzadziej popełniają wykroszenia na zwolnieniu warunkowym niż pozostali. Sprawdźmy jak wygląda udział procentowy:

```
sum(paroleTrain$race=="White" & paroleTrain$violator==1)/sum(paroleTrain$race=="White")
[1] 0.1083916
sum(paroleTrain$race=="Other" & paroleTrain$violator==1)/sum(paroleTrain$race=="Other")
[1] 0.1290323
```

Różnica dwóch punktów procentowych lub 20%. Druga liczba obrazuje, że różnica jest dośc znaczna na korzyść więźniów białych. Może to być odebrane jako niepoprawnośc polityczna. W prawie amrykańskim, jak i polskim, zapisane jest, że nie należy podejmować decyzji prawnych argumentując je kolorem skóry, więc taki czynnnik i tak staje się bezużyteczny.

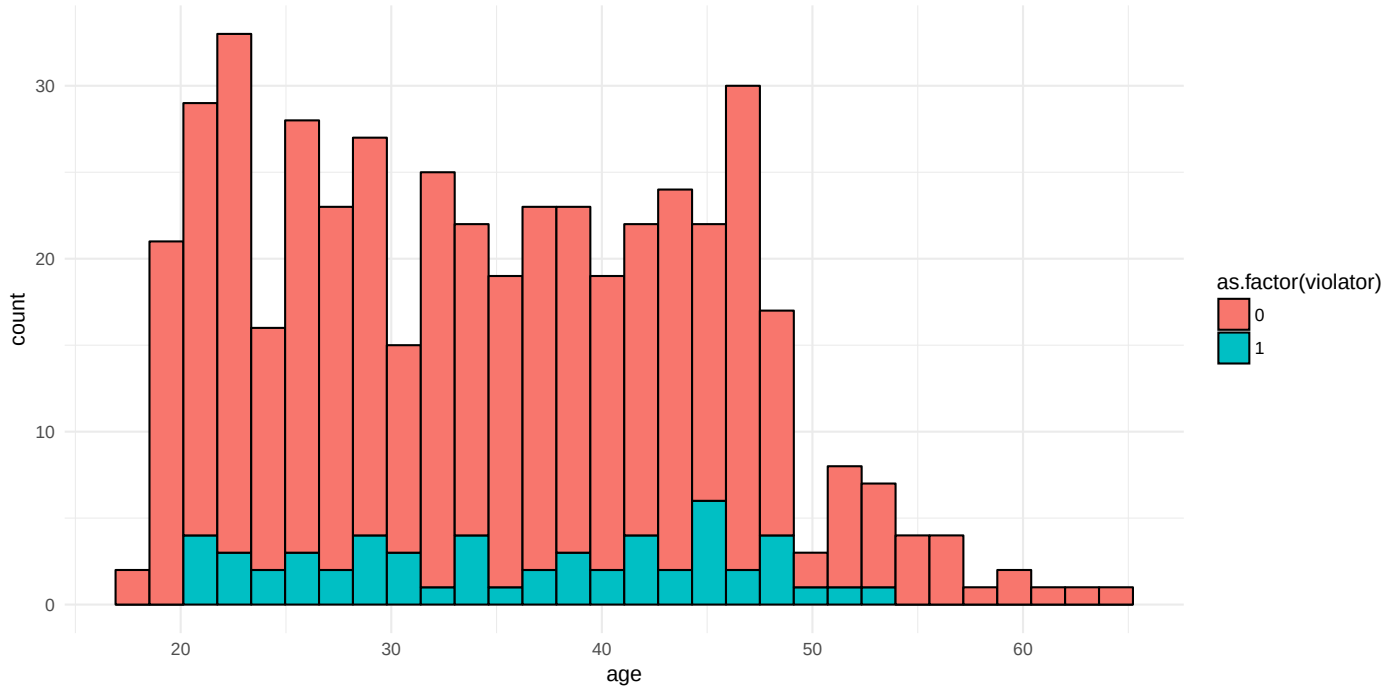
Jedną z dodatkowych obserwacji jest to, że więzienia pewnych stanów są zdominowane przez ludność jednej z ras:

Na chwilę obecną nie wnosi to zbyt wiele do modelu, byc może za chwilę do czegoś dojdziemy. Jest to ciekawa obserwacja, a może być jeszcze ciekawsza, gdybyśmy porównali to z danymi o ludności danego stanu w ogóle. Jak się ma rozkład ras wśród więźniów w stosunku do tego rozkładu w całym stanie?

Przykładowo: w Luizjanie i Wirginii około 60% ludzi to biali amerykanie, a skład więzień odbiega od tego podziału, w Kentucky biali stanowią 90%. Tak analiza i szukanie zależności to nie jest jednak nasz cel, pomimo, że moglibysmy dojść do ciekawych wniosków.

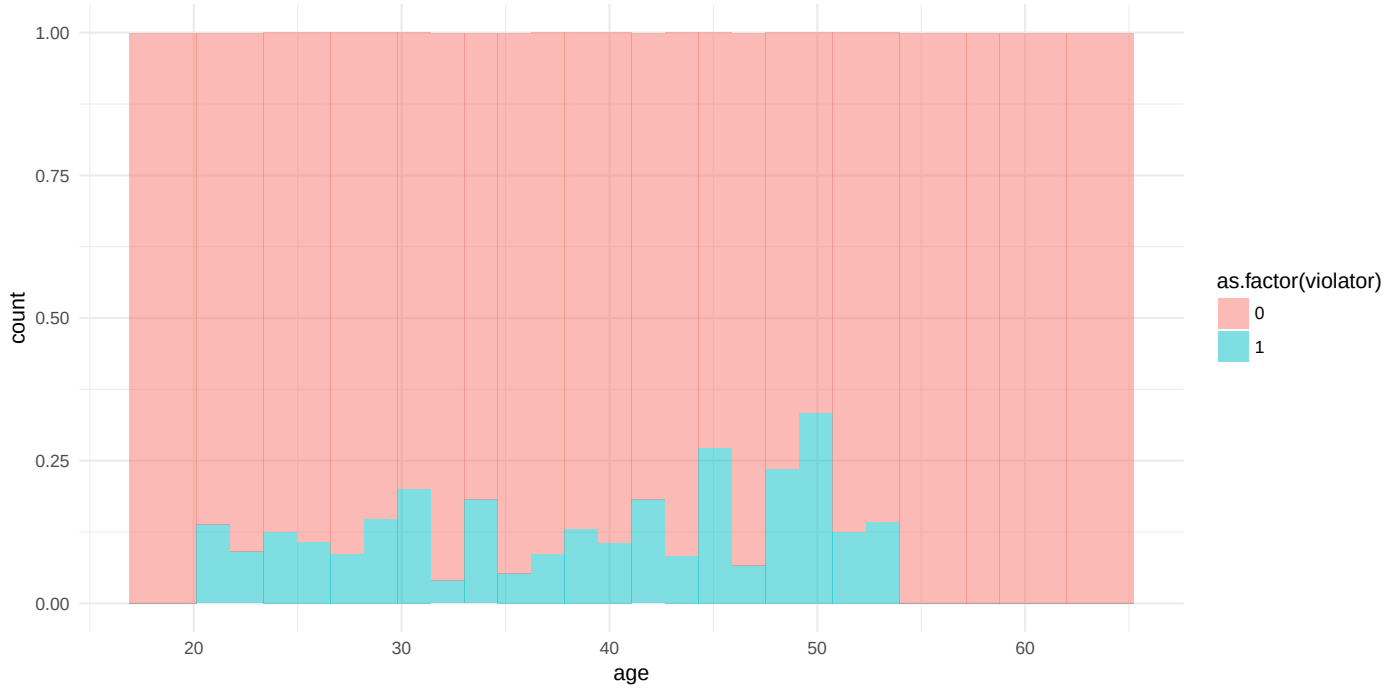
2.4 Wiek

Przyjrzyjmy się jak rozkłada się wiek recydywistów:



Widzimy, że wiek więźniów rozkłada się dość jednostajnie na przedziale 20-50 lat, poza nim obserwujemy wyraźny spadek więźniów. Także na tym przedziale liczba recydywistów jest największa, być moze warto byłoby w pewien sposób skategoryzować zmienną dot. wieku.

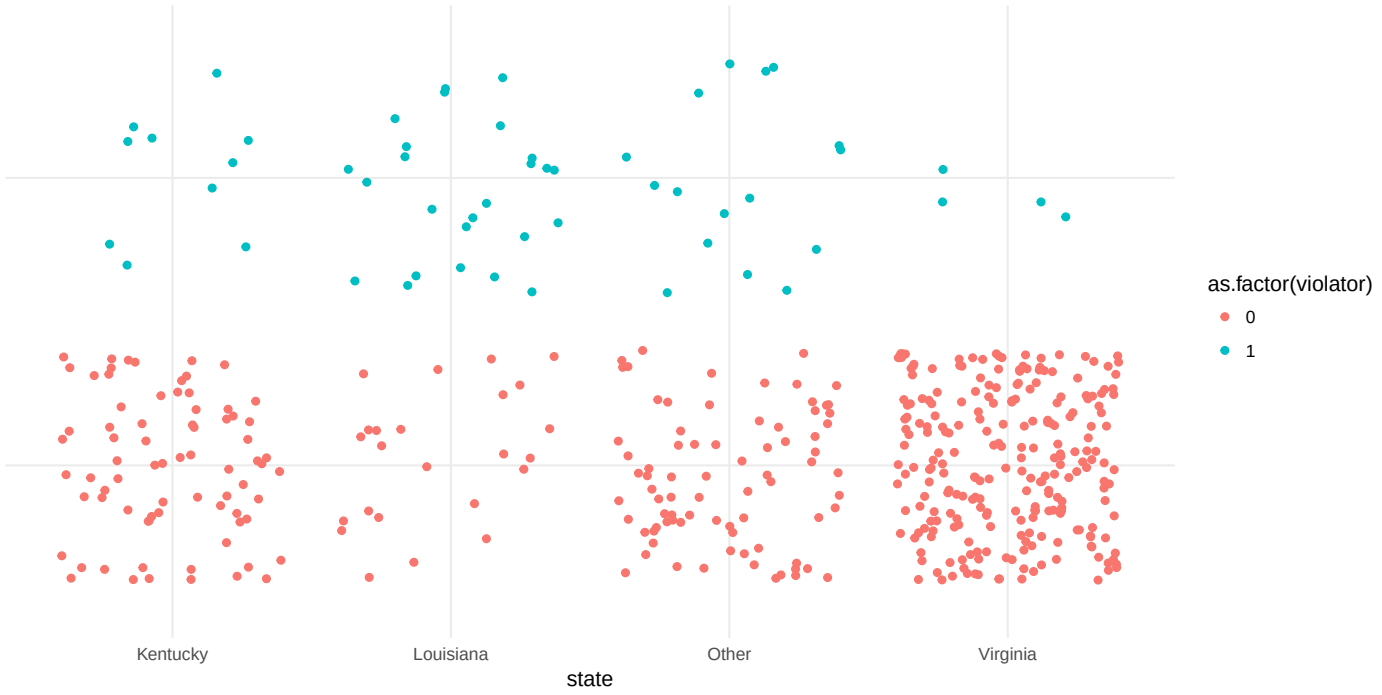
Sprawdźmy jeszcze histogram "skumulowany":



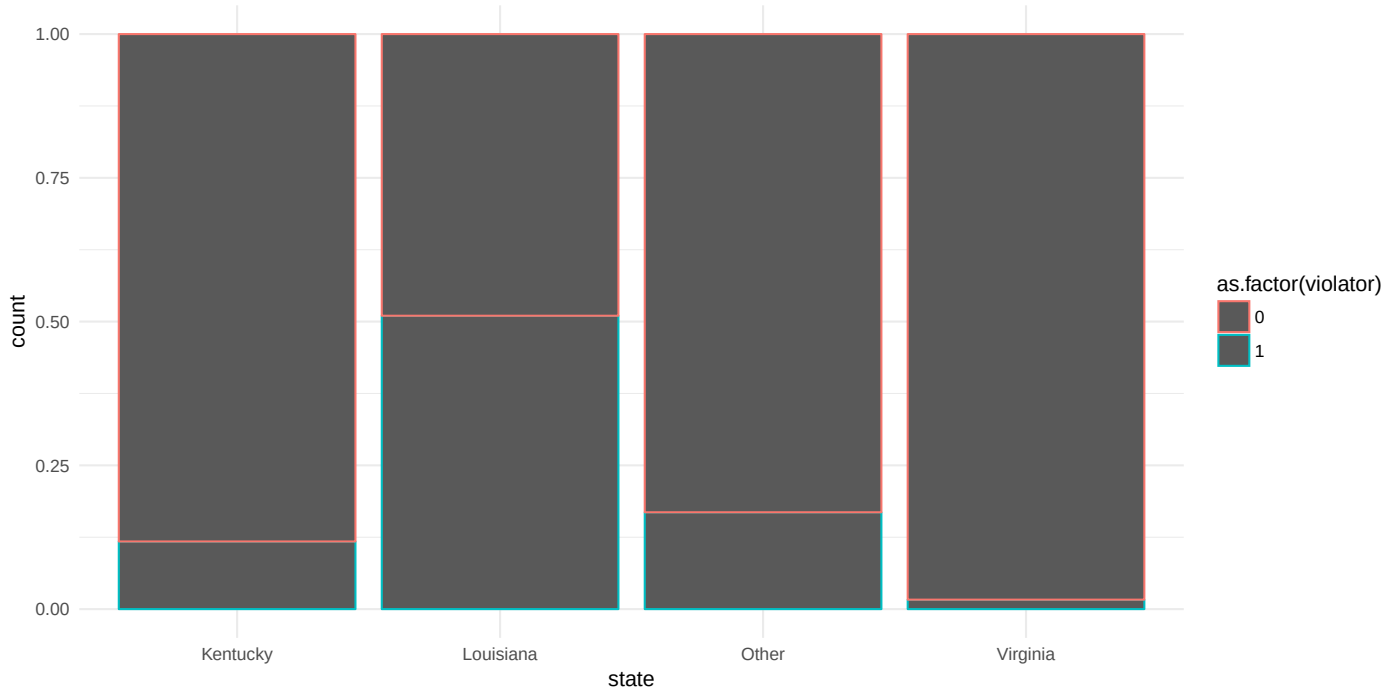
Jak potraktować wiek? Jako zmienną kategoryczną, ilościową czy może pogrupować i wtedy skategoryzować? Stworzymy kilka rozwiązań i porównamy je pomiędzy sobą na etapie budowania modelu.

2.5 Miejsce odsiadki

Przyjrzyjmy się rozkładowi recydywistów ze względu na miejsce odsiadki:



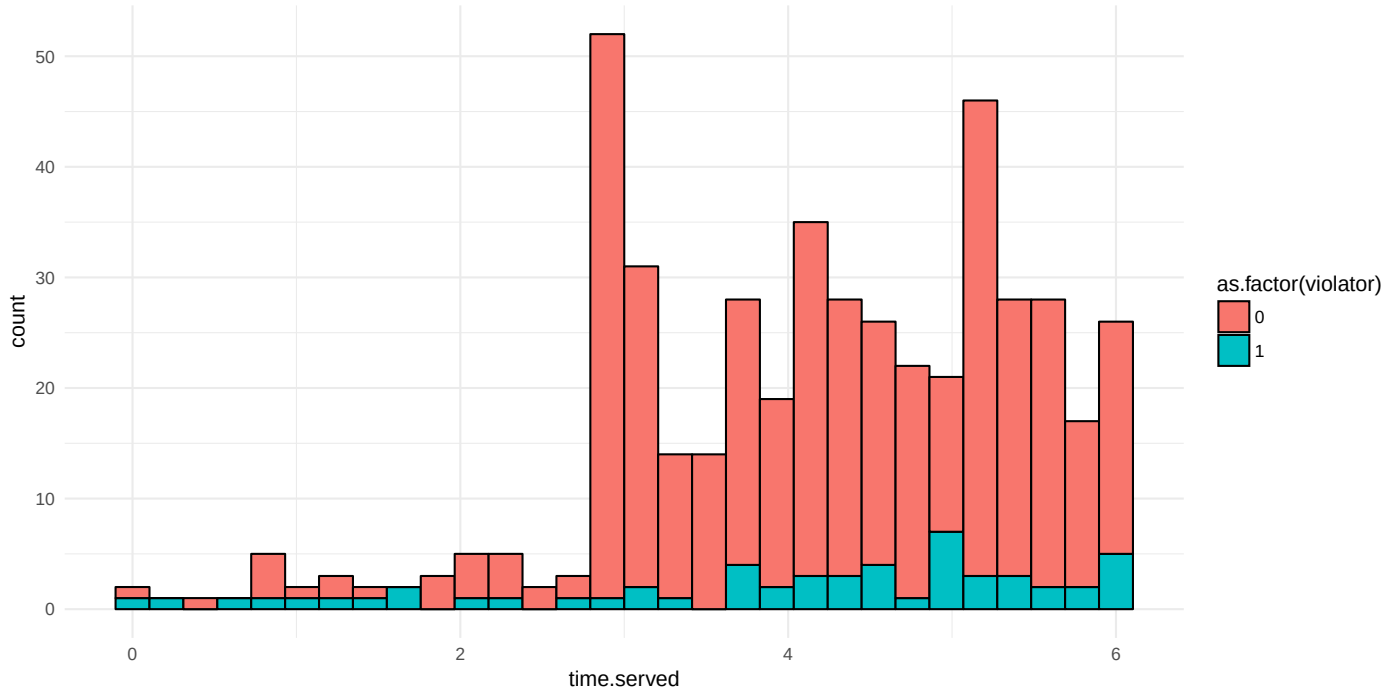
Zdaje się, że w tym przypadku zależnośc jest dość istotna, Virginia wydaje się "spokojniejszym" stanem niż pozostałe, z kolei w Luizjanie liczba więźniów łamiących zasady zwolnienia warunkowego jest względnie wysoka. Spojrzmy jak te wartości rozkładają się po znormalizowaniu:



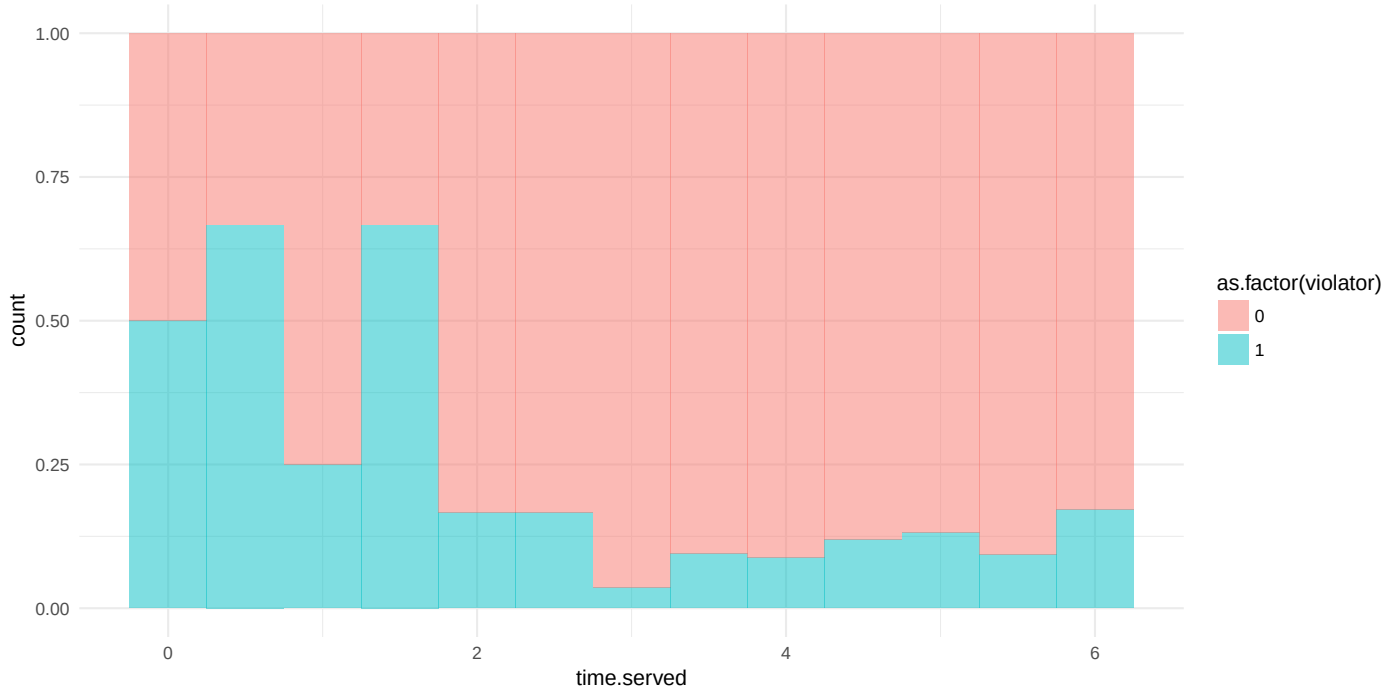
Nasze wstępne obserwacje potwierdzają się - Luizjana to zdecydowanie niespokojny rejon, odsetek więźniów łamiących zwolnienia jest kilkukrotnie wyższy niż w innych stanach. Ta zmienna zdecydowanie powinna znaleźć się w naszym modelu.

2.6 Dotychczasowy czas odsiadki

Przyjrzyjmy jak rozkłada się czas odsiadki do momentu zwolnienia:



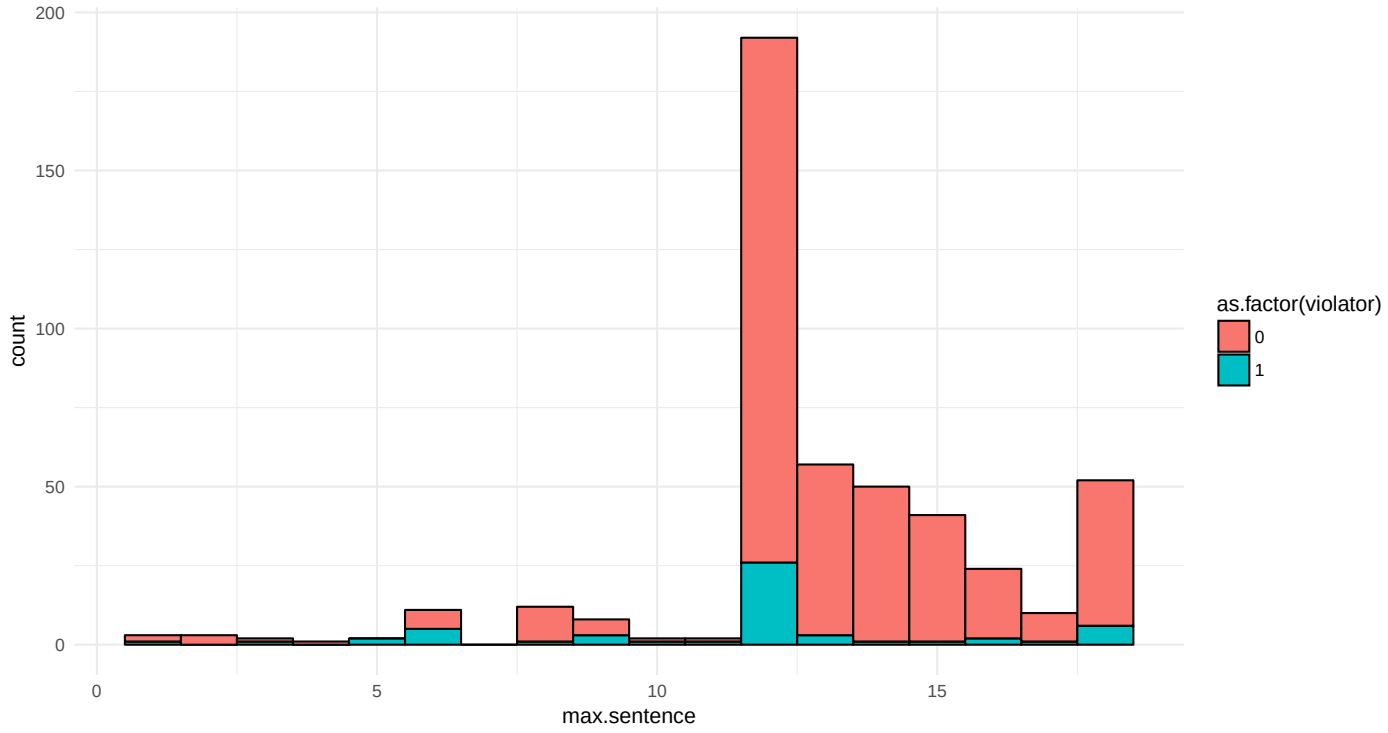
Widzimy, że wśród więźniów którzy krótko odbywali swoją karę, łamanie zwolnienia warunkowego było popularniejsze. Będzie to lepiej widoczne do "znormalizowanym histogramie":



Widoczna jest zależność: im krótsza dotychczasowa odsiadka, tym większe prawdopodobieństwa popełnienia wykroczenia na zwolnieniu warunkowym. Wydaje się to być zgodne z intuicją: więzień po krótszej odsiadce mógł jeszcze nie poznać i zrozumieć co mu odebrano i dlatego nie czuł obaw przed powrotem do celi. Skategoryzuje sobie tą zależność, byc może takie uproszczenie pomoże nam skonstruować dokładniejszy model.

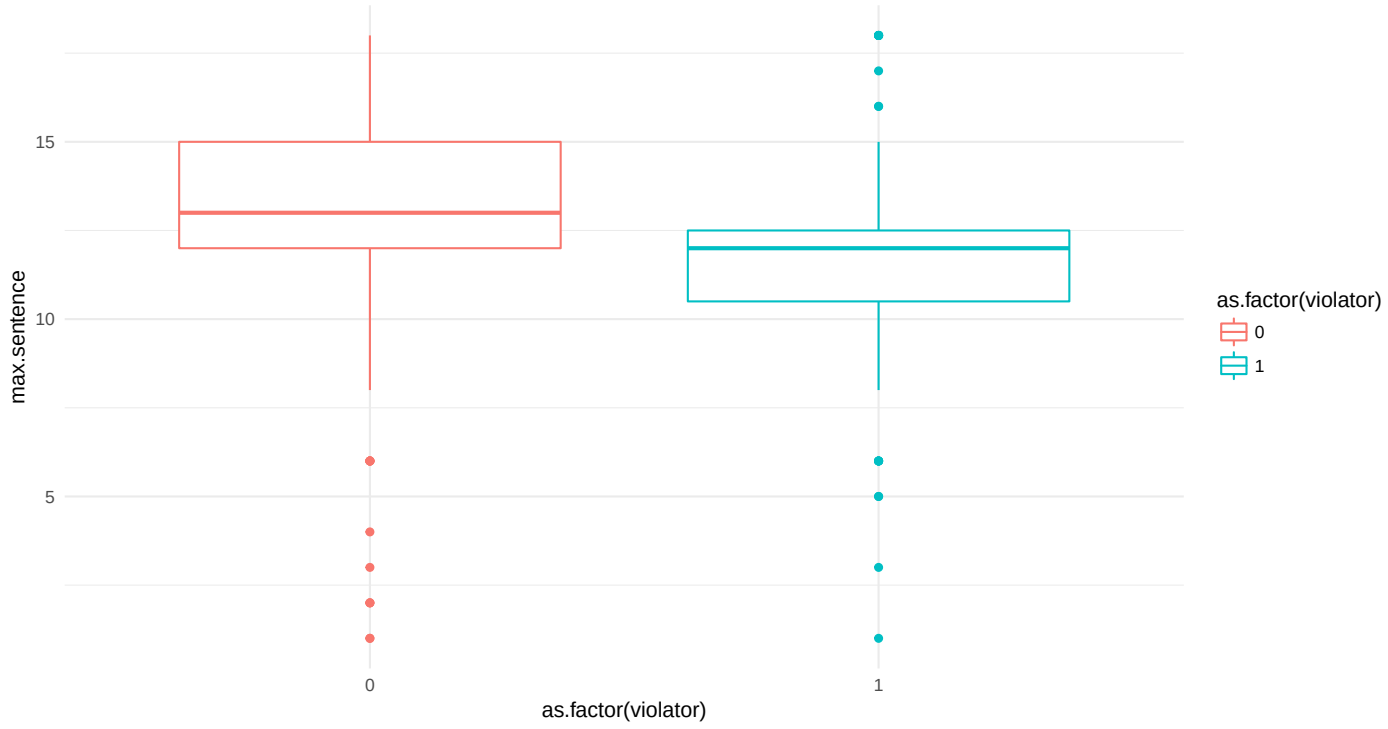
2.7 Całkowity czas odsiadki

Przyjrzyjmy jak rozkłada się czas całkowitej, maksymalnej odsiadki:



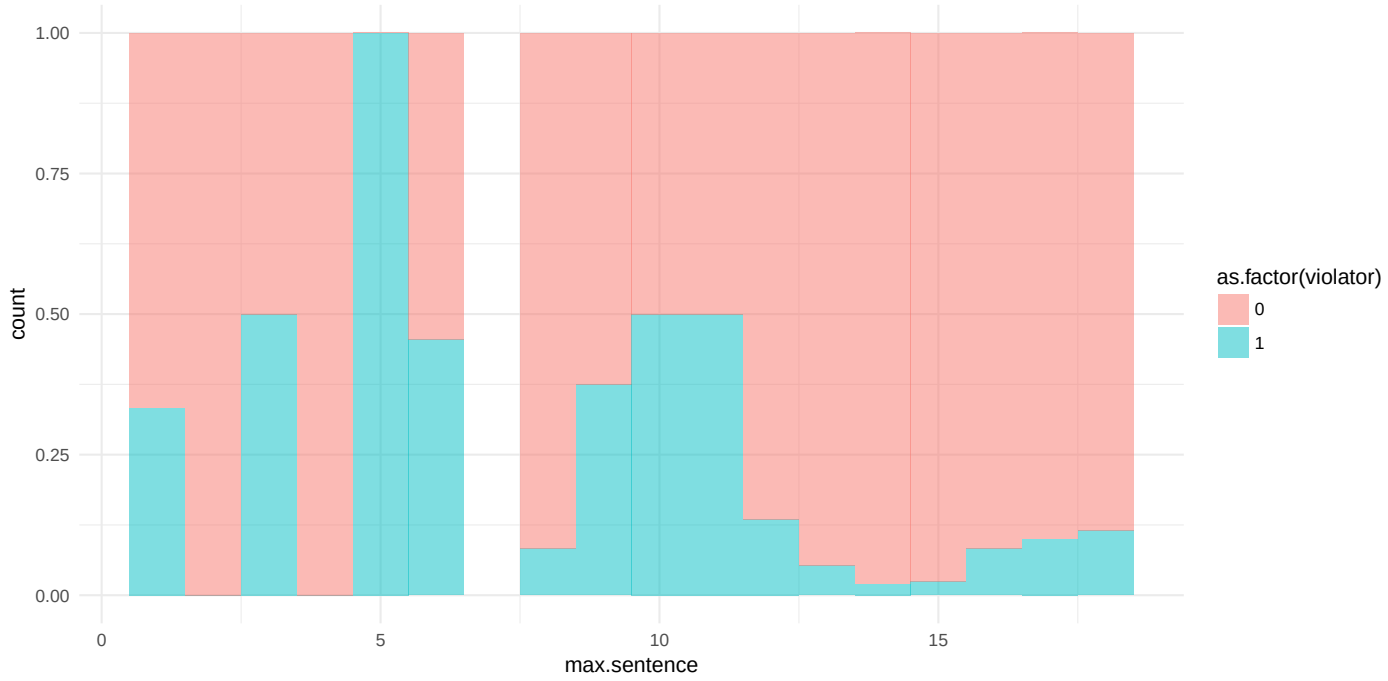
Z racji dominacji wyroków w okolicach roku i większych niewiele jesteśmy w stanie powiedzieć o rozkładzie recydywistów korzystając z tego histogramu. Możemy jednak zauważyć, że wyroki poniżej roku są bardzo rzadkie w stosunku do pozostałych.

Sprawdźmy jak ma się rozkład wieku wsród każdej z grup: łamiących zwolnienie i tych nie:



Widzimy, że wsród młodszych więźniów łamanie zwolnienia zdarza się częściej.

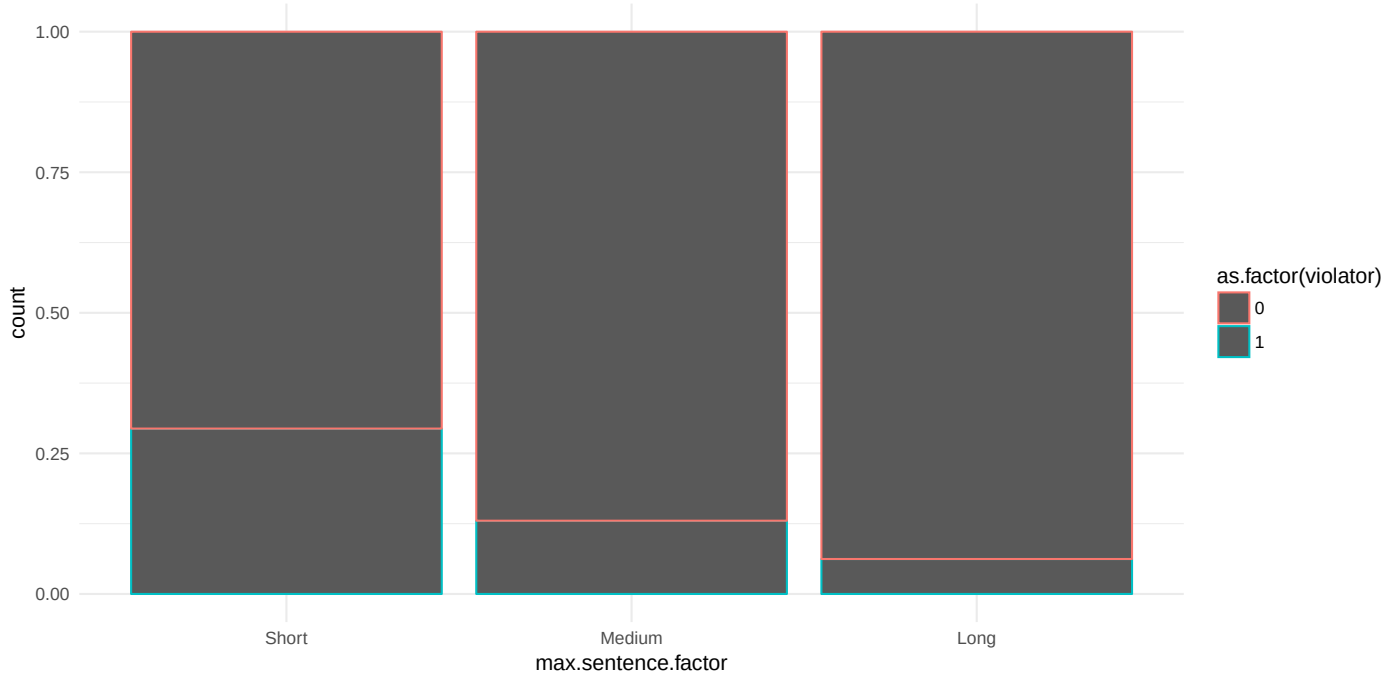
Aby widzieć więcej zależności po raz kolejny posłużymy się "znormalzowanym" histogramem:



Widoczna jest słaba zależnośći drugiego lub wręcz trzeciego stopnia. Tutaj moglibyśmy skategoryzować zmienne w natępujący sposób:

- Wyrok do 8 mies.
- Wyrok powyzej 8, a poniżej 14 mies.
- Wyrok od 14 mies.

co uczynimy. Nie zastąpimy jednak danych, a jedynie stworzymy nową zmienną, która może się przydać przy budowie różnych modeli. Zobaczymy jak teraz wygląda rozkład:

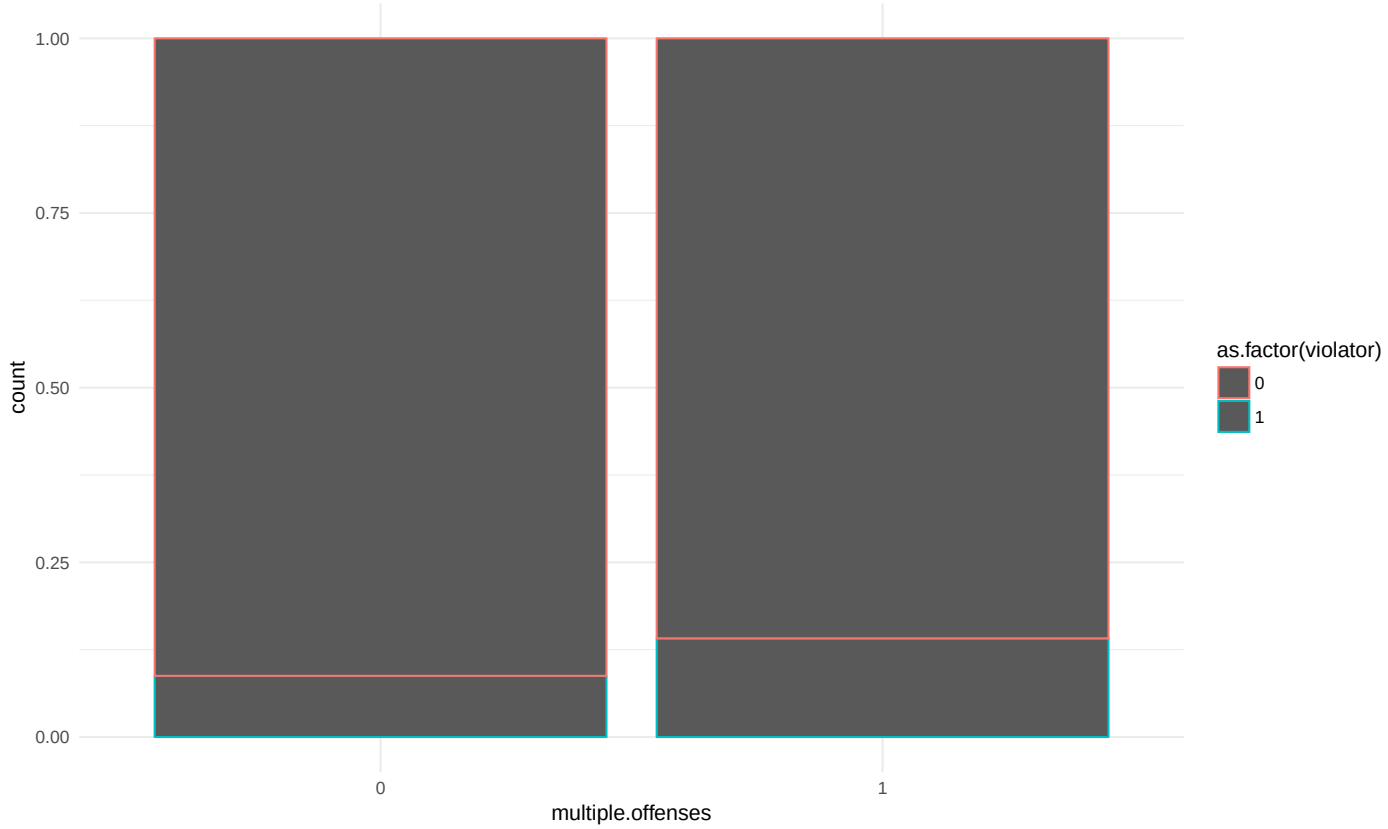


Im krótszy czas całkowitej odsiadki tym "chętniej"" więźniowie wracają do celi.

W końcu wiele nie stracą...

2.8 Liczba przestępstw

Przyjrzyjmy się rozkładowi więźniów ze względu liczbę popełnionych wykroczeń:



Widzimy, że wśród osób, które popełniły więcej niż jedno wykroczenie, ryzyko związane z wypuszczenie mich na zwolnienie warunkowe jest podwyższone, Sprawdźmy jak wyglądają liczby:

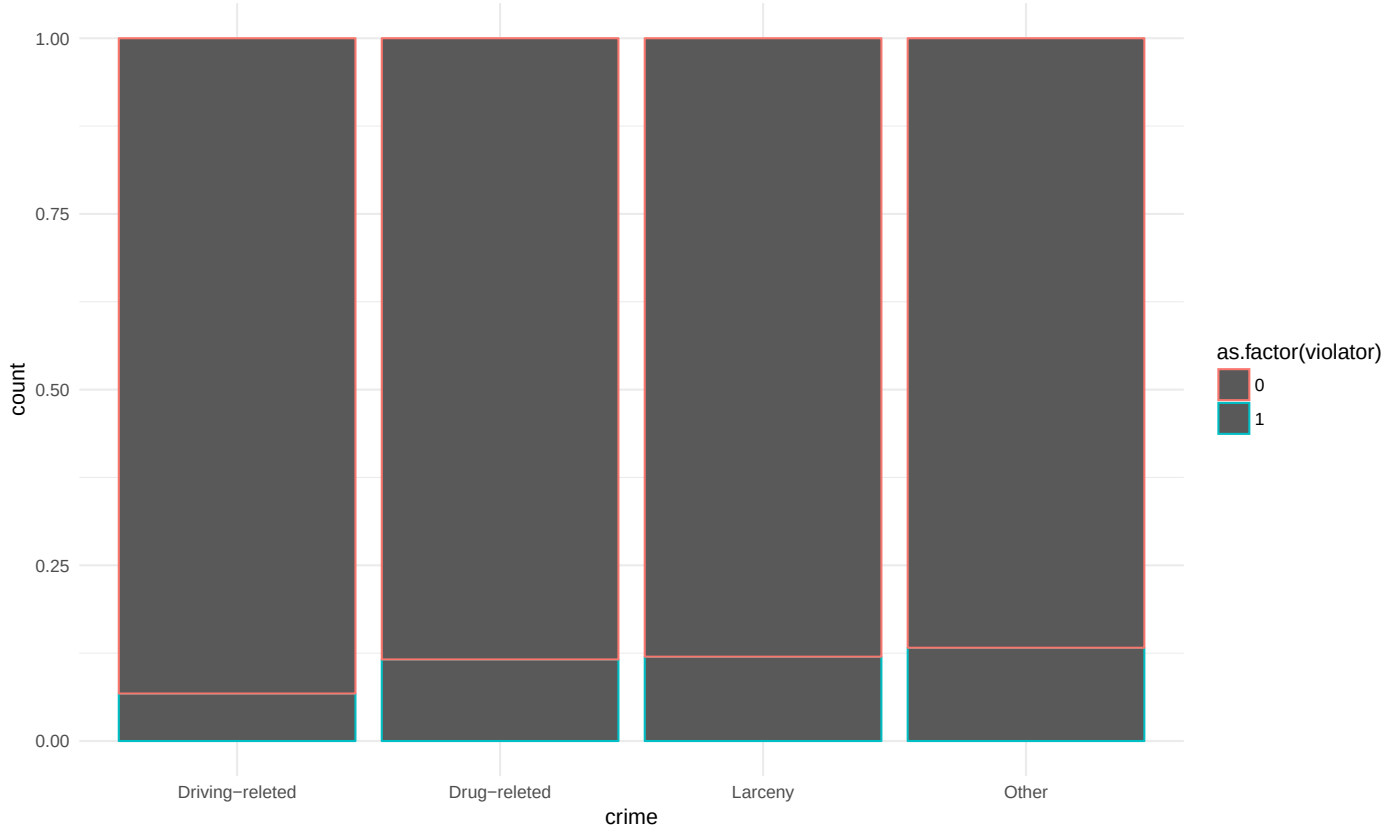
```
sum(paroleTrain$multiple.offenses=="1" & paroleTrain$violator==1)/sum(paroleTrain$multiple.offenses=="1")
[1] 0.1411765
sum(paroleTrain$multiple.offenses=="0" & paroleTrain$violator==1)/sum(paroleTrain$multiple.offenses=="0")
[1] 0.0875576
```

Różnica jest prawie dwukrotna, współczynnik ten powinień mieć istotny wpływ na ponoszone ryzyko.

2.9 Rodzaj przestępstwa

Przyjrzyjmy się rozłozeniu więźniów ze względu na popełnione przestępstwo:

Wydaje się, że kierowcy rzadziej łamią warunki zwolnienia, zobaczmy jak to wygląda na wykresie skumulowanym:



Faktycznie, współczynnik osób łamiących zasady zwolnienia wśród kierowcóW jest prawie dwukrotny niższy niż w pozostałych. Nie znamy dokładnych zasad zwonienia warunkowego, ale pomimo opini, że jeśli raz wsiąknie się w świat narkotyków nie da się z niego uciec, to współczynnik recydiwistów wśróD tej grupy wcale nie jest wyższy niż w pozostałych.

Spójrzmy na udział procentowy w każdej z grup:

```
sum(paroleTrain$crime=="Driving-releted" & paroleTrain$violator==1)/
    sum(paroleTrain$crime=="Driving-releted")
[1] 0.06756757
sum(paroleTrain$crime=="Drug-releted" & paroleTrain$violator==1)/
    sum(paroleTrain$crime=="Drug-releted")
[1] 0.1160714
sum(paroleTrain$crime=="Larceny" & paroleTrain$violator==1)/
    sum(paroleTrain$crime=="Larceny")
[1] 0.12
sum(paroleTrain$crime=="Other" & paroleTrain$violator==1)/
    sum(paroleTrain$crime=="Other")
[1] 0.1327014
```

Obserwacje się potwierdzają, kierowcy są grupą najmniejszego ryzyka. To istotna informacja dla naszego modelu.

3 Model

3.1 Wstęp

Zobaczmy jak prezentują się nasze dane:

```
'data.frame':   472 obs. of  14 variables:
 $ male               : Factor w/ 2 levels "0","1": 2 1 2 2 2 2 2 2 2 2 ...
 $ race               : Factor w/ 2 levels "Other","White": 2 2 1 1 2 2 2 1 2 2 ...
 $ age                : num  33.2 39.7 29.5 46.7 20.5 30.1 37.8 43.5 42.3 21.3 ...
 $ state              : Factor w/ 4 levels "Kentucky","Louisiana",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ time.served        : num  5.5 5.4 5.6 6 5.9 5.3 5.3 5.2 4.8 5.1 ...
 $ max.sentence       : int  18 12 12 18 12 16 8 8 16 8 ...
 $ multiple.offenses  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ crime              : Factor w/ 4 levels "Driving-releted",..: 1 2 2 1 4 2 2 4 2 4 ...
 $ violator           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ age.group.factor   : Factor w/ 3 levels "Young","Middle",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ age.5.factor       : Factor w/ 10 levels "(18,23]","(23,28]",..: 4 5 3 6 1 3 4 6 5 1 ...
 $ time.served.factor : Factor w/ 2 levels "Short","Long": 2 2 2 2 2 2 2 2 2 2 ...
 $ max.sentence.factor: Factor w/ 3 levels "Short","Medium",..: 3 2 2 3 2 3 1 1 3 1 ...
 $ isDriver           : Factor w/ 2 levels "0","1": 2 1 1 2 1 1 1 1 1 1 ...
```

3.2 Budowa i wstępnę porównanie

Zbudujmy kilka podstawowych modeli, będą różniły się użytymi uproszczeniami, żaden z nich nie zawera zmiennej dot. płci oraz rasy, gdyż te odrzucilismy w momencie audytu danych:

```
regLogParole=function(index)
{ return(glm(violator~.,data=paroleTrain[,index], family="binomial")) }

paroleTrain.glmFull          =regLogParole(c(-1,-2,            -10,-11,-12,-13,-14))
paroleTrain.glmFullExtra     =regLogParole()
paroleTrain.glmAgeGroupFactor=regLogParole(c(-1,-2,-3,             -11,-12,-13,-14))
paroleTrain.glmAge5Factor    =regLogParole(c(-1,-2,-3,         -10,    -12,-13,-14))
paroleTrain.glmServed        =regLogParole(c(-1,-2,   -5,      -10,-11,    -13,-14))
paroleTrain.glmSentence      =regLogParole(c(-1,-2,      -6,   -10,-11,-12,    -14))
paroleTrain.glmSimple        =regLogParole(c(-1,-2,-3,-5,-6,-8,    -11            ))
paroleTrain.glmDriver        =regLogParole(c(-1,-2,         -8,-10,-11,-12,-13    ))
```

Zbudowaliśmy bardzo wiele modeli, może nawet zbyt wiele. Nie będziemy ich tutaj po kolei opisywać, powyższe przedstawienie daje możliwość poznania na jakich danych zbudowanych jest model. Szybki rzut oka na średnie, różnice pomiędzy średnimy:

Tablica 1: Średnie dopasowanych wartość dla każdej z grup

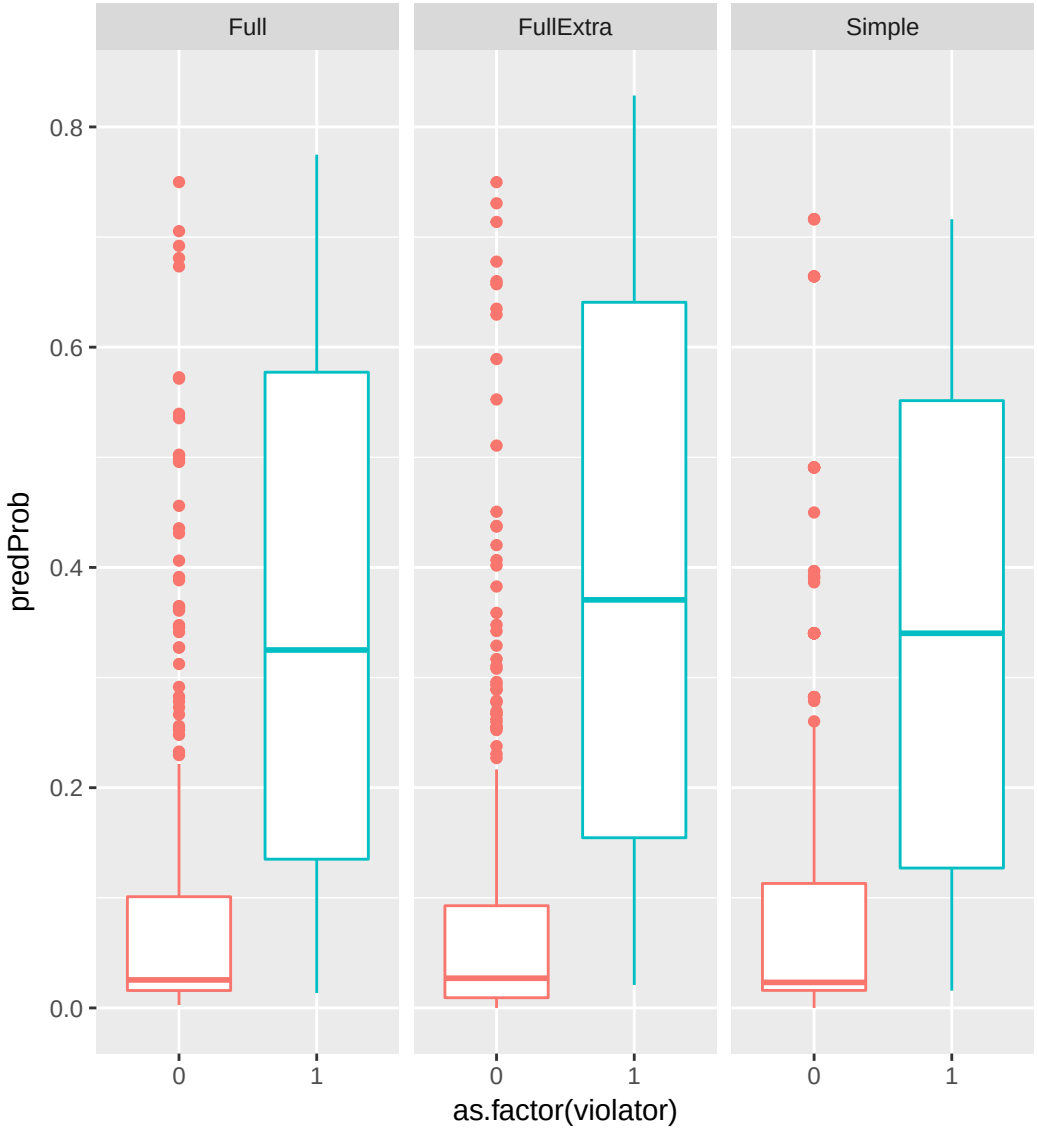
	0	1	diff
Full	0.085	0.354	0.269
FullExtra	0.079	0.397	0.318
AgeGroupFactor	0.085	0.355	0.269
Age5factor	0.082	0.380	0.299
Serve	0.085	0.354	0.269
Sentence	0.085	0.357	0.272
Simple	0.085	0.353	0.268
Driver	0.085	0.352	0.267

Jak widać rożnice między średnimi są dośc znaczące, co dobrze wróży na przyszłość. Wszystkie zbudowane modele są

bardzo do siebie zbliżone, Okazuje się, że najbardziej skomplikowane modele niekoniecznie mają sens.

3.3 Boxplots

Spójrzmy teraz na boxploty przedstawiające informacje o rozkładzie wartości dopasowanych:



Zdecydowaliśmy się przedstawić tylko 4 wykresy pudełkowe, pozostałe cztery niewiele się od nich różnią i tylko sztucznie zwiększają rozmiar raportu. Boxploty wyglądają okej, dość istotna odległość pomiędzy pierwszym i trzecim kwartlem może być problematyczna dla obserwcji recydywistóW, choć duża różnica w medianie i średnich pomiędzy obydwoma stanami powinna z tym pomóc.

3.4 Confusion Matrix

Confusion Matrix pozwalają zobrazować jakie decyzje podejmował model. Przypomnijmy jak jest ona skonstruowana:

	Przewidziane 0	Przewidziane 1
Faktyczne 0	TRUE NEGATIVE (TN)	FALSE NEGATIVE (FP)
Faktyczne 1	FALSE NEGATIVE (FN)	TRUE POSITIVE (TP)

Poniżej przedstawiamy *Confusion Matrix* dla każdego ze zbudowanych modeli:

	FALSE	TRUE
0	406	11
1	35	20

(a) Model Full

	FALSE	TRUE
0	406	11
1	36	19

(b) Model FullExtra

	FALSE	TRUE
0	402	15
1	37	18

(c) Model AgeGroupFactor

	FALSE	TRUE
0	406	11
1	34	21

(d) Model Age5factor

	FALSE	TRUE
0	405	12
1	35	20

(e) Model Serve

	FALSE	TRUE
0	409	8
1	37	18

(f) Model Sentence

	FALSE	TRUE
0	410	7
1	37	18

(g) Model Simple

	FALSE	TRUE
0	406	11
1	35	20

(h) Model Driver

Tablica 3: Porównanie Confusion Matrix

Jak widać wszystkie modele są do siebie bardzo zbliżone, różnice są marginalne, a każdy ze zbydowanych modeli dośc dobrze dopasowuje się do danych (jeżeli możemy tak mówić w przypadku danych, na których model się uczył).

3.5 AIC & AUC

W tym podrozdziale skupimy się na dwóch miarach dobroci modelu, AUC i AIC. Nie będziemy tutaj rozpisywać się na temat tychże wskaźników, wiadome jest, że preferujemy modele z niższym AIC oraz wyższym AUC. Wartość AUC jest ścisle związana z krzywą ROC, która będziemy analizowali w jednym z kolejnych rozdziałóW.

Tablica 4: Miary dobroci modelu

	AIC	AUC
Full	261.5615	0.8651
FullExtra	275.9616	0.8923
AgeGroupFactor	275.9616	0.8802
Age5factor	267.2151	0.8807
Serve	261.8008	0.8632
Sentence	263.2574	0.8645
Simple	259.7152	0.8659
Driver	258.6078	0.8620

Po raz kolejny nie dochodzimy do niczego zaskakującego, wszystkie modele są bardzo do siebie zbliżone, różnice miar dobroci są marginalne. Najbardziej skomplikowane modele mają wysokie AIC i choć wartości AUC są wyższe od pozostałych przestaniemy się nimi interesować w kolejnych analizach. Zrobimy to zgodnie z zasadą, że im prostszy model tym lepiej, a dodawanie kolejnych zmiennych nie przynosi w tym wypadku żadnych wymiernych korzyści, a wręcz szkodzi.

Odrzucamy więc dwa, skomplikowane modele o wysokim AIC, tzn *FullExtra* oraz *Age5Factor*. Model *FullExtra* może się jeszcze przydać podczas zachł;annego algorytmu wybierania modelu - obecność wszystkich zmiennych może być pomocna.

3.6 Sensitivity, specificity oraz precision

Przeanalizujemy 3 współczynniki, które opisują nasz model, *sensitivity*, *specificity* i *precision*, dane one są następującymi wzorami:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

Każdy określa inną zależność na jakiej nam zależy. W naszym problemi najistotniejsz wydaje się *sensitivity* - nie chcemy wypuszczać na wolność więźniów, którzy dopuścili się przestępstwa będąc na zwolnieniu warunkowym. Tzn. chcemy maksymalizować tę wartość.

Wspołczynniki przedstawiają się następująco:

Tablica 5: SSP&TruePositive&FalseNegative

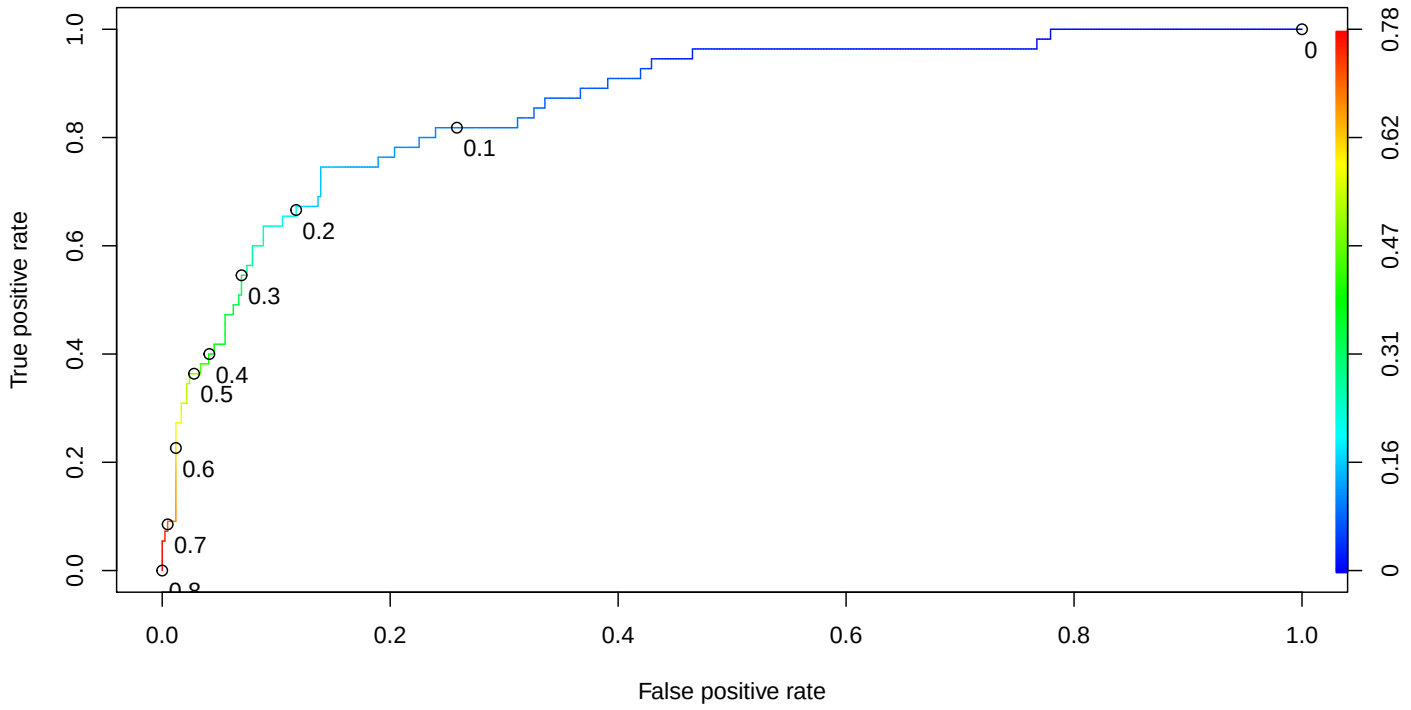
	Sensitivity	Specificity	Precision
Full	0.364	0.974	0.645
AgeGroupFactor	0.327	0.964	0.545
Serve	0.364	0.971	0.625
Sentence	0.327	0.981	0.692
Simple	0.327	0.983	0.720
Driver	0.364	0.974	0.645

Różnice są bardzo małe, dokładneść każdego z modelu jest w 90% co jest zadawalającą wartością.

3.7 Krzywa ROC

Przejdziemy teraz do wykresów krzywej ROC (*Receiver Operator Characteristic*). Przeanalizujemy ją tylko dla jednego modelu, aby nie pokazywać bardzo podobnej krzywej siedem razy.

Krzywa ROC dla modelu *Full*:



Widzimy, że nasz model dobrze dopasowuje się do problemu. Tak jak wcześnie powiedzieliśmy, chcielibyśmy masymalizować *TPR* czyli *sensitivity*. Gdybyśmy "przykręcili śrubę" w naszym modelu wtedy moglibysmy osiągnąć poziom *TPR* na poziomie 50%, czyli w co drugim przypadku nasz model słusznie oceniałby danego więźnia jako "ryzykownego". To powinno zostać poddane dalszej dyskusji, w gronie ekspertów.

Sprawdźmy jednak jak by wyglądała sytuacja współczynników SSP po obniżeniu progu odcięcia do 0.35:

Tablica 6: SSP

	Sensitivity	Specificity	Precision
Full	0.473	0.945	0.531
AgeGroupFactor	0.418	0.945	0.500
Serve	0.436	0.942	0.500
Sentence	0.491	0.940	0.519
Simple	0.382	0.950	0.500
Driver	0.400	0.950	0.512

Po obniżeniu progu, zgodnie z zapowiedziami, wzrosło *sensitivity*, oczywiście kosztem *specificity*, ale głównym priorytet powinno być zapewnienie bezpieczeństwa, a nie obniżenie kosztów utrzymania więzień. Zauważmy, że różnica pomiędzy modelami *Simple*, a *Full* znacząco zmalała.

3.8 Algorytm zachłanny.

Użyjemy algorytmu zachłannego posługującego się miarą dobroci AIC, by wybrać możliwie uproszczony, ale nadal dobry model. Działaniu algorytmu poddamy modele *Full*, *Simple*. Poniżej przedstawiamy jego kroki:

```
paroleTrain.glmFullSTEP=step(glm(violator~.,data=paroleTrain[,c(-1,-2,-10,-11,-12,-13,-14)],
                                 family="binomial"))
```

```
Start:  AIC=261.56
violator ~ age + state + time.served + max.sentence + multiple.offenses +
    crime

                    Df Deviance    AIC
- crime              3   241.82 257.82
- time.served        1   239.94 259.94
- age                1   240.56 260.56
<none>                   239.56 261.56
- max.sentence       1   242.01 262.01
- multiple.offenses  1   250.06 270.06
- state              3   318.53 334.53

Step:  AIC=257.82
violator ~ age + state + time.served + max.sentence + multiple.offenses

                    Df Deviance    AIC
- time.served        1   242.12 256.12
- age                1   242.46 256.46
<none>                   241.82 257.82
- max.sentence       1   243.92 257.92
- multiple.offenses  1   254.30 268.30
- state              3   321.18 331.18

Step:  AIC=256.12
violator ~ age + state + max.sentence + multiple.offenses

                    Df Deviance    AIC
- age                1   242.73 254.73
<none>                   242.12 256.12
- max.sentence       1   244.18 256.18
- multiple.offenses  1   254.61 266.61
- state              3   321.40 329.40

Step:  AIC=254.73
violator ~ state + max.sentence + multiple.offenses

                    Df Deviance    AIC
<none>                   242.73 254.73
- max.sentence       1   244.85 254.85
- multiple.offenses  1   254.97 264.97
- state              3   322.97 328.97
```

Tak wyglądały kroki dla modelu z pełnymi danymi, teraz czas na model z danymi uproszczonymi.

```
paroleTrain.glmSimpleSTEP=step(glm(violator~.,data=paroleTrain[,c(-1,-2,-3,-5,-6,-8,-11)],
                                   family="binomial"))
```

```
Start:  AIC=259.72
violator ~ state + multiple.offenses + age.group.factor + time.served.factor +
    max.sentence.factor + isDriver

                      Df Deviance    AIC
- max.sentence.factor  2   239.49 257.49
- time.served.factor   1   237.90 257.90
- isDriver             1   238.21 258.21
- age.group.factor     2   241.57 259.57
<none>                     237.72 259.72
- multiple.offenses    1   245.42 265.42
- state                3   301.91 317.91

Step:  AIC=257.49
violator ~ state + multiple.offenses + age.group.factor + time.served.factor +
    isDriver

                     Df Deviance    AIC
- time.served.factor  1   239.72 255.72
- isDriver            1   240.04 256.04
<none>                    239.49 257.49
- age.group.factor    2   243.99 257.99
- multiple.offenses   1   247.75 263.75
- state               3   314.00 326.00

Step:  AIC=255.72
violator ~ state + multiple.offenses + age.group.factor + isDriver

                    Df Deviance    AIC
- isDriver           1   240.30 254.30
<none>                   239.72 255.72
- age.group.factor   2   244.20 256.20
- multiple.offenses  1   247.94 261.94
- state              3   326.99 336.99

Step:  AIC=254.3
violator ~ state + multiple.offenses + age.group.factor

                    Df Deviance    AIC
<none>                   240.30 254.30
- age.group.factor   2   244.85 254.85
- multiple.offenses  1   248.95 260.95
- state              3   329.59 337.59
```

Algrytm tenże wyłonił dwa modele które przedstawiamy poniżej
(dla modelu *FullExtra*, ostateczny model uproszczony okazał się tożsamy z uproszczonym modelem *Simple*).

Tablica 7: Model z pełnymi danymi

	Estimate	Std. Error	z value	Pr(>
(Intercept)	-3.125	0.7497	-4.169	3.066e-05
stateLouisiana	1.458	0.5444	2.678	0.007409
stateOther	0.02414	0.4613	0.05233	0.9583
stateVirginia	-3.194	0.6895	-4.633	3.613e-06
max.sentence	0.07447	0.05176	1.439	0.1503
multiple.offenses1	1.377	0.3984	3.456	0.0005479

Tablica 8: Model z uproszczonymi danymi

	Estimate	Std. Error	z value	Pr(>
(Intercept)	-17.02	760.6	-0.02237	0.9822
stateLouisiana	1.312	0.5234	2.507	0.01219
stateOther	0.166	0.4622	0.3592	0.7194
stateVirginia	-2.9	0.6763	-4.287	1.809e-05
multiple.offenses1	1.169	0.3997	2.924	0.003454
age.group.factorMiddle	14.94	760.6	0.01964	0.9843
age.group.factorOld	14.08	760.6	0.01851	0.9852

Takim sposobem otrzymaliśmy kilka sensownych modeli, kolejnym krokiem będzie przetestowanie ich na zbiorze testowym, oraz wybór najlepszego.

3.9 Walidacja

Walidację przeprowadzimy dal 4 modeli, *Full* oraz *Simple* przed i po upraszczaniu algorytmem zachłannym ze względu na AIC. Poniżej przedstawiamy *Confusion Matrix* dla każdego z modeli.

	FALSE	TRUE
0	172	8
1	15	8

(a) Pełne dane, uproszczony model

	FALSE	TRUE
0	173	7
1	17	6

(b) Pełne dane, pełny model

	FALSE	TRUE
0	171	9
1	15	8

(c) Uproszczone dane, uproszczony model

	FALSE	TRUE
0	174	6
1	19	4

(d) Pełne dane, pełny model

Tablica 9: Porównanie Confusion Matrix dla danych na zbiorze walidacyjnym

Widzimy, że model działający na uproszczonych danych lepiej zachowuje się w naszym zadaniu - przy standardowym poziomie odcięcia jego *TPR* było wyższę niz modelu działającego na pełnych danych. Pondto po raz kolejny korzystamy tytaj z zależności prosty model > skomplikowany model.

Zobaczmy jak kształtuje *sensitivity*, *specificity* i *precision*:

Tablica 10: SSP

	Sensitivity	Specificity	Precision	.35 Sensitivity	.35 Specificity	.35 Precision
FullSTEP	0.35	0.96	0.50	0.39	0.92	0.39
Full	0.26	0.96	0.46	0.43	0.89	0.34
SimpleSTEP	0.35	0.95	0.47	0.35	0.95	0.47
Simple	0.17	0.97	0.40	0.48	0.91	0.39

Powyższa tabela potwierdza nasze wcześniejsze obserwacje: model niedoszacowuj ryzyka związanego z danym więźniem. Jednak gdy tylko zmienimy próg odcięc na 0.35 zaczyna działać rozsądniej. Modele STEP, tzn. mo modyfikacji algorytme zachłannym są niewrażliwe na zmienę miejsca odcięcia, są bardziej "stabilne". Z kolei modele sprzed modyfikacji dają niskie *sensitivity*, co jest niepożądane w naszym problemie. Wszystkie modele dają bardzo dobre *specificity*, które oczywiście spada wraz ze zmianą granicy odcięcia.

3.10 Podsumowanie

Uzyskalismy 4 modele, wszystkie na całkiem niezłym poziomie. Wskazanie ostatecznego modelu jest trudne my się tego nie podejmiemy. Każdy z nich ma swoje wady i zalety, dlatego być może warto by było skombinować je jakoś ze sobą. Wśród wszystkich modeli te same czynniki są "silne". Daje nam to podstawy myśleć, że faktycznie to, że dany więzień dopuści się przestępstwa będąc na zwolnieniu warunkowym, jesteśmy wstanie określić z wysokim prawdopodobieństwe tylko za pomocą jego historii. Gdybyśmy mieli więcej danych lub były by one bardziej szczegółowe (więcej informacji o rejonie odsiadki, rodzaju przewinienia) to byc może udałoby się zbudować jeszcze dokładniejszy model. Nie zmienia to jednak fakty, że w dużym stopniu udało nam się rozwiązać problem bezarbitrażowej oceny więźniów, tylko na podstawie suchych faktó. Jest to z jednej strony niebezpieczne, gdyż traktjujemy człowieka bardzo przedmiotowa, ale z drugiej strony zabezpieczamy się przed subiektywnością decyzji ludzkich.